VIDEO TEXT UNDERSTANDING ACROSS DIVERSE SCENES: AN EVIDENCE-GUIDED BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Video text understanding involves interpreting on-screen text together with its visual context. Audits of existing benchmarks have identified questions answerable from a single frame or by text extraction alone, motivating broader assessment of this joint understanding. We introduce an evidence-guided benchmark for video text understanding across diverse scenes. Its 16,165 evaluation items comprise multiple-choice questions and true/false judgments over 2,629 videos, organized into 10 primary scene categories and 34 fine-grained subcategories. The construction workflow plans structured evidence before generating questions, followed by automated verification, text-only filtering, and human correction. We evaluate 26 open-weight multimodal checkpoints spanning 1B–235B parameters. On the matched multiple-choice roster, mean accuracy is 55.15% with video versus 41.49% with questions and options alone. Paired judgments reveal difficulty both accepting supported statements and rejecting their counterfactual counterparts, while performance varies across scenes. We further manually review a subset of 500 base questions and their associated variants for controlled evidence diagnostics: at equal frame counts, sampling distinct evidence windows outperforms repeating one window center on average. The benchmark provides a testbed for evaluating joint text–visual understanding and diagnosing remaining model weaknesses.

1 INTRODUCTION

Video text understanding involves interpreting on-screen text together with its visual context: a scoreboard updates after an event, a road sign refers to a location, and instructions describe visible objects or actions. We use *on-screen text* broadly to include scene text, subtitles, overlays, scoreboards, and interface text. Our focus is their joint interpretation with visual entities, states, and events, rather than text extraction alone.

Early benchmarks do not always require these abilities. An audit of M4-ViteVQA and NewsVideoQA found that 92%/95% of sampled questions were single-frame answerable and 81%/98% were extractive (Jahagirdar et al., 2023). Recent datasets raise evidence requirements through cross-frame fusion and multi-segment cues (Guo et al., 2026; Ben Ami et al., 2026), but scaling such questions while preserving answer validity and evidence provenance remains challenging. Moreover, VideoQA models can exploit linguistic priors and answer-position biases without using the video (Ko et al., 2023; Grunde-McLaughlin et al., 2021; Loginova et al., 2025). Dataset size and balanced labels alone cannot establish visual understanding.

We introduce a multi-scene Video Text QA benchmark targeting this joint text–visual understanding. It contains 2,023 four-option multiple-choice questions (MCQA) and 14,142 true/false statements, totaling 16,165 evaluation items from 7,757 source questions over 2,629 videos. Scene labels organize coverage into 10 primary categories and 34 subcategories; intended ability targets distinguish the kinds of understanding sought within those scenes. Our evidence-structured LLM annotation workflow plans visible text, temporal events, answer slots, and proposed evidence intervals before question generation, automated verification, text-only filtering, and human inspection and repair.

Linked formats share a source question, supporting evidence, and answer semantics. Each judgment pair contains a supported statement and a counterfactual that changes one answer slot; the statements are presented independently. Pair accuracy requires both accepting the supported statement

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.

and rejecting the counterfactual. Models involved in construction, checking, adjudication, or selection are excluded from the reported evaluation. Question-and-option-only controls probe linguistic shortcuts, scene-level results reveal domain variation, and four interventions on a manually reviewed 500-source subset examine evidence use. Equal-frame-count comparisons support complementary window coverage, without establishing general temporal-order sensitivity.

Our contributions are: (1) a multi-scene video text understanding benchmark with linked MCQA/judgment items and traceable source, scene, ability, and evidence records; (2) an evidence-structured LLM annotation workflow combining model generation and automated checks with human quality control; and (3) a large-scale evaluation of 26 open-weight multimodal checkpoints spanning 1B–235B parameters, with diagnostics of textual shortcuts, acceptance/rejection consistency, scene variation, and evidence coverage.

2 Related work

Video Text QA and evidence requirements. NewsVideoQA and M4-ViteVQA establish text-centric video question answering across news, languages, and categories (Jahagirdar et al., 2022; Zhao et al., 2022). EgoTextVQA covers egocentric driving and housekeeping, ViTXT-GQA adds spatiotemporal grounding, and VidText combines tasks and granularities (Zhou et al., 2025; 2024; Yang et al., 2025). MME-VideoOCR spans 25 tasks and 44 scenarios; ViTexQA checks cross-frame text fusion; HERBench requires at least three non-overlapping cues from distinct segments (Shi et al., 2025; Guo et al., 2026; Ben Ami et al., 2026). These works motivate explicit evidence requirements. Our focus is a scalable construction funnel that records intended abilities and temporal evidence, then audits the resulting benchmark. Appendix Table 10 compares the closest designs.

Human and model-assisted construction. M4-ViteVQA, MME-VideoOCR, and ViTexQA rely on human or expert authoring and verification (Zhao et al., 2022; Shi et al., 2025; Guo et al., 2026). Automatic scale has precedents: Just Ask derives 69M training triplets from narrated videos, while EgoTextVQA uses GPT-4o for initial questions followed by human refinement (Yang et al., 2021; Zhou et al., 2025). Our contribution is structured planning of text, temporal evidence, and answer slots before generation, followed by automated checks and human correction; generation counts do not demonstrate greater labor efficiency.

Shortcut and evidence diagnostics. Flipped-VQA and AGQA show that linguistic biases can support non-visual predictions even after balancing (Ko et al., 2023; Grunde-McLaughlin et al., 2021). No-video baselines on MovieQA, TVQA, and NExT-QA further motivate separating language predictability from video understanding (Yang et al., 2020; Xiao et al., 2025); text-context leakage and answer-position bias require distinct controls (Jahagirdar et al., 2023; Loginova et al., 2025; Shi et al., 2025). Our text-filtered MCQA and independent supported/counterfactual judgments address complementary presentation cues, while Pair tests acceptance/rejection consistency. Window interventions follow the diagnostic logic of perturbed video evidence (Liu et al., 2024; Cai et al., 2024; Li et al., 2026). Scene breakdowns have precedents in M4-ViteVQA, EgoTextVQA, and MME-VideoOCR; our contribution is the measured variation, not category count alone.

3 Benchmark construction

3.1 Scope and provenance

A *source question* is a base question with supporting evidence and answer semantics shared by its format variants; an *evaluation item* is one independently presented MCQA question or judgment statement. The evaluated bank contains 2,023 four-option MCQA items and 14,142 judgment statements (7,071 supported/counterfactual pairs), totaling 16,165 items from 7,757 sources and 2,629 video paths. Of 1,337 sources shared by both formats, 873 belong to the cross-frame pool. Source and parent-video identifiers prevent linked variants or clips from being treated as independent sources.

Before format-specific filtering, the source core contains 7,081 VidText and 4,005 non-VidText questions. Other inputs include RoadText-3K, NewsVideoQA, and DSText V2 (Garcia-Bordils et al.,

2022; Jahagirdar et al., 2022; Wu et al., 2023). These are counts of our derived records, not original dataset sizes. We retain clip intervals, original language fields, generation stages, and evidence provenance.

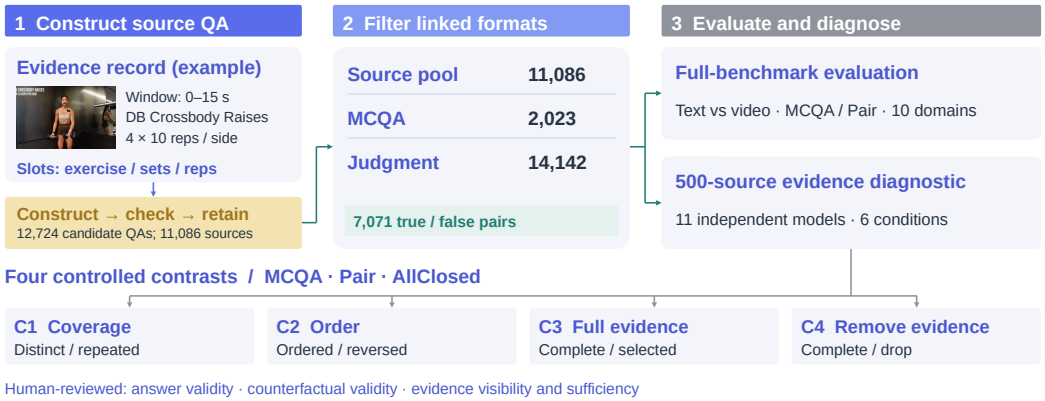


Figure 1: Benchmark construction and evaluation overview. (1) Construct and verify source QA using structured evidence. (2) Derive and filter linked MCQA and judgment items. (3) Evaluate models on the full benchmark and a manually reviewed 500-source subset for evidence diagnostics. Human correction precedes evaluation, which excludes recorded pipeline participants.

