SEEING THE EVIDENCE IS NOT ENOUGH: REMOVAL TARGETS IN HARMFUL VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Harmful video understanding increasingly pairs video-level judgments with temporal annotations of where harmful content occurs, and removing the annotated content offers a natural test of whether a model relies on it. Harmful videos make this test fragile: a slur, threat, or hateful message can recur in speech, on-screen text, and later scenes, so an annotation can mark valid evidence while harmful support persists elsewhere in the model's observation. We distinguish the *intervention target*, the information a deletion withholds, from the *claim target*, the support to which the resulting prediction change is interpreted, and call their correspondence *removal fidelity*. We study it with HARMFULFAITH on three harmful-video datasets, four models, and sampled-frame and ASR/OCR text observations. Removing the annotated content changes far fewer `HARMFUL` predictions than retaining it preserves, and a clear harmful cue survives annotation-exact removal in 43–54% of audited observations. Source annotations cover a median 69.2% of independently audited text support and fully cover it in only 18.2% of cases with audited support; under the same 16-frame construction, visual harmful support is fully covered in 6.0% of such cases, versus 38.2% for Charades-STA event support. This residual support matters: source removal changes 13.3–19.5% of text predictions when a clear residual cue remains, versus 95.7–100% when none does, and the gap to audited-support removal widens as coverage falls. Under the same budget, removing audited support changes 93.8% and 88.9% of DeepSeek and GPT predictions, against 34.2% and 31.7% for matched random deletion. A weak response to removing annotated harmful content therefore need not mean that a model ignores harmful content: deletion-based conclusions about harmful-video models should be conditioned on what support the intervention actually withholds.

1 INTRODUCTION

Harmful video understanding (Jo & Wojcieszak, 2025) asks whether a video contains content that meets a task-specific definition of harmfulness[1]. Unlike a static image, the information behind such a judgment may appear only briefly or be spread across visual content, speech, and overlaid text. Recent tasks therefore go beyond video-level classification to localize when harmful content occurs (Wang et al., 2025; Chen et al., 2025), characterize harmful elements, or produce supporting rationales (Lu et al., 2026; Wu et al., 2026), and Large Multimodal Models (LMMs) are increasingly used to reason over these heterogeneous signals (OpenAI, 2023; Gemini Team, 2023; Maaz et al., 2024; Lin et al., 2024). Temporal annotations thereby acquire a second role. Once a dataset marks where the harmful content is, removing that interval offers a natural test of the model: if the prediction survives, the model appears insensitive to the evidence; if it changes, the model appears to depend on it. The test is attractive because it needs no access to model internals: the annotation fixes what to remove, and the resulting change in prediction measures the model's response.

This test quietly assumes that the annotated interval contains the support for the judgment. Annotations are rarely collected to satisfy that assumption. An annotator marking a hateful segment identifies content that justifies the label, not every moment that would justify it: an insult may be repeated later, an on-screen caption may restate what is spoken, or the surrounding context may make the intent unmistakable. Figure 1 shows such a case. The annotated intervals, which we call

[1] **Disclaimer**: This paper contains content that may be disturbing to some readers.

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

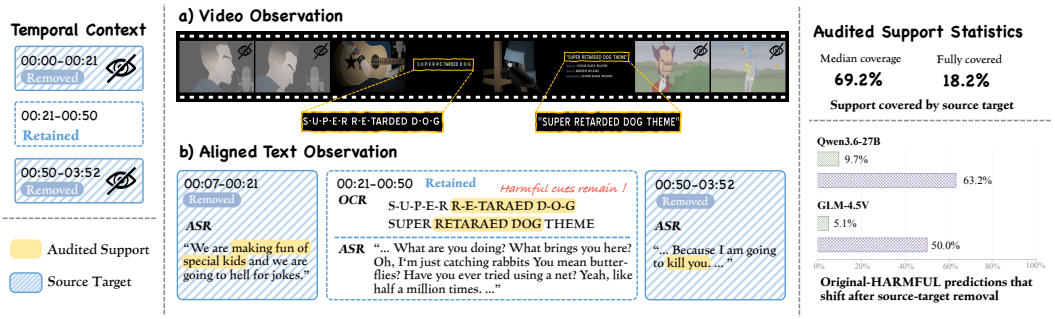


Figure 1: **Removing the annotated target can leave other support available.** Removing the annotated source target (blue hatching) leaves a middle segment whose overlaid text remains judgment-relevant; audited support (yellow) is identified by annotators independently of the annotation. Right: in the text audit, source targets cover a median 69.2% of audited support and fully cover it in only 18.2% of cases with audited support; after visual source removal, predictions shift far less often when a clear residual cue remains (9.7% and 5.1%) than when none does (63.2% and 50.0%).

the *source target*, contain valid evidence, yet the segment left between them still carries judgment-relevant text. Removing the source target follows the annotation exactly and still leaves *support*, information relevant to the same judgment, in the model's observation. A weak response to that removal is then ambiguous: the model may be insensitive to the harmful content, or it may simply find the same content elsewhere, and the removal alone cannot distinguish the two explanations.

Interpretability research has long recognized related issues. Information sufficient to sustain a prediction need not be the information whose removal changes it (DeYoung et al., 2020; Jacovi & Goldberg, 2020), human rationales can be incomplete, and redundant cues can blunt deletion-based signals (Carton et al., 2020; Madsen et al., 2022). We study a temporal setting in which this distinction becomes an evaluation problem: an annotation collected as valid evidence is later reused as a deletion target and read as though it exhausts the support relevant to a broader claim about model sensitivity. This failure differs from redundancy. Under redundancy, the prediction survives because other information substitutes for the support that was removed; under target mismatch, part of the support the claim concerns was never removed at all. The second can be measured directly by asking what the intervention covers, independently of how the model responds to the deletion.

