What Breaks Integrity of VideoQA Answers After Evidence Retrieval?

Anonymous ACL submission

Abstract

High VideoQA accuracy alone does not show that a correct answer depends on the evidence selected after retrieval. We audit selected evidence integrity with EVIDENCEBREAK: retrieval is fixed, adjacent retrieved blocks are grouped into temporal segments, important segments are found through leave one segment out option scoring, and controlled defects are applied only to selected evidence. We introduce Core Dependence Gap (CDG), which compares answer flips caused by editing core segments with flips caused by matched edits outside the core set. The answerer receives keyframes with dataset specific text, namely generated captions for NExT-QA and aligned subtitles for Video-MME, and we count only predictions that are correct before corruption but wrong after corruption. NExT-QA shows limited fragility, while Video-MME shows stronger selected evidence dependence: core omission and mis-stitching flip 24.0% and 20.1% of valid clean correct examples. Matched controls yield CDGs of 11.03 and 7.39 points for omission and mis-stitching, and a full Video-MME rerun with Qwen3-VL-4B preserves the same defect ranking.

1 Introduction

Retrieval based VideoQA systems usually answer from a small set of selected moments rather than the full video. This improves tractability but makes it unclear whether a correct answer actually depends on the selected evidence. We study this dependency after retrieval is fixed and call it selected evidence integrity: when one selected segment is altered, does a clean correct prediction remain correct or break?

Prior grounded VideoQA work has shown that a correct prediction alone does not guarantee grounding in human annotated evidence (Xiao et al., 2024). We ask a complementary question: if an important selected segment is removed or replaced, does the answer still hold? When the edit turns a clean correct prediction into a wrong one under the same system configuration, the failure indicates dependence on the edited evidence.

To this end, EVIDENCEBREAK fixes retrieval, groups adjacent retrieved blocks into segments, selects core segments by the drop in option likelihood gap after segment removal, and applies four controlled evidence defects. We use NExT-GQA only for temporal grounding calibration, NExT-QA for question type analysis, and Video-MME for duration analysis; human grounding labels are never used for retrieval, core selection, or answer scoring.

Contributions. We make four contributions. First, we define selected evidence integrity as an evaluation target for VideoQA systems that answer after retrieval. Second, we propose EVIDENCEBREAK, a counterfactual audit that keeps retrieval fixed while editing selected segments. Third, we introduce Core Dependence Gap (CDG), a paired metric for comparing edits to core segments with matched edits outside the core set. Fourth, we show that Video-MME exhibits strong dependence on core segments under omission and mis-stitching, while NExT-QA shows limited flips among valid recipes.

2 Related Work

Grounded VideoQA and VideoQA with retrieval. NExT-QA introduced causal and temporal VideoQA (Xiao et al., 2021), NExT-GQA adds temporal grounding annotations (Xiao et al., 2024), and Video-MME evaluates multimodal LLMs across video durations (Fu et al., 2025). Retrieval augmented video systems retrieve clips to reduce long video inference cost (Gia et al., 2025; Jeong et al., 2025). Rather than improve retrieval, we audit whether selected evidence continues to support clean correct answers after controlled edits.