3.2 Evidence-structured generation

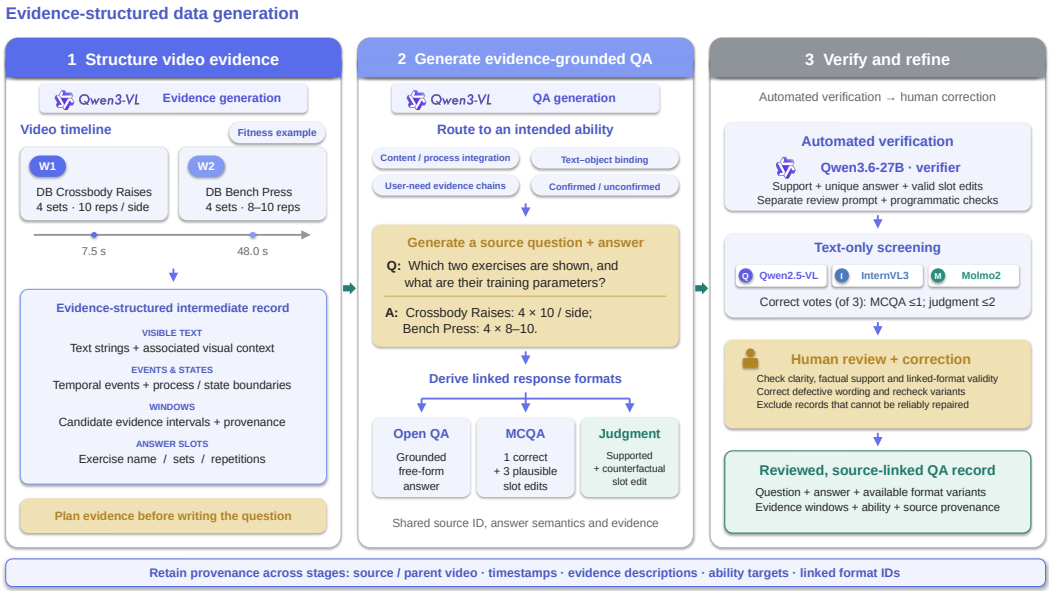


Figure 2: Evidence-structured data generation: video text, events, windows, and answer slots guide question generation, linked formats, automated checks, and human correction. Qwen3-VL generates structured evidence and QA, with questions produced directly in English; Qwen3.6-27B performs automated verification. Text-only selectors are Qwen2.5-VL-7B, InternVL3-8B, and Molmo2-4B. The fitness example is abridged; timestamps mark evidence-window centers.

Figures 1–2 summarize the workflow. Evidence plans represent visible text, temporal events, process/state boundaries, proposed evidence intervals, and answer slots. Generation is routed to the intended ability targets in Table 1. These labels guide question design; they do not certify multi-frame necessity (Figure 3).

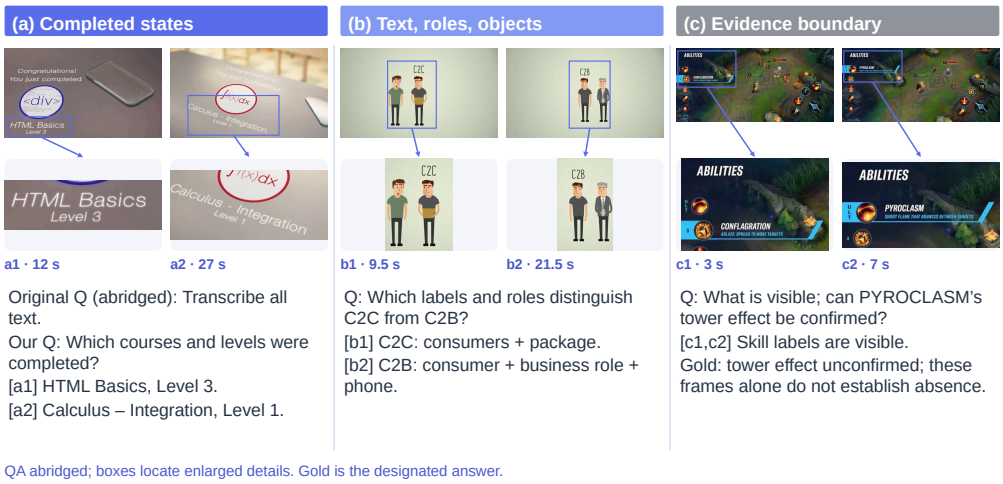


Figure 3: Human-reviewed source-bank examples of (a) completed-state selection, (b) labels tied to roles and objects, and (c) visible labels versus an unconfirmed effect. Boxes enlarge evidence; panel IDs link abridged answers. Appendix K provides full records and limitations. These cases illustrate intended demands, not their prevalence.

The upstream pool contains 12,724 candidate open-answer questions (8,269 VidText and 4,455 other sources). MCQA construction and automatic validation retain 11,086 source questions (87.13%). Questions and options are generated directly in English, without a translation stage. For each, the generator produces a supported answer and three plausible, type-matched alternatives that change the same semantic slot. Qwen3.6-27B checks support, contradictions, answer uniqueness, option count, and position/length constraints; programmatic checks follow. Human review then corrects or excludes defective records. Slot edits can still leave linguistic cues, as with caption perturbations (Cai et al., 2024), motivating additional text-only filtering.

3.3 Format-specific filtering

Let c_{ij} indicate whether text-only selector j answers MCQA source i correctly. We retain

$$\mathcal{D}_{\text{MCQA}} = \{i : \textstyle\sum_{j=1}^{3} c_{ij} \leq 1\}. \tag{1}$$

Qwen2.5-VL-7B, InternVL3-8B, and Molmo2-4B receive only questions and options (Bai et al., 2025; Zhu et al., 2025; Clark et al., 2026). Incorrect or unparseable responses count as no correct vote. The rule retains 803 zero-vote and 1,220 one-vote items. No video result enters selection; the cutoff was chosen after inspecting question profiles and was not preregistered.

Judgment generation pairs each supported statement with a counterfactual changing one answer slot. It yields 10,598 valid pairs (9,975 strict passes and 623 recovered; 488 abandoned sources). The same selectors receive each pair as a two-option contrastive question; at most two correct votes retains 7,071 pairs (66.72%). Evaluation presents statements as independent single-turn requests. True/false labels are exactly balanced, and randomized Accurate/Inaccurate order gives 7,071 correct A and 7,071 correct B positions. This removes four-way distractor comparison, but not linguistic or binary-label priors.

3.4 Scene and ability categories

A scene category describes the video's content domain, whereas an F label describes a question's intended ability target (Table 1); MCQA and judgment specify response formats. We organize scene metadata into 10 primary and 34 fine-grained secondary categories using a benchmark-specific hierarchy. Source-dataset priors are stored separately; generated scene descriptions guide labeling rather than treating dataset names or question roles as scene ground truth. Model agreement, larger-model adjudication, and 101 human metadata overrides assign labels to 2,823 canonical records corresponding to 2,770 pre-filter video paths. These overrides are separate from QA review. Road and

Table 1: Intended ability targets. F0 denotes basic direct QA; F1–F6 guide richer question construction. Labels specify design intent, not proven minimal evidence requirements.

Code	Ability target	Intended question demand
F0	Natural direct QA	Read or retrieve directly supported information.
F1	Content/process integration	Relate text content to visible actions or processes.
F2	State/information boundary	Distinguish stages, current states, and final states.
F3	Four-stage timeline	Organize key stages using temporal anchors.
F4	Text-carrier spatial binding	Associate text with its carrier, object, or location.
F5	User-need evidence chain	Combine evidence to answer a practical user need.
F6	Confirmed/unconfirmed boundary	Separate supported facts from unconfirmed claims.

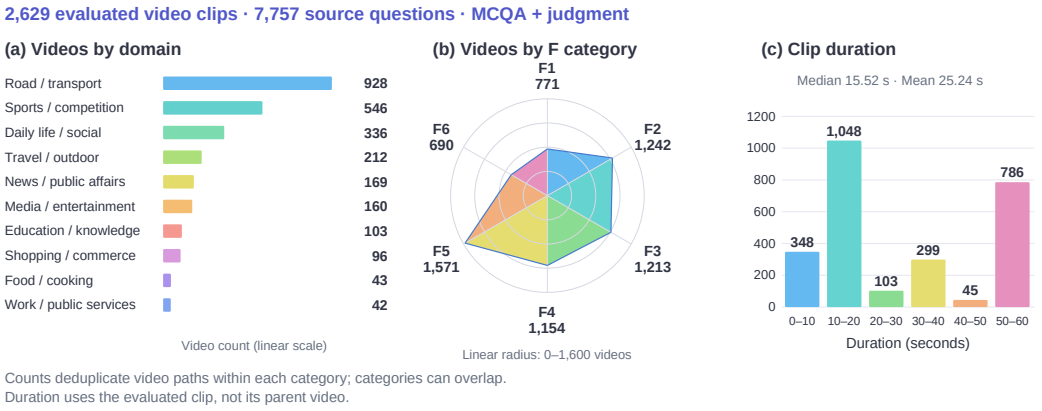


Figure 4: Video coverage in the evaluated MCQA/judgment union. (a) Videos by content domain. (b) Videos associated with questions targeting F1–F6 (Table 1), using a linear radial count scale; F0 is omitted here. Videos are deduplicated within categories, which can overlap. (c) Video-stream durations of all 2,629 evaluated clips, each counted once in 10-second bins.

sports comprise 58.04% of the pre-filter source core, so coverage does not imply balance. Complete response-format examples appear in Appendix Figures 12–13.

Figure 4 profiles the 2,629 evaluated video clips. Median/mean clip duration is 15.52/25.24 seconds; category-specific duration distributions appear in Appendix Figure 10.

3.5 Human Quality Assurance

Reviewers inspect each source question, linked variants, video, and evidence records without using evaluation outcomes. Checks cover answerability, gold-answer support and uniqueness, text readability, intended-ability fit, distractor/counterfactual validity, English wording, and cross-format consistency. Unsupported, inaccurate, or ambiguous content is corrected and rechecked; sources without a reliable evidence-grounded repair are excluded, keeping judgment pairs intact.

All 500 evidence-diagnostic sources and their closed-form variants received complete manual review before use as the experimental benchmark. Review used source videos and decoded inputs, and corrections were incorporated into the experimental set. This validates content but does not establish minimal window necessity. Appendix Table 8 details the review scope; we report neither a human-performance baseline nor inter-annotator agreement.