To make this precise, we separate two targets that deletion-based evaluation usually conflates. The *intervention target* is the information a deletion actually withholds; the *claim target* is the support scope to which the resulting prediction change is interpreted. In source-target removal, the source target is the intervention target, whereas a conclusion about sensitivity to the harmful content in the observation takes that broader support as its claim target. We call the correspondence between the two *removal fidelity*. When fidelity is low, a small prediction change measures sensitivity to the removed interval, not to the support the claim is about. Removal fidelity is therefore a property of the pairing between an intervention and a claim, not of the annotation or the model alone.

We study removal fidelity with HARMFULFAITH, which evaluates four models on sampled frames and ASR/OCR text from three harmful-video datasets. The starting point is an asymmetry: retaining only the source target preserves 90.0–93.1% of HARMFUL predictions, whereas removing it changes only 19.1–42.6%. Preservation alone says little about the source target, because in a visual control an equal number of random frames preserves the prediction almost as often. The informative question is what removal leaves behind. Annotators blind to the source target and model behavior mark the support in each observation, which we call *audited support*. Source targets cover a median 69.2% of it and cover it fully in only 18.2% of cases with audited support. The uncovered remainder matters: when a clear residual cue survives source removal, predictions rarely change, and the gap between source and audited-support removal grows as coverage falls. Under a matched deletion budget, removing audited support changes 88.9–93.8% of predictions, against 31.7–34.2% for random deletion. Removal response thus tracks how much relevant support the intervention covers, not merely whether an annotated interval is removed. Our contributions are summarized as follows:

- We show that a valid temporal annotation need not license a broader deletion-based claim, and make this testable through *removal fidelity*, operationalized as coverage of independently audited support, which predicts how far source removal understates sensitivity.

- Across visual and text observations of harmful videos, we show that source targets often leave audited support available, fully covering it far less often than in general temporal grounding, and that clear residual support is associated with much weaker removal response; in the text audit, lower target coverage is associated with a larger gap between source and audited-support removal for both text models, even after source adjustment.
- Under matched deletion budgets, audited-support removal produces substantially larger prediction changes than random deletion, and random deletions of the same size change predictions more often when they cover more audited support: response depends on which support is withheld, not on the number or length of the deleted ASR/OCR units alone.
- We provide common-base, repeatability, negative-safety, and transfer analyses that test the robustness and scope of these conclusions across models and datasets (Section 5).

2 Related Work

Harmful video understanding and evidence grounding. Harmful and hateful video understanding has progressed from video-level classification toward finer-grained analysis of when and why content is harmful. HateMM studies multimodal hate-video classification (Das et al., 2023); HateClipSeg and DeHate provide segment-level temporal annotations (Wang et al., 2025; Zhang et al., 2025); and recent work considers weakly supervised hate localization, broader harmful categories, multimodal harmful elements, contextual rationales, and multiple levels of video understanding (Sun et al., 2026; Chen et al., 2025; Liu et al., 2025; Lu et al., 2026; Wu et al., 2026; Zhang et al., 2026). These annotations serve their grounding objectives well. We study what happens when they are reused as deletion targets for claims about model sensitivity, a use they were not collected for.

Faithfulness and deletion-based evaluation. Interpretability work distinguishes plausible explanations from those that reflect model behavior (Jain & Wallace, 2019; Adebayo et al., 2018; Jacovi & Goldberg, 2020), and ERASER operationalizes this through sufficiency and comprehensiveness (DeYoung et al., 2020). Two facts are established: human rationales need not score well on these measures (Carton et al., 2020), and redundant information can mask the importance of deleted content, which Madsen et al. (2022) address by recursive masking with retraining; video explanation methods apply analogous perturbation tests (Lin et al., 2021; Chen et al., 2021), and video faithfulness work targets generated content (Jing et al., 2026). What remains open is the target: a weak deletion response is read as a property of the rationale or the model, without measuring how much of the support a claim concerns the deletion target covers, or whether response follows that coverage at a fixed budget. HarmfulFaith supplies these measurements: we audit support independently of both annotation and model, quantify target coverage, and show that response rises with it across cases, within cases at a fixed budget, and against same-budget random deletion (Appendix A).

3 Evidence Interventions and Removal Targets

We treat a removal response as the response to a specified intervention on the model's observation. Its interpretation depends on what that intervention withholds and what it leaves in place.

3.1 Interventions and Paired Measures

For instance i, let E_i denote the source target: temporal evidence derived from the dataset's annotations and fixed before model evaluation. We do not assume that E_i is unique, minimal, or exhaustive. The visual track uses sampled video frames and the text track uses normalized units from timestamped ASR/OCR observations; both are constructed independently of evaluated-model outputs, so that no part of the input construction adapts to the evaluated model's behavior.