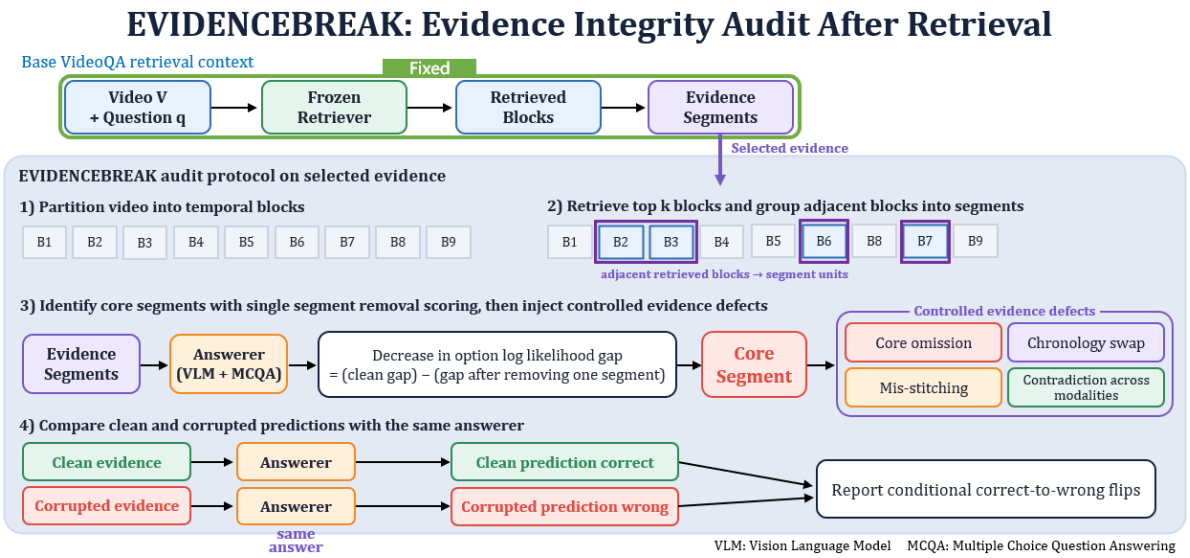

Figure 1: EVIDENCEBREAK audits evidence integrity after retrieval by fixing retrieval, grouping adjacent retrieved blocks into segments, selecting core segments, injecting controlled evidence defects, and measuring paired answer flips caused by corruption.

Evidence alignment and robustness. Grounding work asks whether predictions are supported by human annotated evidence, while robustness studies perturb pixels, text, or full inputs. Recent diagnostics also examine evidence misalignment and induced corruption in long video models (Qiu et al., 2026; Yang et al., 2026). EVIDENCEBREAK differs by fixing retrieval and applying paired edits to selected segments from the same video.

3 Method

Audit target. We audit selected evidence integrity after retrieval rather than retrieval quality. Given a clean evidence pack, we edit one selected segment and compare clean and corrupted predictions under the same answerer, prompt, frame budget, and scoring rule. A clean prediction is the answer before any defect is applied, and a corrupted prediction is the answer after the selected segment is edited. A core segment is a selected segment whose removal causes one of the largest drops in the gap between the top two option log-likelihoods. We count a flip only when a clean prediction is correct and its corrupted counterpart is wrong.

Clean evidence construction. Given a video V and question q, we divide V into temporal blocks and retrieve the top k blocks with a frozen SigLIP2 image and text retriever (Tschannen et al., 2025). Blocks are ranked by dot product similarity between question and block embeddings; we set $k = 4$ to keep multiple moments within the VLM input while leaving unused blocks for perturbation. For each retrieved block, we retain up to two frames selected by question similarity as keyframes. The VLM receives these keyframes with dataset specific text: for NExT-QA, question agnostic block captions generated once from two sampled frames per block with Qwen2.5-VL-7B-Instruct; for Video-MME, subtitles aligned to the retrieved block. Visual clue text available in Video-MME is not used for inference.

Segment grouping and core search. Retrieved blocks are sorted by time and merged into a segment when their indices differ by at most one. Retrieval keeps the block budget fixed, while the audit edits segments. For each clean correct example, we remove each segment, recompute option log-likelihoods, and measure the drop in the gap between the largest and second largest values. The top 25% of segments by this drop are core segments. This marks model reliance, not human rationales; NExT-GQA labels are used only to check overlap with human temporal evidence.

Evidence defects. We use four defects. **Core omission** replaces a core segment with a low relevance segment from the same video while preserving the number of blocks and images. **Chronology swap** changes the order of selected segments and their displayed timestamps. **mis-stitching** replaces a core segment with a temporally distant segment from the same video that is visually plausible but irrelevant to the question. Replacement candidates

are drawn from segments outside the retrieved evidence and outside the core set. **Contradiction across modalities** keeps Video-MME keyframes fixed but rewrites local subtitle evidence. Because NExT-QA has no subtitles, we report a caption diagnostic that edits generated block captions while keeping keyframes fixed. These synthetic edits are limited to simple action, count, or spatial changes and are not treated as human validated defects.

Scoring. For each multiple choice question, Qwen2.5-VL-7B-Instruct (Qwen Team, 2025a) scores each option by conditional log-likelihood under a fixed prompt. NExT-QA and NExT-GQA use an 8,192 token budget, and Video-MME uses 10,240 tokens because it includes subtitles; increasing the Video-MME budget from 8,192 to 10,240 did not change clean predictions or selected core segments.

4 Experimental Setup

Calibration. We compared 8s, 12s, and 16s blocks on NExT-GQA and use 16s in all main experiments because it gives the best clean accuracy: 73.17%, compared with 73.02% for 12s and 71.62% for 8s. Among NExT-GQA examples that are correct with clean evidence, selected core segments overlap human temporal grounding in 92.2% of cases; the retrieved segment with the highest human overlap is included in the core set for 91.1% of examples. Because mean IoU and IoP remain moderate at 23.2% and 24.4%, we treat core segments as model reliance evidence rather than human rationales.