4 Evaluation Protocol

4.1 Release-level evaluation

We evaluate multimodal checkpoints with publicly downloadable weights (open-weight); this term does not imply that training data and all training code are open. The checkpoints span 1B–235B

parameters, including dense and mixture-of-experts architectures. All reported models are excluded from recorded construction, checking, metadata adjudication, and selection stages. "Independent" denotes pipeline independence, not disjoint pretraining data or model families. MCQA covers 26 video models; its text–video comparison uses the same 24 matched models and 2,023 items. Judgment uses 23 models with complete text/video predictions on 7,071 pairs. We average per-model accuracies equally and score unparseable outputs as incorrect. Video handling differs across implementations, so leaderboard comparisons are not compute-matched.

4.2 Frozen evidence diagnostic

From the 873 shared cross-frame sources, outcome-blind requirements—two to eight merged windows, at least three seconds of evidence span, a valid judgment pair, and bounded option-length imbalance—retain 564 eligible source questions. Selection uses scene, ability, window count, dataset, video/parent grouping, and historical-overlap metadata, without current or historical predictions. The reviewed 500-source set contains 424 video paths and 273 parents, with at most two sources per video and eight per parent. A reversible deterministic MCQA permutation assigns exactly 125 gold answers to each of A–D.

Each source supplies MCQA and two independent judgments. Before new inference, we froze the 11-model pipeline-external roster, four contrasts, scoring, statistics, and report-all policy. Five new conditions yield $11 \times 500 \times 5 \times 3 = 82{,}500$ predictions; 16,500 reused complete-window predictions bring the total to 99,000. Appendix D records the conditions and run integrity.

4.3 Evidence interventions

For K windows, center inputs use one frame per window; window inputs use the 25% and 75% positions. A metadata heuristic selects one window by token overlap between its evidence description and the changed judgment slot, supported value, and counterfactual value, without model outcomes. We compare (C1) distinct versus repeated centers at equal K; (C2) ordered versus reversed centers with the same frame multiset; (C3) complete $2K$-frame input versus the selected two-frame window; and (C4) complete input versus dropping that window, leaving $2(K-1)$ frames. C1/C2 control frame count or identity; C3/C4 also change input frame counts, so their effects cannot isolate evidence content from the amount of visual input. The selected window is not an oracle or minimal-evidence label. All contrasts are exploratory because the source set and intervention family were developed previously; one reversal does not test arbitrary permutations or eliminate content-based chronology cues.

4.4 Scoring and uncertainty

Item accuracy scores individual statements; Pair requires both statements from a source to be correct, and AllClosed additionally requires its MCQA answer to be correct. Let a_i, t_i, f_i indicate correctness for MCQA, supported and counterfactual statements:

$$\text{Pair} = \frac{1}{N}\sum_i t_i f_i, \qquad \text{AllClosed} = \frac{1}{N}\sum_i a_i t_i f_i. \tag{2}$$

These conjunctions measure operational consistency, related to grouped scoring in TemporalBench and Video-MME-v2 (Cai et al., 2024; Fu et al., 2026). Always predicting true yields 50% Item but zero Pair; independent fair guessing yields 50% Item and 25% Pair. Correlated decisions can score below this Pair reference, so we also report supported/counterfactual accuracy. Paired source differences use 10,000 bootstrap draws over 273 parent-video clusters with equal model weights. Intervals condition on the fixed roster and are unadjusted for four contrasts and three metrics.

5 Results

5.1 Text filtering reduces, but does not eliminate, shortcuts

Using saved predictions from the same 24 pipeline-independent models, mean text-only accuracy falls from 72.52% on the 11,086-item pre-filter MCQA pool to 41.49% on the retained 2,023 items, a

Table 2: MCQA and judgment accuracy (%) for the 23 models with complete results in both formats, sorted by video MCQA. Hatched dark blue and solid light blue bars denote video and text-only input, respectively. MCQA uses 2,023 items; judgment uses 7,071 pairs from a different source set. The displayed intersection excludes incomplete rows; MCQA aggregates in the text use 24 matched models (full roster: Appendix Table 4).



31.03-point reduction. On the retained set, video accuracy is 55.15%, exceeding text-only accuracy by 13.66 points; both exceed the 25% chance level. Published no-video baselines under different protocols appear in Appendix Table 6. Across the roster, text/video accuracy correlates strongly ($\rho = 0.863$). Our text-only input contains no video transcription (Jahagirdar et al., 2023); residual accuracy therefore reflects question–option signal, not proof of visual necessity.

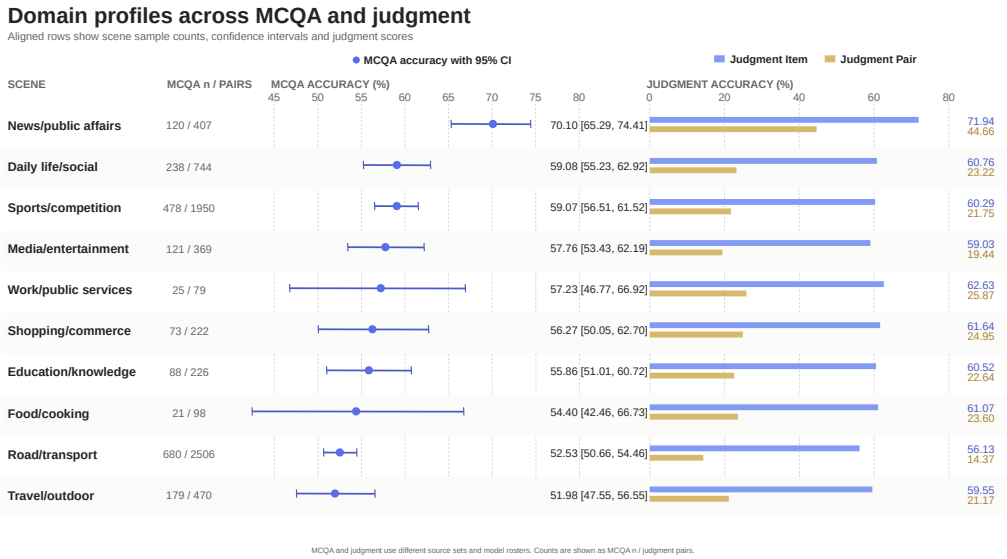
Table 2 aligns 23 fully evaluated models; Appendix Table 4 gives the full MCQA roster.

Repository-year and parameter-scale trends are reported in Appendix Figure 6; these associations are descriptive and confounded by roster composition.

5.2 Balanced judgment pairs expose acceptance/rejection asymmetry

On the 23-model judgment roster, video raises model-macro item accuracy from 50.72% to 59.50%, but raises Pair more sharply from 3.91% to 20.65%, a 16.75-point gain. Under video, supported-statement accuracy is 78.96% while counterfactual accuracy is 40.05%, a 38.90-point asymmetry. Thus, balanced labels do not ensure balanced acceptance and rejection. The text-only Pair score of 3.91% further exposes errors hidden by item averages. Judgment and MCQA cover different item sets, so their values are not a causal format comparison.

Table 3: Video-input scene profiles. Rows show sample counts, MCQA accuracy with parent-video-cluster 95% confidence intervals, and judgment Item and Pair accuracy. The formats use different source sets and model rosters.



Complete per-model results appear in Appendix Table 5.

Appendix Figure 7 illustrates how accepting both statements yields Item $= 1/2$ but Pair $= 0$.

5.3 A SINGLE OVERALL SCORE HIDES SCENE-CONDITIONED GAPS

Table 3 reports scene profiles for 26 MCQA models. Model-macro accuracy ranges from 70.10% for news/public affairs to 51.98% for travel/outdoor, an 18.12-point gap. Their parent-video-cluster intervals do not overlap. Across models, the median difference between each model's best and worst scene is 22.86 points, showing variation hidden by a single overall score.

Judgment shows domain variation as well: across 23 independent models and 7,071 pairs, news/public affairs reaches 71.94% Item and 44.66% Pair, versus 56.13% and 14.37% for road/transport (gaps of 15.82 and 30.29 points). These are within-format comparisons; the MCQA and judgment source sets and model rosters differ.

Overall and scene-macro model rankings remain strongly correlated (Spearman $\rho = 0.967$), so scenario reweighting does not overturn the broad leaderboard. It nevertheless changes individual conclusions: among the full 26-model MCQA roster, Molmo2-8B (2025; 8B) moves from sixth by overall accuracy to twelfth by scene macro and spans 37.38 points across scenes. These are conditional descriptions, not causal scene effects: scene labels are metadata/model-derived, scene sizes are imbalanced, and scene remains confounded with source, ability, and video/text properties.

5.4 PIPELINE-EXTERNAL MODELS SUPPORT COMPLEMENTARY EVIDENCE COVERAGE

On the human-reviewed 500-source set, distinct window centers at equal K outperform a duplicated center by 1.9 points on MCQA, 2.5 on Pair, and 1.9 on AllClosed; all intervals exclude zero, and MCQA/AllClosed differences are positive for 9/11 and 10/11 models. This equal-frame-budget comparison supports complementary temporal coverage. Appendix Table 7 reports all four effects for each of the 11 models.

Complete windows outperform one selected window by 3.0/4.0/3.5 points and the drop-window input by 3.6/6.3/4.9 points on MCQA/Pair/AllClosed; MCQA improves for all 11 models in both comparisons, but unequal frame counts confound them. Ordered-minus-reversed effects are only

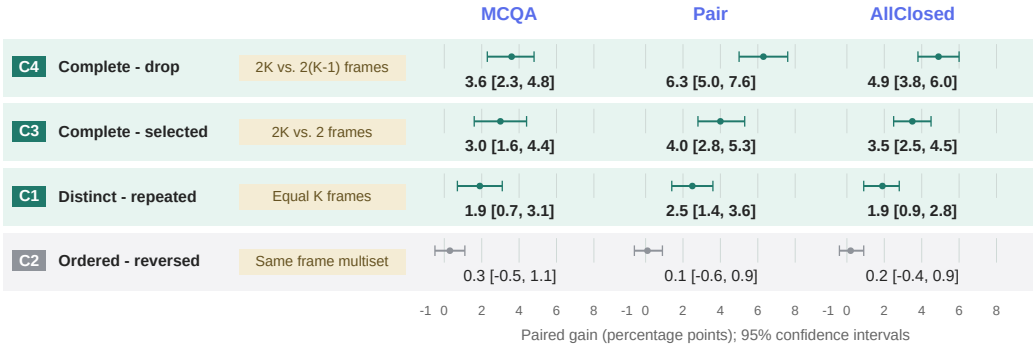


Figure 5: Evidence interventions on 500 reviewed sources and 11 pipeline-external models. Points show equal-weight model-macro paired differences; bars show unadjusted 95% intervals from 10,000 parent-video-cluster bootstrap draws. Positive values favor the first condition. C1/C2 control frame count or identity; C3/C4 change the frame budget.