For world-specific input x_i^w,

$$\hat{y}_i^w = f(x_i^w), \qquad w \in \{\text{orig}, \text{only}, \text{removed}\},$$

where H denotes `HARMFUL`. The original world provides the reference observation; the evidence-only and source-removed worlds are materialized from E_i and its temporal complement under the track-specific construction. For complete cases predicted `HARMFUL` in the original world, let

$$\mathcal{I}_H = \left\{ i : \hat{y}_i^{\text{orig}} = H \text{ and all three predictions are valid} \right\}.$$

We define

$$R_{\text{pres}} = \frac{1}{|\mathcal{I}_H|} \sum_{i \in \mathcal{I}_H} \mathbb{1}[\hat{y}_i^{\text{only}} = H], \qquad R_{\text{shift}} = \frac{1}{|\mathcal{I}_H|} \sum_{i \in \mathcal{I}_H} \mathbb{1}[\hat{y}_i^{\text{removed}} \neq H].$$

Evidence-only preservation R_{pres} asks whether the source target can sustain the prediction; removal-sensitive shift R_{shift}, which we also call the *removal response*, asks whether removing it changes the prediction. We report their difference $\Delta_{\text{PR}} = R_{\text{pres}} - R_{\text{shift}}$ as the PR gap. High preservation shows that the source target can sustain the prediction, not that it is uniquely responsible for it; what the removal response says beyond the intervention depends on the support the removal withholds.

3.2 Removal Targets and Coverage

Let O_i denote the model's observation, a fixed set of frames or text units, and $E_i^O \subseteq O_i$ the units falling within the source target. We write $W_i \subseteq O_i$ for the *intervention target*, the units actually withheld, and $T_i \subseteq O_i$ for the *claim target*, the support scope about which the response is interpreted. Removal fidelity describes how well W_i corresponds to T_i; we operationalize one direction of it through *coverage*, the fraction of the claim target that the intervention withholds:

$$C^{\text{cov}}(T_i, W_i) = \frac{|T_i \cap W_i|}{|T_i|}, \qquad |T_i| > 0.$$

Full coverage means $T_i \subseteq W_i$; exact matching, $W_i = T_i$, is stronger.

The support a model uses is not observable, so we estimate the claim target with *audited support* $S_i^{\text{obs}} \subseteq O_i$: the units that independent annotators judge supportive of the harmfulness decision within the observation. Annotators see neither E_i nor model predictions or intervention outcomes, so audited support is defined independently of how the model responds to its deletion. Throughout, *coverage* refers to the fraction of audited support that an intervention target withholds. For the source target,

$$C_i^{\text{src}} = C^{\text{cov}}(S_i^{\text{obs}}, E_i^O) = \frac{|S_i^{\text{obs}} \cap E_i^O|}{|S_i^{\text{obs}}|}.$$

This quantity places every intervention we study on one axis. Source removal covers C_i^{src} of audited support, a random deletion covers whatever fraction it happens to include, and audited-support removal, which withholds exactly S_i^{obs} from the original text observation, covers all of it. If removal response is governed by which support is withheld rather than by how much is deleted, response should rise along this axis at a fixed deletion budget. The visual audit uses a nested variant that withholds $E_i^O \cup S_i^{\text{obs}}$, extending source removal by the identified residual support frames.

Audited support is a human judgment bounded by the observation, not the model's complete causal evidence. We read each removal response as sensitivity to the specified intervention, and use coverage, varied across cases by the source annotation and within cases at a fixed deletion budget, to test how far that sensitivity extends to the support a broader deletion-based claim concerns.

4 Experimental Setup

4.1 Models, Data, and Primary Worlds

The direct-visual cohort contains 672 videos: 292 from HateClipSeg (Wang et al., 2025) and 380 from HateMM (Das et al., 2023). We evaluate Qwen3.6-27B and GLM-4.5V, with GPT-5.6-sol as an independent visual replication on the same cohort. The two GPT visual conditions use identical videos and frame sets but inherit the within-input frame orderings of the Qwen and GLM tracks, respectively; prompts and inference settings are otherwise identical. The text cohort, constructed later, adds MultiHateClip and contains 963 videos: 321 from HateClipSeg, 431 from HateMM, and 211 from MultiHateClip (Wang et al., 2024). It is evaluated with DeepSeek-V4-Pro and GPT-5.6-sol. All intervention videos are harmful under the source-to-task mapping, and reported rates condition on complete valid cases predicted `HARMFUL` in the original world (Section 3.1).

In the visual track, eight frames are sampled from E_i (E_8) and eight from its temporal complement (R_8). The original world combines four frames from each, while the evidence-only and source-removed worlds use E_8 and R_8, respectively. Because these worlds are not all literal subsets of one original observation, we additionally evaluate two common-base constructions (Section 4.2).

The text track projects E_i onto timestamped ASR/OCR observations generated with `faster-whisper` and EasyOCR and frozen before model evaluation. After masking, retained text is normalized and rechunked; boundary-crossing ASR words are dropped rather than clipped, and timestamps and ASR/OCR channel labels are not shown to the models, so the model-facing text carries no temporal metadata. Source provenance, observation construction, prompts, inference settings, and denominator accounting for every evaluation condition are detailed in Appendix B.

4.2 Support Audits and Controls

Support audits. A two-annotator residual-support audit labels 190 visual and 230 text source-removed observations, sampled independently of model predictions and removal outcomes, as containing a clear residual cue, an ambiguous cue, or no clear residual cue. To measure coverage, a separate source-balanced 120-case text audit labels every ASR/OCR unit in the original observation as supportive of the harmfulness decision or not, yielding S_i^{obs}. Audited-support removal deletes exactly S_i^{obs}. A same-budget random control deletes exactly the same numbers of ASR and OCR units from the same observation, using up to 20 unique draws per case. A parallel 120-video visual audit labels support frames within the U_{16} observations described below; there, the broader-removal target $E_i^O \cup S_i^{\text{obs}}$ extends source removal by the identified residual support. Annotation protocols, agreement statistics, and non-empty sensitivities for both audits are reported in Appendix C.