Evaluation. Clean accuracy is measured before any evidence defect is applied. Defect results are restricted to clean correct examples for which the corresponding edit can be constructed. We report coverage, conditional flip, full split flip, Stable Q+Opt, and 95% Wilson score intervals (Wilson, 1927). Corrupted accuracy is the complement of conditional flip on the clean correct valid subset.

Controls. Baselines compare full evidence with text only, caption or subtitle only, reduced frame, reduced block, random replacement, and replacement outside the core set. For Video-MME omission and mis-stitching, we also evaluate length matched outside core controls and summarize the paired contrast as *Core Dependence Gap*, $\mathrm{CDG} = \mathrm{Flip}_{core} - \mathrm{Flip}_{outside}$. CDG intervals use paired bootstrap (Koehn, 2004), and significance is tested with an exact directional McNemar test (McNemar, 1947).

Dataset / group	Clean	Q+Opt	Opt only
NExT-QA overall	75.70	45.51	22.17
causal	76.94	48.07	21.90
temporal	70.17	43.92	21.43
descriptive	83.46	39.59	24.82
Video-MME overall	48.19	31.19	27.67
short	55.78	31.56	26.11
medium	46.11	31.11	27.67
long	42.67	30.89	29.22

Table 1: Clean accuracy and language prior accuracy. All entries are percentages. Q+Opt removes all video evidence and keeps only question and options.

Replication. We also rerun Video-MME with Qwen3-VL-4B (Qwen Team, 2025b), regenerating core search and defect recipes from that model's own clean correct predictions while keeping retrieval and evidence construction fixed.

5 Results

Clean performance and language priors. To validate the effectiveness of visual inputs, we test whether the answerer can solve the same questions without selected video evidence. Motivated by a preliminary observation that the model often chooses plausible answers from question and answer choices alone, we include Q+Choices and Choices only as diagnostic probes. Q+Choices keeps only the question and answer choices, while Choices only removes the question as well. Table 1 shows stronger priors in NExT-QA than in Video-MME: Q+Choices reaches 45.51% on NExT-QA but 31.19% on Video-MME.

Evidence defects. Table 2 summarizes the main defect results. NExT-QA yields low conditional flips, ranging from 5.08% to 6.62%, and very small full split flips, so we avoid using it as evidence for broad fragility. Video-MME shows stronger dependence on selected evidence: core omission and mis-stitching flip 23.96% and 20.10% of valid clean correct examples, while contradiction across modalities remains weak. A language prior overlap check further shows that roughly half of stable corrupted predictions are also correct with only question and options, so stability after corruption should not be treated as faithful evidence preservation.

Matched segment controls and replication. Table 3 shows that Video-MME flips are target dependent. Over 934 matched recipes, omission and

Dataset	Defect	Cov. (%)	n_{valid}	Cond. Flip (%) [95% CI]	Full Flip (%)	Stable Q+Opt (%)
NExT-QA	Chronology swap	13.8	492	5.08 [3.47, 7.39]	0.53	54.8
NExT-QA	Core omission	7.7	272	5.88 [3.65, 9.34]	0.34	52.3
NExT-QA	mis-stitching	7.7	272	6.62 [4.23, 10.22]	0.38	52.0
NExT-QA	Caption–visual diag.	98.8	3514	0.46 [0.28, 0.74]	0.34	–
Video-MME	Chronology swap	82.2	1069	11.51 [9.73, 13.56]	4.56	49.7
Video-MME	Core omission	75.7	985	23.96 [21.40, 26.72]	8.74	52.7
Video-MME	mis-stitching	75.7	985	20.10 [17.72, 22.72]	7.33	52.4
Video-MME	Contradiction across modalities	55.2	718	1.53 [0.86, 2.72]	0.41	46.3

Table 2: Defect results. Cov. is the percentage of clean correct examples for which the edit can be constructed, and n_{valid} is the resulting count. Cond. Flip is the clean to wrong flip rate among valid recipes with 95% Wilson intervals. Full Flip normalizes the same flips by the full split size. Stable Q+Opt is, among examples that remain correct after corruption, the percentage also correct with only the question and options.