0.3/0.1/0.2 points and all intervals include zero. Thus the exploratory results support evidence coverage, not general temporal-order sensitivity; all four contrasts are reported regardless of direction.

A reconstruction of all six inputs and model outputs for one fitness source appears in Appendix Figure 8; a separate scoreboard trace appears in Appendix Figure 9.

6 Discussion and Limitations

Evidence-structured generation combines scalable authoring with human repair, but throughput does not establish a labor-efficiency advantage. Text filtering reduces rather than eliminates shortcuts; independent judgments remove MCQA-specific comparison cues while retaining linguistic and label priors. Year/scale trends and scene gaps depend on model roster, source, ability, and video/text properties. Model-specific video preprocessing also prevents compute-matched leaderboard comparisons.

Equal-K interventions support complementary evidence coverage. Single/drop-window comparisons also change input frame counts, and reversal gives no reliable temporal-order effect. Human review of all 500 diagnostic sources validates content without making sparse decoding exhaustive or proving minimal multi-window necessity.

Conclusion. We present a multi-scene benchmark designed to assess the joint understanding of on-screen text and visual context. Its 16,165 multiple-choice and judgment items complement text extraction with questions targeting processes, states, spatial associations, and evidence-based conclusions across diverse scenes. Evaluation of 26 open-weight multimodal checkpoints reveals substantial remaining errors, difficulty consistently accepting supported statements and rejecting counterfactuals, and performance differences across scenes. Controlled diagnostics further support the value of complementary evidence coverage at equal frame counts. Together, these results provide a testbed for tracking progress in video text understanding and diagnosing model weaknesses beyond a single aggregate score. The structured-evidence annotation workflow combines LLM generation, automated checks, and human correction; adapting this workflow to additional tasks and scenarios is a promising direction, whose transfer effectiveness remains to be evaluated.

AI USE STATEMENT

Generative models were used in the data workflow for evidence-oriented planning, direct English question and counterfactual generation, automated verification by Qwen3.6-27B, metadata adjudication, and text-only selection. Human reviewers inspected generated benchmark records against their source evidence and corrected or removed unsupported, inaccurate, ambiguous, or otherwise inappropriate content; all 500 sources in the evidence diagnostic received manual quality review. Every recorded model that participated in the automated stages is excluded from release-level benchmark results and from the current 500-source evidence diagnostic. AI assistance was also used to organize existing materials, inspect implementation details, search and organize references, draft this manuscript, prepare figure/table generation code, and execute the frozen evidence-diagnostic queue.

ETHICS STATEMENT

This study derives questions from existing video datasets whose frames and text may contain personal or copyrighted information. Access to and use of source videos are governed by their original licenses and terms; derived annotations do not alter ownership of the underlying content. Human quality control is part of dataset curation. We report no external human-solving experiment or participant-performance results.

REPRODUCIBILITY STATEMENT

Section 3 describes construction, filtering, and counting units; Section 4 defines the evaluation scope, visual interventions, scoring, and statistical analysis. The appendix provides per-model results and the six evidence-input conditions. Figures and tables are computed from saved predictions and frozen statistics without additional model inference during figure generation. The evidence diagnostic analyzes 82,500 new condition predictions and 16,500 reused complete-window predictions.

Appendix J documents diagnostic decoding, the shared prompt, output allowances, and answer parsing.

REFERENCES

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

Dan Ben Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. HERBench: A Benchmark for Multi-Evidence Integration in Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ami_HERBench_A_Benchmark_for_Multi-Evidence_Integration_in_Video_Question_Answering_CVPR_2026_paper.html.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. TemporalBench: Benchmarking Fine-grained Temporal Understanding for Multimodal Video Models. *arXiv preprint arXiv:2410.10818*, 2024. URL https://arxiv.org/abs/2410.10818.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding. *arXiv preprint arXiv:2601.10611*, 2026. URL https://arxiv.org/abs/2601.10611.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu

Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-MME-v2: Towards the Next Stage in Benchmarks for Comprehensive Video Understanding. *arXiv preprint arXiv:2604.05015*, 2026. URL https://arxiv.org/abs/2604.05015.

Sergi Garcia-Bordils, George Tom, Sangeeth Reddy, Minesh Mathew, Marçal Rusiñol, C. V. Jawahar, and Dimosthenis Karatzas. Read While You Drive – Multilingual Text Tracking on the Road. In *Document Analysis Systems*, pp. 756–770, 2022. doi: 10.1007/978-3-031-06555-2_51. URL https://doi.org/10.1007/978-3-031-06555-2_51.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Grunde-McLaughlin_AGQA_A_Benchmark_for_Compositional_Spatio-Temporal_Reasoning_CVPR_2021_paper.html.

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. URL https://arxiv.org/abs/2606.24602.

Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Watching the News: Towards VideoQA Models that can Read. *arXiv preprint arXiv:2211.05588*, 2022. URL https://arxiv.org/abs/2211.05588.

Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Understanding Video Scenes through Text: Insights from Text-based Video Question Answering. *arXiv preprint arXiv:2309.01380*, 2023. URL https://arxiv.org/abs/2309.01380.

Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. Large Language Models are Temporal and Causal Reasoners for Video Question Answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4300–4316, 2023. doi: 10.18653/v1/2023.emnlp-main.261. URL https://aclanthology.org/2023.emnlp-main.261/.

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*, 2026. URL https://arxiv.org/abs/2602.00288.

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? *arXiv preprint arXiv:2403.00476*, 2024. URL https://arxiv.org/abs/2403.00476.

Olga Loginova, Oleksandr Bezrukov, Ravi Shekhar, and Alexey Kravets. Addressing Blind Guessing: Calibration of Selection Bias in Multiple-Choice Question Answering by Video Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3216–3246, 2025. doi: 10.18653/v1/2025.acl-long.162. URL https://aclanthology.org/2025.acl-long.162/.

Yang Shi, Huanqian Wang, Wulin Xie, Huanyao Zhang, Lijie Zhao, Yi-Fan Zhang, Xinfeng Li, Chaoyou Fu, Zhuoer Wen, Wenting Liu, Zhuoran Zhang, Xinlong Chen, Bohan Zeng, Sihan Yang, Yushuo Guan, Zhang Zhang, Liang Wang, Haoxuan Li, Zhouchen Lin, Yuanxing Zhang, Pengfei Wan, Haotian Wang, and Wenjing Yang. MME-VideoOCR: Evaluating OCR-Based Capabilities of Multimodal LLMs in Video Scenarios. *arXiv preprint arXiv:2505.21333*, 2025. URL https://arxiv.org/abs/2505.21333.

Weijia Wu, Yiming Zhang, Yefei He, Luoming Zhang, Zhenyu Lou, Hong Zhou, and Xiang Bai. DSText V2: A Comprehensive Video Text Spotting Dataset for Dense and Small Text. *arXiv preprint arXiv:2312.01938*, 2023. URL https://arxiv.org/abs/2312.01938.

Junbin Xiao, Nanxin Huang, Hangyu Qin, Dongyang Li, Yicong Li, Fengbin Zhu, Zhulin Tao, Jianxing Yu, Liang Lin, Tat-Seng Chua, and Angela Yao. VideoQA in the Era of LLMs: An Empirical Study. *International Journal of Computer Vision*, 133(7):3970–3993, 2025. doi: 10.1007/s11263-025-02385-8. URL https://arxiv.org/abs/2408.04223.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just Ask: Learning To Answer Questions From Millions of Narrated Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1686–1697, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Yang_Just_Ask_Learning_To_Answer_Questions_From_Millions_of_Narrated_ICCV_2021_paper.html.

Jianing Yang, Yuying Zhu, Yongxin Wang, Ruitao Yi, Amir Zadeh, and Louis-Philippe Morency. What Gives the Answer Away? Question Answering Bias Analysis on Video QA Datasets. *arXiv preprint arXiv:2007.03626*, 2020. URL https://arxiv.org/abs/2007.03626.

Zhoufaran Yang, Yan Shu, Jing Wang, Zhifei Yang, Yan Zhang, Yu Li, Keyang Lu, Gangyan Zeng, Shaohui Liu, Yu Zhou, and Nicu Sebe. VidText: Towards Comprehensive Evaluation for Video Text Understanding. *arXiv preprint arXiv:2505.22810*, 2025. URL https://arxiv.org/abs/2505.22810.

Minyi Zhao, Bingjia Li, Jie Wang, Wanqing Li, Wenjing Zhou, Lan Zhang, Shijie Xuyang, Zhihang Yu, Xinkun Yu, Guangze Li, Aobotao Dai, and Shuigeng Zhou. Towards Video Text Visual Question Answering: Benchmark and Baseline. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e726197ffd401df4013cd9f81007b5cf-Abstract-Datasets_and_Benchmarks.html.