Construction controls. A strict nested control filters the $4E+4R$ original observation into its existing E_4 and R_4 subsequences, so that both intervention inputs are literal subsets of the same original. An evidence-agnostic control instead samples a chronological 16-frame observation U_{16} without access to E_i and only then partitions the frames by temporal membership, so that the observation itself is independent of the source target. On U_{16}, source-only retention and source removal are also compared with multi-draw random subsets of the same frame count, frozen before inference.

4.3 Transfer and Statistical Analysis

To test whether the mismatch extends beyond harmful-video annotation, we use 120 Charades-STA test query–video pairs (Gao et al., 2017). On U_{16} observations, Qwen3.6-27B and GLM-4.5V predict whether the queried event is `PRESENT`, `ABSENT`, or `UNCERTAIN`; two annotators, blind to the source interval and model outcomes, identify event-support frames. We compare source-interval removal, broader-target removal, and same-frame-count random removal on these observations.

Paired comparisons report risk differences with case-level bootstrap confidence intervals and exact McNemar tests for binary outcomes. Multi-draw random controls are averaged within case before aggregation, and the same-cohort matched text control uses a centered randomization test. Models, tracks, target constructions, and valid denominators are analyzed separately, and all random draws are fixed before inference; the full statistical procedures, including seeds, are given in Appendix B.9.

5 Removal Response Tracks Support Coverage

5.1 A Motivating Asymmetry

We begin with the pattern that motivates the analysis. Across all six evaluation conditions, retaining only the source target preserves 90.0–93.1% of original `HARMFUL` predictions, whereas removing it changes only 19.1–42.6% (Figure 2). Read naively, the source target looks both sufficient and largely dispensable. Two further observations point to a more careful reading of this asymmetry.

First, preservation is not specific to source-selected content. On the U_{16} observations, retaining a same-count random frame subset preserves `HARMFUL` almost as often as retaining the source-covered frames: 86.52% versus 84.68% for Qwen (+1.83 points, 95% CI $[-0.73, 4.36]$) and 91.55% versus 89.00% for GLM (+2.55 points, $[-0.07, 5.25]$; Table 2B). This control cannot distinguish distributed or redundant harmful cues, broadly informative video content, and model-level label tendency. It does limit what evidence-only preservation can establish: the source target can sustain the prediction, but so can comparably sized content drawn elsewhere in the observation. This is one reason the analyses below focus on what the removal target covers rather than on preservation.

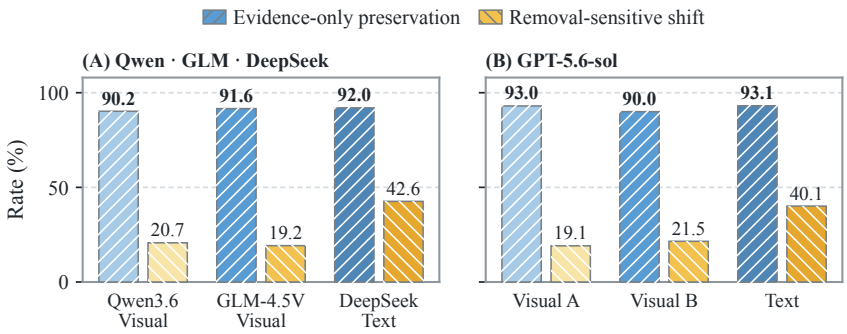


Figure 2: **A motivating asymmetry.** Evidence-only preservation (whether the source-only input remains HARMFUL) is much higher than removal response (whether source removal changes the HARMFUL prediction). Rates condition on complete cases predicted HARMFUL in the original world. Preservation is not specific to source-selected frames (Section 5.1), so the asymmetry motivates, rather than answers, the question of what source removal withholds. The two GPT visual conditions share the cohort and frame sets but use different within-input frame orderings, labeled A and B.

Second, the removal response varies with how much of the video the source target spans. On MultiHateClip, whose source intervals cover a median 95.2% of video duration, text removal changes 83.3% of DeepSeek and 77.8% of GPT predictions, and the PR gap narrows to 1.2 and 16.7 points; on HateClipSeg, the gap ranges from 52.1 to 68.8 points across the six conditions. Temporal extent is only a proxy for coverage, because what matters is whether the support in the observation falls inside the target. The next sections measure coverage directly with independent human audits.

5.2 Source Targets Incompletely Cover Audited Support

Source removal is mechanically exact: none of the 1,520 audited visual frames or 103,134 retained ASR/OCR units lies inside E_i, and all 2,073 boundary-crossing ASR words are dropped. It is not exhaustive. Among 99 text-audit cases with at least one audited support unit, the source target covers a median 69.2% [IQR 36.1, 90.4] of audited support and fully covers it in only 18.2%. Across all 1,611 adjudicated support units, 914 lie inside the source target and 697 outside it. Coverage is lowest on HateClipSeg, whose segment annotations cover a median 44.9% of audited support, with only 1 of 33 support-defined cases fully covered. The visual audit is similar: median coverage is 75.0% [41.9, 81.3] and 5 of 83 support-defined cases (6.0%) are fully covered. Because visual agreement is moderate ($\kappa \approx 0.42$, versus 0.79 for text), the graded coverage analyses below rest on the text audit. A source target is thus typically a valid but partial account of the support in the observation, and removing it leaves the remaining support available to the evaluated model.