Check	Chron.	Omit.	Stitch	Contr.
Qwen2.5 all	11.51	23.96	20.10	1.53
Qwen3-VL-4B	9.66	25.52	20.80	2.20
Core matched	–	23.34	19.81	–
Outside core	–	12.31	12.42	–
CDG	–	+11.03	+7.39	–

Table 3: Video-MME conditional flips. Qwen2.5 all and Qwen3-VL-4B use full valid recipe sets. Core matched and Outside core use the same 934 paired recipes; CDG is core minus outside core.

mis-stitching yield CDGs of 11.03 points (95% CI: 8.24–13.81) and 7.39 points (4.60–10.17), with significant McNemar tests ($p = 9.03 \times 10^{-15}$ and $p = 1.84 \times 10^{-7}$). The Qwen3-VL-4B rerun preserves the same defect order: core omission, mis-stitching, chronology swap, and contradiction across modalities. The matched design is therefore the main control for the Video-MME claim. Core and outside core edits use the same answerer, retrieval output, recipe pool, and example set, so the contrast isolates where the edited segment sits rather than whether an edit was possible.

Subgroups and baselines. Table 4 shows that NExT-QA flips concentrate in temporal and causal questions, Video-MME omission is larger for medium and long videos, and reduced frame or random replacement baselines stay close to clean accuracy. Table 3 therefore gives the direct control for target location. By contrast, subgroup and baseline checks serve as supporting context: they identify where flips concentrate and show that simple evidence reduction or random replacement does not reproduce the core edit pattern.

Dataset	Group	Cond. Flip (%)
NExT-QA	temporal	7.61
NExT-QA	causal	7.25
NExT-QA	descriptive	2.38
Video-MME	short	17.20
Video-MME	medium	25.78
Video-MME	long	25.26

Condition	NExT-QA	Video-MME
Clean evidence	75.70	48.19
Text only	65.61	40.26
Caption/subtitle only	65.31	40.48
Reduced frame	74.47	48.11
Reduced block	73.36	45.26
Random replacement	75.45	48.26
Outside core replacement	75.83	49.00

Table 4: Subgroup flips and baseline checks. Upper block reports conditional flip rates: NExT-QA uses mis-stitching by question type, and Video-MME uses core omission by duration. Lower block reports full split baseline accuracy.

6 Discussion and Conclusion

Audit interpretation. EVIDENCEBREAK should be read as an audit of selected evidence after retrieval, not as a general robustness benchmark. NExT-QA suggests that short caption based evidence and language priors can hide dependence on core segments, whereas Video-MME provides stronger evidence of selected evidence dependence. CDG and Qwen3 replication further show that the main effect is tied to core segments rather than to peripheral evidence or a single answerer.

Stable predictions. Stable corrupted answers should not be read as faithful robustness, because many are also correct with only question and option text. Evidence preservation therefore requires accuracy plus counterfactual controls that test whether answers change when selected evidence is damaged.

Limitations

This study uses one retriever and one main retrieval budget; Qwen3-VL-4B checks answerer specificity but not retriever generality. Core segments indicate model reliance, not human rationales, and NExT-GQA shows frequent grounding overlap but moderate localization. Defect estimates apply only to clean correct examples with valid recipes. misstitching is a same video distractor test rather than a human validated corruption, and generated captions may couple the captioner with the answerer. The weak contradiction result leaves modality reliance unresolved; future work should perturb visual frames, subtitles, metadata, and option priors separately.

Ethical Considerations

We use public benchmarks and publicly available model checkpoints for research evaluation. We do not collect personal data, redistribute raw videos, annotations, subtitles, or model weights, identify individuals, infer sensitive attributes, or report example level personal information. All results are reported as aggregate statistics.

References

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

Bao Tran Gia, Khiem Le, Tien Do, Tien-Dung Mai, Thanh Duc Ngo, Duy-Dinh Le, and Shin'ichi Satoh. 2025. VRAG: Retrieval-augmented video question answering for long-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3728–3737.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. 2025. VideoRAG: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21278–21298, Vienna, Austria. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Chenhao Qiu, Yechao Zhang, Xin Luo, Shien Song, and Xusheng Liu. 2026. VideoSEAL: Mitigating evidence misalignment in agentic long video understanding by decoupling answer authority. *arXiv preprint arXiv:2605.12571*.

Qwen Team. 2025a. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Qwen Team. 2025b. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Edwin Bidwell Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13204–13214.

Junqi Yang, Yuecong Min, Jie Zhang, Shiguang Shan, and Xilin Chen. 2026. INFACT: A diagnostic benchmark for induced faithfulness and factuality hallucinations in video-LLMs. *arXiv preprint arXiv:2603.11481*.