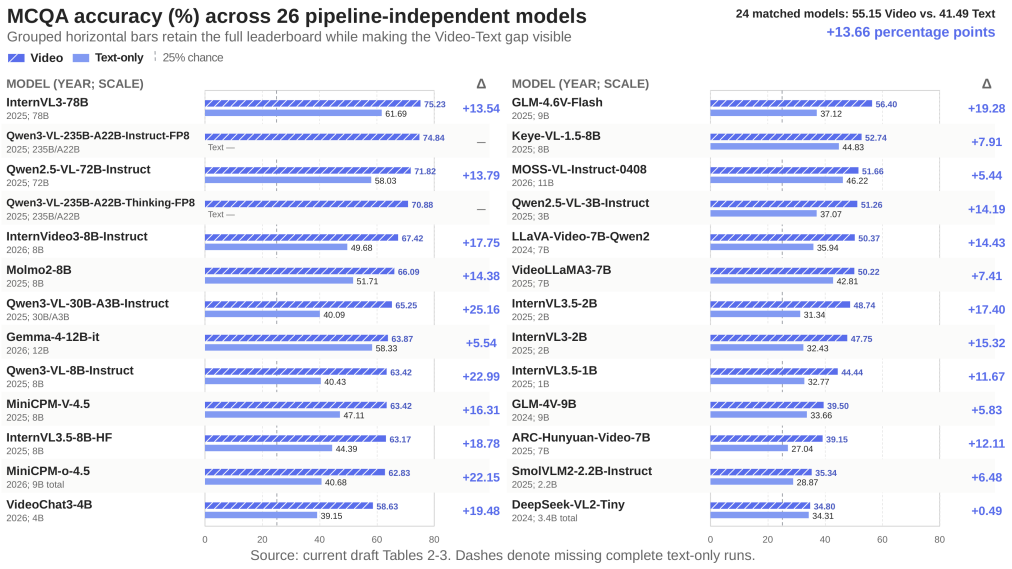
Sheng Zhou, Junbin Xiao, Xun Yang, Peipei Song, Dan Guo, Angela Yao, Meng Wang, and Tat-Seng Chua. Scene-Text Grounding for Text-Based Video Question Answering. *arXiv preprint arXiv:2409.14319*, 2024. URL https://arxiv.org/abs/2409.14319.

Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. EgoTextVQA: Towards Egocentric Scene-Text Aware Video Question Answering. *arXiv preprint arXiv:2502.07411*, 2025. URL https://arxiv.org/abs/2502.07411.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

A SUPPLEMENTARY MODEL RESULTS AND CASE ANALYSES

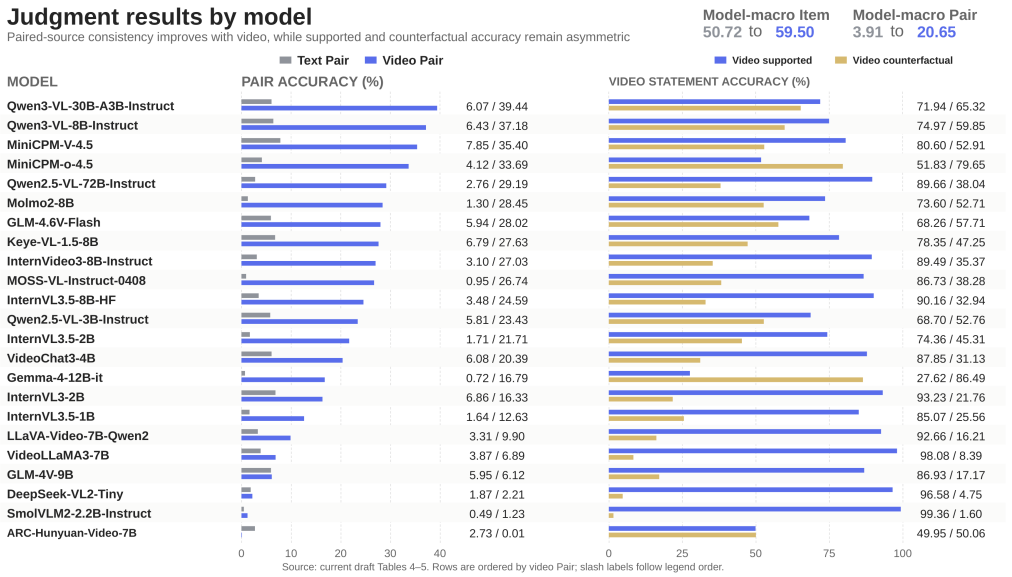
Table 4: Per-model MCQA accuracy (%) on the 2,023-item release. Grouped horizontal bars compare video and text-only accuracy for all 26 pipeline-independent checkpoints, ordered by video MCQA accuracy from highest to lowest; the dashed line marks 25% chance, and Δ is Video − Text. Dashes denote missing complete text-only runs.



Source: current draft Tables 2-3. Dashes denote missing complete text-only runs.

Table 4 reports the complete 26-model video roster in the appendix; the 24 rows with complete text-only runs define all matched text–video aggregates. Dashes denote missing text-only evaluations rather than zero accuracy. Checkpoint-year and scale definitions are documented in Appendix B.

Table 5: Per-model judgment results on 7,071 balanced source pairs. The left panel compares text-only and video Pair accuracy; the right panel separates video supported/true and counterfactual/false statement accuracy. Exact percentages are printed beside the bars, and models are ordered by video Pair accuracy.



Source: current draft Tables 4–5. Rows are ordered by video Pair; slash labels follow legend order.

Table 5 places all 23 per-model text/video Pair results alongside the video supported and counterfactual slices. It shows directly that near-50% text-only Item often averages opposite tendencies to accept or reject statements rather than consistent pair resolution.

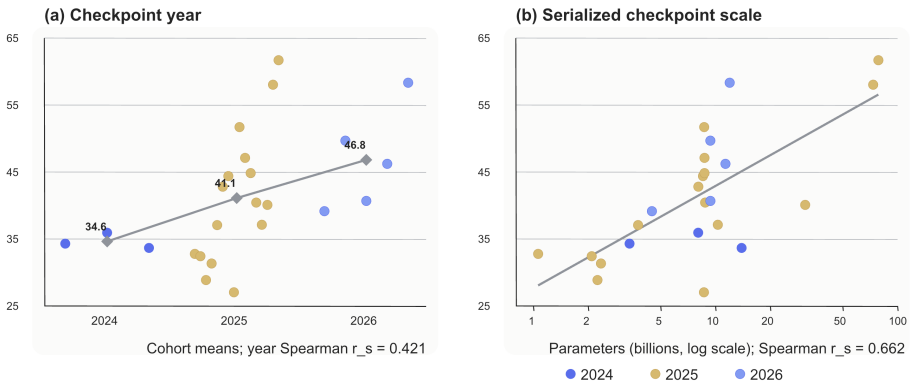


Figure 6: Text-only MCQA accuracy for the 24 matched pipeline-external models. Each point is one evaluated checkpoint. Left: individual results and equal-weight means by public-checkpoint repository year (34.64/41.11/46.81% for 2024/2025/2026; $n = 3/16/5$). Right: accuracy against the frozen Hugging Face safetensors total parameter count on a log scale. Spearman associations are 0.421 with year and 0.662 with scale. The roster is heterogeneous and non-random, and year, scale, architecture, and family are confounded; the trends are descriptive evidence that newer and larger systems recover more question–option signal, not causal effects of time or parameter count.

Figure 6 makes the capability trend explicit: cohort means rise across repository years, and larger serialized checkpoints tend to achieve higher text-only accuracy. This supports the concern that a filter frozen around a few earlier selectors can weaken as models improve, although it does not identify whether year, scale, model family, training data, or instruction following causes the association.

Figure 7 shows how one correct acceptance can coexist with failure to reject a minimally changed counterfactual. These selected outputs explain Pair, rather than estimate the frequency or cause of acceptance/rejection bias.

In Figure 8, MiniCPM-V-4.5 and VideoChat3-4B change from incorrect MCQA with repeated inputs $[7.5, 7.5]$ seconds to correct with $[7.5, 48.0]$, while retaining correct judgments. Reversal also succeeds for both; complete windows do not uniformly help every model. Appendix Figure 9 provides a separate three-state scoreboard trace. Neither case establishes single-window necessity over the bank.

B Evaluation-model metadata and supplemental context

Appendix Table 4 lists the 26 video-evaluated models that did not participate in benchmark construction, automated checking, metadata adjudication, or text-only selection. Results are percentages on the current 2,023-item filtered bank; each row is recomputed from existing predictions. The exclusion policy removes four pipeline participants from the original result archive. It does not claim independence of model family, training data, or upstream benchmark exposure. All 26 checkpoints have publicly downloadable weights; "open-weight describes weight availability, not a shared license or availability of all training data. The parenthetical year is the creation year of the evaluated public checkpoint repository; scale is the published/nameplate checkpoint scale (total/active parameters for named MoE checkpoints). The frozen metadata snapshot also retains each repository URL

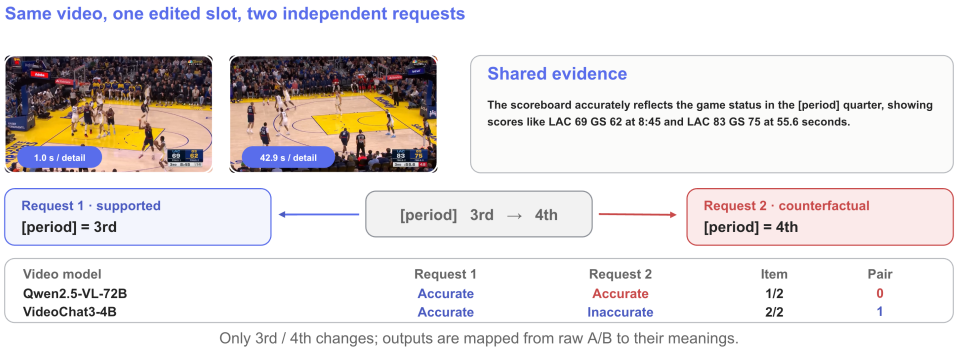


Figure 7: A source-linked judgment pair differs only in quarter (3rd/4th). The shared template branches into two independent requests; only the 3rd/4th slot changes. Existing video predictions are linked to Item and Pair scoring. Scoreboard details use inspection frames, not necessarily model inputs. Accepting both statements gives Item $= 1/2$ but Pair $= 0$.