5.3 Residual Support Predicts Weaker Source Removal

If the uncovered support matters to the model, source removal should be weakest where support visibly remains. The residual-support audit shows exactly this. After adjudication, 82 of 190 visual (43.2%) and 125 of 230 text source-removed observations (54.3%) contain a clear residual cue. In those cases, source removal changes only 5.1–9.7% of visual and 13.3–19.5% of text predictions; when no clear residual cue remains, it changes 50.0–63.2% and 95.7–100%. The contrast holds for all four model–track pairs and under either assignment of ambiguous cases to the two strata.

These are between-case contrasts, so we also vary the intervention within case. Additionally deleting the identified residual text units raises the removal response on clear-residual cases from 19.47% to 78.76% for DeepSeek and from 13.33% to 69.17% for GPT. Where an exact matched random deletion is feasible (69 and 75 cases), deleting the same numbers of random ASR and OCR units yields 21.74% and 13.33%, compared with 73.91% and 58.67% for the targeted deletion (Appendix C.9).

5.4 Removal Response Rises with Coverage

Residual support gives a binary view; coverage makes the relation graded. We examine it across cases, where coverage is set by the source annotation, and within cases at a fixed budget.

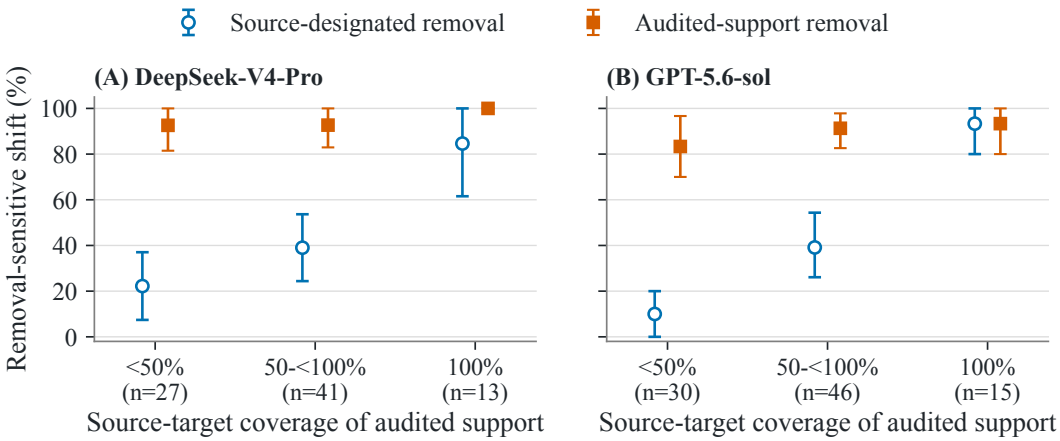


Figure 3: **Removal response rises with coverage.** Text-audit cases are stratified by coverage, the fraction of audited support included in the source target. The gap between source and audited-support removal shrinks as coverage increases. Full coverage means that all audited support lies within the source target, not that the two targets are identical. Error bars show 95% paired-bootstrap confidence intervals; stratum sizes (n) appear below each coverage bin for both models.

Across cases. Figure 3 stratifies text-audit cases by source-target coverage. When less than half of audited support lies inside the source target, source removal changes 22.2% of DeepSeek and 10.0% of GPT predictions, while audited-support removal changes 92.6% and 83.3%, gaps of +70.4 and +73.3 points. The gaps fall to +53.7 and +52.2 points under partial coverage and to +15.4 and 0.0 points under full coverage ($n = 13$ and 15), where source removal itself reaches 84.6% and 93.3%. Treating coverage continuously, each additional 25 percentage points of audited support outside the source target widens the paired gap by 12.5 points for DeepSeek (95% CI [4.3, 20.5]) and 13.6 points for GPT ([6.2, 20.7]) after adjusting for source dataset. Word-weighted coverage, consensus-only support, and exclusion of empty inputs give similar estimates for both models (Appendix C.8).

Within cases, at a fixed budget. Cases in different strata may differ in other ways, so we also use the same-budget random draws, which hold the case and the ASR/OCR deletion counts fixed but differ in how much audited support they happen to include. Each additional 10 percentage points of coverage raises the probability of a removal shift by 8.68 points for DeepSeek (95% CI [6.72, 10.45]) and 6.72 points for GPT ([4.42, 8.93]), with within-case permutation $p_{\text{emp}} < .001$ for both. Because equal unit counts can still differ in length, we additionally adjust for the fraction of words deleted; the coverage effect remains +8.35 points ([5.96, 10.38]) and +5.65 points ([2.82, 7.94]). Coverage thus predicts response even when the deletion budget is held fixed per case.

5.5 Coverage, Not Deletion Budget, Explains the Audited-Support Response

On the text-audit cohort, removing audited support changes 93.8% of DeepSeek and 88.9% of GPT predictions, compared with 40.7% and 38.9% under source removal. Audited-support removal is the full-coverage endpoint of the axis in Section 3.2, but it also deletes a different amount of text. The question is whether coverage or deletion budget explains its larger response. Because audited support is fixed by blinded annotators before any model is run, this comparison is not constructed from deletion outcomes, and no unit is selected for its effect on the prediction of either model.