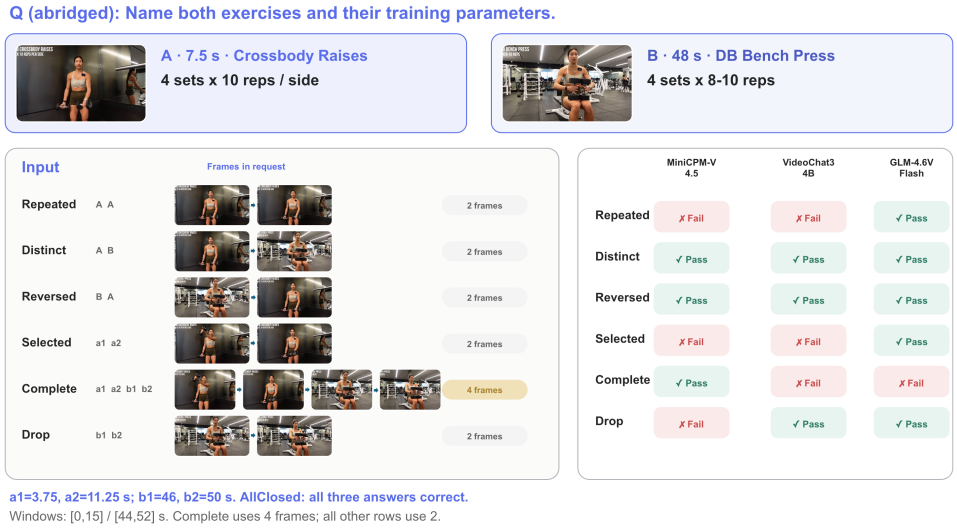


Figure 8: Complementary exercise evidence and existing AllClosed outcomes. Thumbnails reconstruct all six input sequences from logged timestamps. Reversals and full-window failures are retained. The figure shows three independent models; all 11 models for this source appear in Appendix Table 11, and aggregate evidence in Figure 5.

Table 6: Published no-video question-plus-option baselines on other video MCQA benchmarks. Values are exact primary-paper results but are not directly comparable across datasets, splits, training regimes, or models.

Benchmark	Split	Choices	Text-only system	Acc.	Chance
MovieQA (A5)	validation	5	fine-tuned RoBERTa QA-only	37.33	20
TVQA (A5)	validation	5	fine-tuned RoBERTa QA-only	48.91	20
NExT-GQA (NExT-QA-derived)	test	5	SeViLA language-only (BlindQA)	59.60	20

and its raw serialized parameter count. This operational policy avoids silently assigning a later paper date to an earlier public checkpoint. Across the 24 matched rows, the lower/middle/upper video-ranked thirds have 32.90/40.68/50.88% mean text-only accuracy. The 2024/2025/2026 repository-year cohorts contain 3/16/5 models and average 34.64/41.11/46.81% text-only. Both summaries are descriptive and roster-confounded; the per-model rows permit inspection without treating year or size as causal.

C External text-only context

Table 6 transcribes only settings that omit video while retaining the question and candidate answers. The MovieQA/TVQA values are validation results from Table 1 of Yang et al. (2020); the NExT-GQA value is the SeViLA BlindQA point in Figure 5(a) of Xiao et al. (2025). Different choice counts, splits, training, and model families preclude a direct ranking against our 41.49%.

D Evidence-diagnostic visual conditions

The current diagnostic analyzes 500 manually reviewed sources under the six conditions below. Here K is the number of frozen evidence windows. The display name "selected" avoids implying that the internal `best` IDs are accuracy oracles.

The fixed roster and each checkpoint's year/scale are reported in Table 7. All 54 shards completed without runtime-error rows; 32/82,500 new outputs were unparseable and scored incorrect.

Condition	Input	Frames
Repeated center	Selected center repeated	K
Ordered centers	One center per window	K
Reversed centers	Reverse of ordered centers	K
Selected window	Two samples within selected window	2
Complete selected windows	Two samples per selected window, ordered	$2K$
Drop selected window	Two samples in remaining windows	$2(K-1)$

The four reported contrasts are distinct versus repeated centers, ordered versus reversed centers, complete versus selected window, and complete versus drop selected. The first holds the K-frame budget fixed, the second preserves the exact frame multiset, and the final two alter the frame count. All were frozen before inference of the five new conditions and all are reported as exploratory. Extremely short windows can repeat their starting frame. Decoding may backtrack when a requested frame cannot be read; the completed manual review therefore used the decoded inputs and source videos rather than relying only on timestamp manifests.

E Traceable multi-frame example

Figure 9 makes the three scoreboard states in the designated answer directly inspectable.

Table 7: Per-model paired effects (percentage points) in the manually reviewed 500-source evidence diagnostic. Models are ordered by the mean of the 12 displayed effects, from highest to lowest. Columns group MCQA, Pair, and AllClosed results for C1–C4; green favors the first condition, red favors the second, and gray is near zero. C1 compares distinct with repeated centers at equal K; C2 compares ordered with reversed centers using the same multiset; C3 compares complete with selected-window input; C4 compares complete with drop-selected input.

Per-model effects in the 500-source evidence diagnostic

Green favors the first condition, red favors the second and gray marks zero or near-zero effects

positive | near zero | negative

MODEL	MCQA				Pair				AllClosed			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
MiniCPM-V-4.5	5.2	2.0	5.6	6.4	6.4	0.8	8.8	14.4	5.6	2.0	9.2	12.8
MiniCPM-o-4.5	4.2	1.2	1.8	3.8	7.0	1.0	11.0	13.4	4.8	0.6	7.2	10.6
MOSS-VL-Instruct-0408	1.2	-0.6	3.8	6.0	8.4	0.2	8.0	10.2	4.2	0.8	6.4	7.2
Gemma-4-12B-it	3.6	1.6	3.2	4.0	1.4	0.4	7.4	5.2	1.6	0.2	6.0	4.2
Keye-VL-1.5-8B	1.8	0.8	3.6	3.4	1.8	-1.0	4.2	8.0	0.6	-0.8	2.8	5.4
LLaVA-Video-7B-Qwen2	3.4	2.2	3.4	4.0	1.0	1.0	1.2	2.8	0.8	1.0	1.0	3.0
GLM-4.6V-Flash	-0.6	-0.8	2.0	1.0	0.4	-0.8	4.8	7.8	1.6	0.4	4.2	4.4
InternVideo3-8B-Instruct	-3.2	-1.6	3.0	1.8	0.6	0.2	3.4	5.4	0.2	0.2	4.0	3.8
VideoChat3-4B	2.6	0.2	0.4	2.0	3.2	0.6	0.2	3.6	2.2	-1.0	-0.4	2.2
SmolVLM2-2.2B-Instruct	1.8	1.6	4.4	3.0	0.0	-0.4	-0.2	0.0	0.2	0.0	0.0	0.0
VideoLLaMA3-7B	1.2	-3.4	2.2	4.0	-2.4	-0.8	-4.4	-1.6	-1.0	-0.8	-1.8	0.2

C1 distinct minus repeated centers (equal K) C2 ordered minus reversed centers (same multiset) C3 complete minus selected window C4 complete minus drop selected

Values are paired effects in percentage points. Every cell reproduces Table 8.

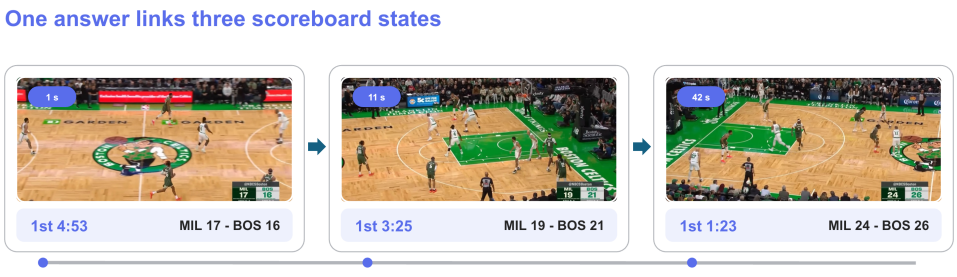


Figure 9: Three scoreboard states linked by one human-reviewed answer (source MCQ-008e4217b65702051990). Insets enlarge the original scoreboards; timestamps denote inspection points within the video. Review used the source clip in addition to these displayed frames.

F Completed human quality control

The audit unit is a source question with its full context, evidence windows, answer slots, and format variants. Human reviewers assess whether the question is answerable, whether the gold answer

Table 8: Completed human quality-control scope. The review corrects defective content but is not used as a human-performance baseline or a certificate of minimal evidence.

Scope	Completed checks and actions
Generated QA records	Inspected question clarity, answer support and uniqueness, option validity, and appropriate wording against source evidence.
Linked formats	Checked semantic consistency across available MCQA, supported-statement, and counterfactual-statement variants.
Quality repair	Corrected unsupported, inaccurate, ambiguous, misleading, or otherwise inappropriate expressions; rechecked affected variants and excluded records without a reliable repair.
500-source diagnostic	Manually reviewed all 500 selected sources, their closed-form variants, source videos, and supplied evidence before treating them as the experimental set.
Interpretation boundary	Treated evidence windows and ability labels as structured support, not as automatic proof that no smaller evidence subset can answer the question.

Table 9: Intended ability targets in the current MCQA bank.