Table 1 reports the same-budget random control. Audited-support removal exceeds the random expectation by +59.63 points for DeepSeek (95% CI [51.09, 68.18]) and +57.23 for GPT ([48.34, 66.09]), with $p_{\text{emp}} < .001$ for both. Two checks address simpler explanations. The random baseline is not an irrelevant control: random targets include about 65% of audited support on average. Nor does deleted length explain the gap: audited targets are somewhat longer, yet matching deleted word length as well as unit count still leaves +47.20 and +46.91 points. Under the same budget, what is withheld matters far more than how much, as the coverage axis of Section 3.2 anticipates.

Other constructions preserve the ordering. Source and audited targets use different pipelines (temporal masking versus item deletion); deleting both literally from the same item base removes this confound, and gaps remain +37.0 to +68.8 points. Restricting the target to units both annotators in-

Table 1: **Audited-support removal versus same-budget random deletion.** Removal response (%) on the text-audit cohort. Random targets delete the same numbers of ASR and OCR units as S_i^{obs} and are averaged within case; markers follow Figure 3. Parentheses give Δ = audited − random in pp; all $p_{\text{emp}} < .001$ for both models (10,000-replicate two-sided centered randomization test).

			Removal response (%)			
Cohort	**Model**	n	○ **Source**	**Random**	■ **Audited** (Δ)	**95% CI of** Δ
Primary	DeepSeek-V4-Pro	81	40.74	34.19	**93.83** (+59.63)	[51.09, 68.18]
	GPT-5.6-sol	90	38.89	31.66	**88.89** (+57.23)	[48.34, 66.09]
Non-empty	DeepSeek-V4-Pro	75	38.67	28.93	**93.33** (+64.41)	[55.88, 72.62]
	GPT-5.6-sol	82	36.59	25.00	**87.80** (+62.81)	[53.59, 71.80]

Table 2: **Visual construction and same-count random controls. A:** the source target under three input constructions (uniform U_{16} is evidence-agnostic); parentheses give paired nested-minus-equal-count changes in PR gap. **B:** source target versus same-count random frame subsets on U_{16}, with differences and 95% CIs. Rates are in %; gaps and differences are in pp, from unrounded rates.

	Qwen3.6-27B				GLM-4.5V			
Input	n	**Pres.**	**Shift**	**PR gap**	n	**Pres.**	**Shift**	**PR gap**
A. Construction controls: the same source target under three input constructions								
Equal-count E_8/R_8	357	90.2	20.7	69.5	313	90.7	21.4	69.3
Nested E_4/R_4	357	86.8	28.3	**58.5** (−10.9)	313	83.1	30.4	**52.7** (−16.6)
Uniform U_{16}	357	86.6	26.3	**60.2**	355	91.5	20.0	**71.6**
B. Random controls on U_{16}: source target versus same-count random frame subsets								
Random subset	356	84.68	19.05		355	89.00	13.95	
Source target	356	86.52	26.40		355	91.55	20.00	
Source − random		**+1.83** [−0.73, 4.36]	**+7.35** [3.55, 11.26]			**+2.55** [−0.07, 5.25]	**+6.05** [2.75, 9.40]	

dependently labeled as support gives consensus-minus-random gaps of +58.90 and +60.97 points. In the visual audit, extending source removal by residual support exceeds same-count random deletion by +53.01 points for Qwen and +47.24 for GLM on 11 and 10 non-empty cases (Appendix C).

5.6 Construction and Repeatability Controls

The primary visual worlds are not all literal subsets of a single original observation, so we check that construction does not drive the pattern. Table 2A shows that the preservation–removal gap remains 58.5 and 52.7 points for Qwen and GLM under the strict nested construction, and 60.2 and 71.6 points on the evidence-agnostic U_{16} observation; the nested construction narrows the gap by 10.9 and 16.6 points but leaves it above 50 in every case. The separation also persists when U_{16} frames and aligned ASR/OCR are supplied jointly to Qwen: 95.28% preservation against 19.81% removal ($n = 106$; Appendix D.6). Panel B compares the source target with same-count random frames on U_{16}: for removal it is more consequential, by 7.35 and 6.05 points, but by far less than the audited-support margins above. This modest margin is consistent with partial coverage: the source target and random frames each withhold only part of the support. For retention, by contrast, it is barely more sustaining than random frames (+1.83 and +2.55 points; both confidence intervals include zero). Exact-input reruns keep no-op instability below the removal shift in every condition, and shifts are dominated by transitions to `NOT_HARMFUL` rather than `UNCERTAIN` (Appendix B.10).

5.7 The Same Mismatch Appears in Temporal Event Grounding

We next ask whether the mismatch is specific to harmful-video annotation. On Charades-STA, where the claim concerns whether the queried event occurs, source intervals cover a median 66.7% of the observed event support and fully cover it in 38.2% of 110 support-defined cases, compared with 6.0% for harmful videos under the same 16-frame construction (Section 5.2). Removing the audited support together with the source interval raises the removal response from 53.6% to 84.5% for Qwen (+31.0 points) and from 37.4% to 60.7% for GLM (+23.4 points), and this broader-target