Code	Ability	Items
F0	Natural direct QA	9
F1	Content/process integration	211
F2	State/information boundary	432
F3	Four-stage timeline	372
F4	Text-carrier spatial binding	272
F5	User-need evidence chain	511
F6	Confirmed/unconfirmed boundary	216

is supported and unique, whether each distractor or false statement changes the intended slot, and whether all formats preserve the intended semantics. They correct unsupported, inaccurate, ambiguous, misleading, or otherwise inappropriate wording and recheck linked variants after repair. Records without a reliable evidence-grounded repair are excluded.

The 500-source experimental set was reviewed in full using the source videos, decoded inputs, and recorded evidence windows. The completed review covers semantic and presentation quality but remains distinct from a proof of minimal evidence necessity: an answer can be valid and well supported even when a different frame subset would also suffice. The illustrative example in Figure 9 makes one reviewed trace visible but is not an estimate of the frequency of multi-frame necessity.

G Capability distribution

Capability labels record the intended generation target. Human review checks semantic validity and evidence support, but the labels are not certificates of minimal evidence requirements. The current MCQA bank contains 1,310 cross-frame-pool and 713 simple-pool items. Unequal capability and scene sizes should accompany any subgroup leaderboard.

H Related-work comparison protocol

The comparisons in Table 10 concern documented design choices. We have not evaluated all cited systems under common checkpoints, frame budgets, or scoring, so the table cannot establish a performance advantage. Source counts also differ in meaning: evaluation items in our release include format variants, whereas TempCompass reports task instructions and ViTexQA reports QA pairs (Liu et al., 2024; Guo et al., 2026). These denominators must be preserved in any scale comparison.

Table 10: Direct Video Text QA neighbors. The table records documented designs rather than claiming that any individual feature is exclusive to us.

Work	Construction	Difficulty control	Scenario/task scope	Quality audit
M4-ViteVQA (Zhao et al., 2022)	Human authored and verified	Easy/hard; hard uses at least two frames	9 categories; held-out-category test	Manual QA verification
EgoTextVQA (Zhou et al., 2025)	GPT-4o initial QA, then human revision	Multi-frame and temporal-grounding analyses	Driving and housekeeping activities	Human review and refinement
MME-VideoOCR (Shi et al., 2025)	Manual annotation and expert verification	Includes cross-frame text understanding	25 tasks and 44 scenarios	Expert verification
VidText (Yang et al., 2025)	Dense text annotation and human QA design	Paired perception/reasoning tasks	27 categories; 8 tasks; 3 granularities	Human-authored QA
ViTexQA (Guo et al., 2026)	Trained human annotators	All QA checked for cross-frame fusion	30 categories in 6 domains	Human answerability checks and CoT
This work	Evidence-structured LLM-first workflow with human revision	Ability routing plus text/evidence controls	10 primary and 34 secondary scenarios	Automated/program checks and completed human quality review

I VIDEO COVERAGE AND DURATION BY CATEGORY

The counting unit is a unique video path in the union of evaluated MCQA and judgment sources, excluding unevaluated open-answer items. Scene labels and F targets are joined by source ID. Six videos have source records assigned to more than one domain; 2,078 videos support questions from multiple F categories. Each video is counted once within each associated category, so category totals need not equal the 2,629-video union. Durations are read from video-stream timestamps (container duration is the fallback), and refer to evaluated clips rather than unsliced parent videos. F0 has only 11 associated videos. These distributions describe coverage and sampling; they do not establish balanced domains or the necessity of a particular reasoning ability.

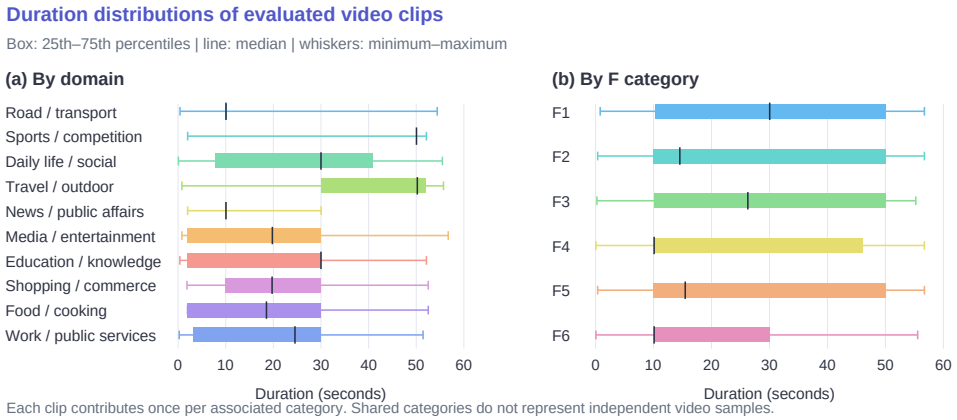


Figure 10: Clip duration by domain and intended F category (F1–F6 shown). Boxes span the 25th–75th percentiles, central lines mark medians, and whiskers show minima and maxima. Boxes can collapse when many clips have the same duration. Videos are deduplicated within categories; overlapping categories are not independent samples.

J EVALUATION IMPLEMENTATION DETAILS

Diagnostic decoding and prompts. The 500-source diagnostic reads the ordered timestamps in its condition manifest rather than uniformly resampling the video. Before checkpoint-specific preprocessing, decoded frames are resized with aspect ratio preserved to at most 200,704 pixels per frame. Checkpoint processors may subsequently resize or tile these images, so this cap is not a matched visual-token budget. The closed-form runner uses the larger of 16 new tokens and the checkpoint-

specific output allowance. Each judgment statement is a separate request with its recorded Accurate/Inaccurate options. The shared prompt is reproduced below; its A–D instruction is retained even when a judgment provides only A/B options.

> Watch the entire video carefully and answer the multiple-choice question.
> Question: [question text]
> Options:
> [label]. [option text; repeated for each supplied option]
> Return exactly one uppercase option letter (A, B, C, or D). Do not output any other text.

Parsing and aggregation. The diagnostic parser accepts a single option letter with supported punctuation/wrappers, two to four repetitions of the same letter, and specified answer-marker patterns. A final fallback accepts only short outputs (at most 32 characters) whose standalone uppercase option letters all agree. Unparseable responses score zero; parsed letters are compared with the stored gold option. Pair and AllClosed join the independently scored variants by source identifier, and the reported intervals resample parent-video clusters. These implementation details describe the evidence diagnostic; the release-level leaderboard uses its recorded model-specific video pipelines and is not compute-matched.

K Qualitative cases and source traces

The cases in the main text and appendix are selected illustrations, not a random audit. The following question and designated-answer fields reproduce the human-reviewed English evaluation records; main-text shortenings are for display only. The boundary example belongs to the open/judgment source bank and is not in the 2,023-item MCQA release. Displayed timestamps refer to source-clip inspection points. They are not all the frames used by a model. For the game example, the unconfirmed claim was checked against the full clip during manual review rather than inferred from two screenshots.

Completed course states (F2; MCQ-772367542437e8a00976). Q: Based on the text projected on the desk in the video, what are the course content and level completed by the child?

Designated answer: The child completed 'HTML Basics Level 3' and 'Calculus - Integration Level 1'.

Original VidText question: "Please provide all the text content from the video." Original answer excerpts: "HTML Basics Level 3" and "Calculus-Integration Level 1"; intervening transcription omitted. The same source video is used in both tasks.

Text, roles, and objects (F1; MCQ-de845291c345cef8e452). Q: Based on the video, which of the following most accurately describes the combination of text labels and character actions that drive the progression of the Consumer-to-Consumer (C2C) and Consumer-to-Business (C2B) transaction models?

Designated answer: The C2C model is presented through the text labels 'Consumer to Consumer' and 'C2C', accompanied by the action of two characters passing a package; the C2B model is advanced through the text labels 'Consumer to Business' and 'C2B', accompanied by the action of the character in gray passing a mobile phone to the character in a suit.

Confirmed and unconfirmed information (F6; MCQ-9c422fb7ac529f2c4bae). Q: Based on the video, which of the following most accurately describes what the game streamer can confirm and why the effect of the skill 'PYROCLASM' on defensive towers cannot be confirmed?

Designated answer: Confirmed: The text 'ABILITIES', 'ABLAZE: SPREAD TO MORE TARGETS', 'PYROCLASM', and 'SET THE WORLD ABLAZE' appears in sync with the visuals; Unconfirmed: Due to the absence of textual explanations such as 'effect on defensive towers' or 'skill range limitations'.

Complementary exercise evidence (F2; MCQ-be8f23aeff2d20ce650d). **Q:** Based on the video, which of the following most accurately describes the names and training parameters of the two main exercises demonstrated?

Designated answer: DB Crossbody Raises: 4 sets of 10 reps per side; DB Bench Press: 4 sets of 8 to 10 reps.

Actual input timestamps (seconds): repeated [7.5, 7.5]; distinct [7.5, 48.0]; reversed [48.0, 7.5]; selected [3.75, 11.25]; complete [3.75, 11.25, 46.0, 50.0]; drop selected [46.0, 50.0].

Supported/counterfactual pair (F5; MCQ-00015c1af1ab9d4d7ce0). **Supported:** The scoreboard accurately reflects the game status in the 3rd quarter, showing scores like LAC 69 GS 62 at 8:45 and LAC 83 GS 75 at 55.6 seconds.

Counterfactual: The scoreboard accurately reflects the game status in the 4th quarter, showing scores like LAC 69 GS 62 at 8:45 and LAC 83 GS 75 at 55.6 seconds.

Table 11: AllClosed outcomes for the selected fitness source across all 11 independent models. Models are ordered by total passes across the six displayed conditions, from highest to lowest; ties retain the source order. A pass requires MCQA and both judgment statements to be correct. Columns show repeated-center, distinct-center, reversed-center, selected-window, complete-window, and drop-selected inputs. This is one selected source, not an aggregate.