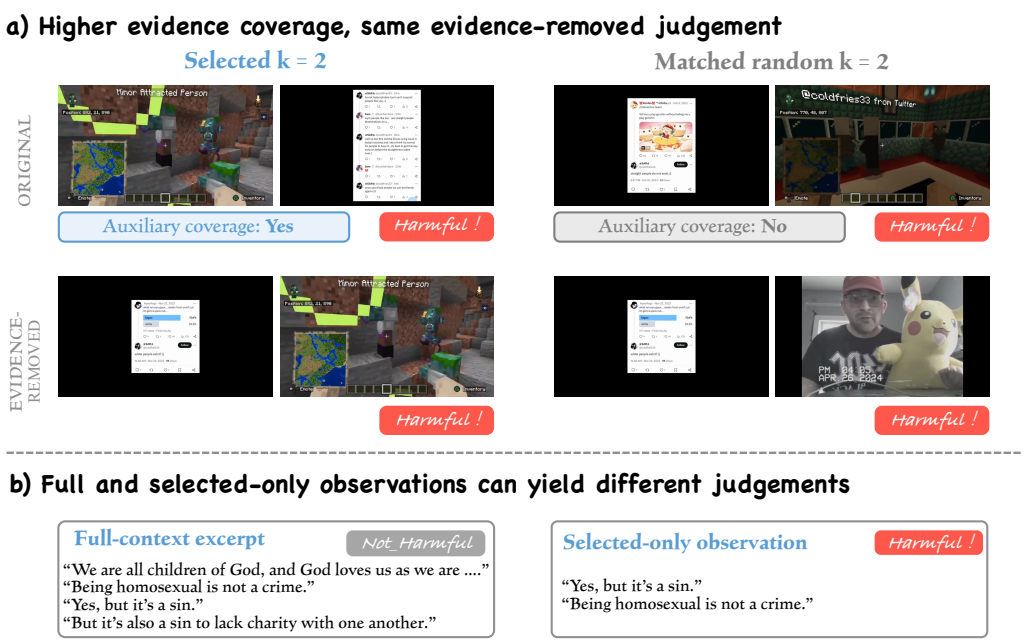


Figure 4: Case studies of evidence coverage and behavioral response. **(A)** GPT-5.6-sol visual example from the selected-evidence analysis (Appendix F): the selected $k = 2$ subset contains at least one item tagged as harmfulness-relevant under the auxiliary annotation, whereas its matched-random counterpart does not, yet both receive `HARMFUL` judgments in the original and evidence-removed worlds. Subsets are selected separately within each world, so the example is descriptive rather than a literal deletion comparison. **(B)** GPT-5.6-sol text example from the separate negative-safety analysis (Appendix G): the full-context observation is judged `NOT_HARMFUL`, whereas its selected-only observation is judged `HARMFUL`. Full-context text in (B) is excerpted here for readability.

removal exceeds same-frame-count random removal by 50.9 and 46.6 points. The mismatch is thus not confined to harmful-video annotation, but full coverage is far rarer there (Appendix E).

5.8 Case Study

Figure 4 illustrates why neither preservation nor the presence of annotated content settles what a judgment depends on. In (A), a subset containing annotated harmfulness-relevant content and a random subset without it lead to the same `HARMFUL` judgment, consistent with harmful cues that are distributed across the observation. In (B), a locally relevant excerpt is judged `HARMFUL` although the full observation is judged `NOT_HARMFUL`: content that sustains a judgment in isolation need not be what determines the judgment in context, a limit of preservation-based evidence. The main analyses therefore rest on removal targets whose coverage of audited support is measured directly.

6 Conclusion and Future Work

Temporal deletion results are interpretable only relative to what the intervention removes. Source annotations in harmful-video datasets provide valid evidence while incompletely covering harmful support that recurs across speech, on-screen text, and scenes in the model's observation. Across our audits and controls, removal responses become larger when interventions cover more of that support, and when audited support is removed rather than an equal-budget random subset. In harmful-video evaluation, a weak removal response is therefore easily misread as a model ignoring harmful content. Whenever deletion supports broader claims about model sensitivity, evaluations should report what fraction of audited support the removal target covers and whether a clear residual cue remains. Future work should study removal fidelity under denser multimodal observations and broader temporal grounding settings, develop methods for constructing high-fidelity removal targets at scale, and jointly consider prediction correctness and the model's behavioral response to intervention.

ETHICS STATEMENT

This work studies harmful and hateful video understanding using existing research datasets and controlled model evaluations. Human annotation is used for residual-support audits and observation-bounded support audits in text and sampled visual observations; a separate transfer audit uses ordinary temporal event grounding. For harmful-video annotation, exposure is limited to the material required for the analysis. Independent annotators are blinded to model predictions, intervention outcomes, and the source target when the protocol requires support identification. We report aggregate results and do not use the analysis to make claims about individual people or demographic groups. The purpose of this work is to improve the validity of evidence-grounded evaluation for video understanding.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the evaluation protocol and reported analyses. Appendix B documents the source annotations, visual and text observation construction, exact model-facing prompts, inference settings, and evaluation denominators. Appendix B.9 specifies the statistical procedures, randomization protocols, and frozen seeds, while Appendix B.10 reports exact-input repeatability. Appendix C gives the support-audit procedures and additional target-control analyses, and Appendix E describes the Charades-STA transfer protocol. We also provide the corresponding code and supporting resources in the supplementary material.

AI USE STATEMENT

Generative AI tools were used in a supportive manner during manuscript preparation, primarily for language polishing, code implementation, paper review, formatting suggestions, and improving clarity of presentation for the paper.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9505–9515, 2018.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9294–9307. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.747.