AllClosed outcomes for the selected fitness source

A pass requires correct MCQA and both judgment statements; this matrix describes one source, not an aggregate

✓ Pass × Fail

	CENTER-FRAME INPUTS			WINDOW INPUTS		
MODEL	R Repeated	D Distinct	V Reversed	S Selected	C Complete	X Drop selected
InternVideo3-8B-Instruct	✓	✓	✓	✓	✓	×
GLM-4.6V-Flash	✓	✓	✓	✓	×	✓
Keye-VL-1.5-8B	✓	✓	×	✓	✓	×
MOSS-VL-Instruct-0408	✓	×	✓	✓	×	✓
MiniCPM-V-4.5	×	✓	✓	×	✓	×
VideoChat3-4B	×	✓	✓	×	×	✓
MiniCPM-o-4.5	✓	✓	×	×	✓	×
Gemma-4-12B-it	×	×	✓	×	✓	×
SmolVLM2-2.2B-Instruct	×	×	×	×	×	×
LLaVA-Video-7B-Qwen2	×	×	×	×	×	×
VideoLLaMA3-7B	×	×	×	×	×	×

R repeated center; D distinct centers; V reversed centers; S selected window; C complete windows; X drop selected.

Source: current draft Table 12. The pattern must not be interpreted as an aggregate ranking across conditions.

K.1 Illustrative response formats.

Figures 12 and 13 show complete released items alongside source-frame evidence. The MCQA examples distinguish completed course levels and indoor/outdoor labels; the judgment examples alter exercise repetitions or the scoreboard quarter. Questions, options, statements, and gold labels retain their released wording and order. These selected illustrations make the response formats inspectable rather than estimate bank-wide quality.

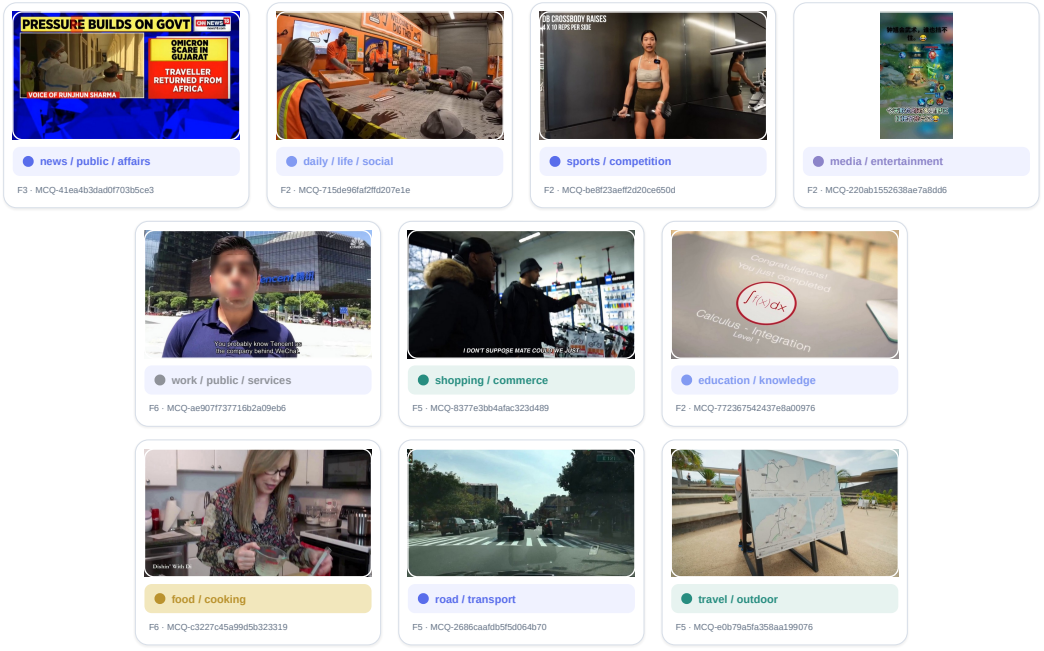


Figure 11: One source screenshot per primary scene, with traceable source IDs and recorded F labels. Frames preserve the source aspect ratio; the gallery illustrates scene variety rather than complete question–answer examples for every F category. Selected cases do not establish balanced coverage or estimate bank-wide quality.

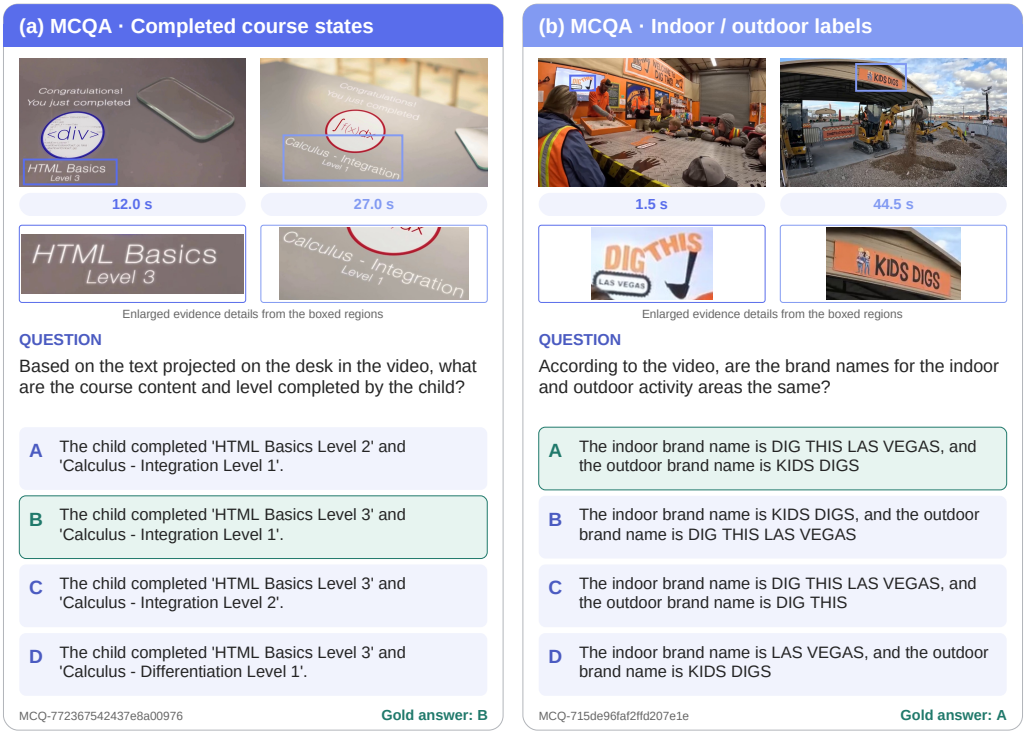


Figure 12: MCQA examples with key video frames: (a) completed course content and levels; (b) indoor and outdoor activity-area labels. All four options are shown, with the designated answer in green. Insets enlarge the boxed evidence regions; timestamps denote inspection points within each source clip.

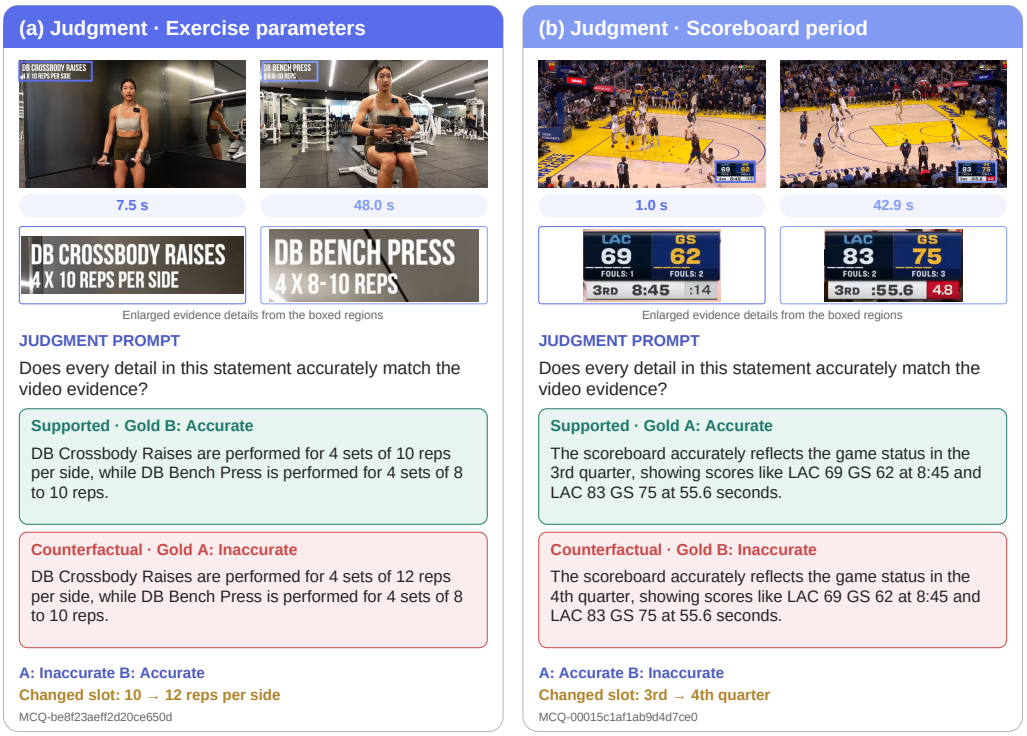


Figure 13: Supported/counterfactual judgment examples: (a) changing repetitions per side; (b) changing the scoreboard quarter. Each pair is displayed together for illustration, but its statements are evaluated independently. Statement wording, Accurate/Inaccurate option order, and gold letters are preserved. Insets enlarge source-frame details; clip-relative timestamps are inspection points, not necessarily model input frames.