Yanbei Chen, Thomas Hummel, A. Sophia Koepke, and Zeynep Akata. Where and when: Space-time attention for audio-visual explanations. *arXiv preprint arXiv:2105.01517*, 2021. doi: 10.48550/arXiv.2105.01517. URL https://arxiv.org/abs/2105.01517.

Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. SafeWatch: An efficient safety-policy following video guardrail model with transparent explanations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xjKz6IxgCX.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. HateMM: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pp. 1014–1023, 2023. doi: 10.1609/icwsm.v17i1.22209.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.408.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285, 2017. doi: 10.1109/ICCV.2017.563.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. doi: 10.48550/arXiv.2312.11805. URL https://arxiv.org/abs/2312.11805.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.386.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357.

Liqiang Jing, Viet Dac Lai, Seunghyun Yoon, Trung Bui, and Xinya Du. FIFA: Unified faithfulness evaluation framework for text-to-video and video-to-text generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11418–11456, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.555. URL https://aclanthology.org/2026.findings-acl.555/.

Wonjeong Jo and Magdalena Wojcieszak. MetaHarm: Harmful YouTube video dataset annotated by domain experts, GPT-4-Turbo, and crowdworkers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pp. 2496–2509, 2025. doi: 10.1609/icwsm.v19i1.35951.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.342. URL https://aclanthology.org/2024.emnlp-main.342/.

Xinmiao Lin, Wentao Bao, Matthew K. Wright, and Yu Kong. Gradient frequency modulation for visually explaining video understanding models. In *British Machine Vision Conference*, 2021. doi: 10.5244/C.35.15.

Xuannan Liu, Zekun Li, Zheqi He, Peipei Li, Shuhan Xia, Xing Cui, Huaibo Huang, Xi Yang, and Ran He. Video-SafetyBench: A benchmark for safety evaluation of video LVLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 152459–152500, 2025. doi: 10.52202/085713-4587. Datasets and Benchmarks Track.

Junyu Lu, Deyi Ji, Liqun Liu, Xiaokun Zhang, Youlin Wu, Roy Ka-Wei Lee, Peng Shu, Huan Yu, Jie Jiang, Bo Xu, Liang Yang, and Hongfei Lin. Decoding multimodal cues: Unveiling the implicit meaning behind hateful videos. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1255–1265. Association for Computing Machinery, 2026. doi: 10.1145/3805712.3809637.

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, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.679. URL https://aclanthology.org/2024.acl-long.679/.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1731–1751. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.125.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774. URL https://arxiv.org/abs/2303.08774.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, pp. 510–526, 2016.

Qiyue Sun, Tailin Chen, Yinghui Zhang, Yuchen Zhang, Jiangbei Yue, Jianbo Jiao, and Zeyu Fu. MultiHateLoc: Towards temporal localisation of multimodal hate content in online videos. In *Proceedings of the ACM Web Conference 2026*, pp. 9024–9032. Association for Computing Machinery, 2026. doi: 10.1145/3774904.3793032.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. MultiHateClip: A multilingual benchmark dataset for hateful video detection on YouTube and Bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7493–7502. Association for Computing Machinery, 2024. doi: 10.1145/3664647.3681521.

Han Wang, Zhuoran Wang, and Roy Ka-Wei Lee. HateClipSeg: A segment-level annotated dataset for fine-grained hate video detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13304–13310. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3758289.

Jiajun Wu, Haoyu Kang, Yining Sun, Jiacheng Hou, Heng Zhang, Danyang Zhang, Zhenjun Zhao, Haochi Zhang, Leixin Sun, Eric Hanchen Jiang, Yushan Li, Ruiyu Li, Mengkai Huang, Yan Gao, Xu Zhang, and Guancheng Wan. HarmVideoBench: Benchmarking harmful video understanding in large multimodal models. *arXiv preprint arXiv:2606.27187*, 2026. doi: 10.48550/arXiv.2606.27187. URL https://arxiv.org/abs/2606.27187.

Yuchen Zhang, Tailin Chen, Jiangbei Yue, Yueming Sun, Rahul Singh, Jianbo Jiao, and Zeyu Fu. DeHate: A holistic hateful video dataset for explicit and implicit hate detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13177–13183. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3758272.

Yuchen Zhang, Shuang Dai, Zeyu Fu, Yunfei Long, Ravi Shekhar, and Haralambos Mouratidis. CLARA: Clip-level multimodal alignment with VLM-derived rationales for hateful video detection. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026. doi: 10.1145/3767308.3836438. URL https://arxiv.org/abs/2608.15905.

A Relation to Existing Deletion-Based Evidence Evaluation

Deletion-based faithfulness has a well-established operational form. In rationale evaluation, *comprehensiveness* measures how much a prediction changes after a supplied rationale is removed (DeYoung et al., 2020; Carton et al., 2020). Prior work also shows that human rationales may be incomplete and that redundant predictive information can weaken deletion-based signals (Carton et al., 2020; Madsen et al., 2022). Video explanation methods use analogous interventions by deleting or perturbing salient spatiotemporal regions and measuring the resulting prediction change (Lin et al., 2021; Chen et al., 2021). Recent video faithfulness work also studies factual consistency and hallucination in generated video or video-conditioned text (Jing et al., 2026), while temporal and harmful-video work increasingly provides intervals, segments, harmful elements, or rationales associated with video-level judgments (Wang et al., 2025; Zhang et al., 2025; Sun et al., 2026; Lu et al., 2026; Wu et al., 2026). These directions address different outputs and notions of faithfulness from the target correspondence studied here. Table 3 summarizes how we interpret the intervention responses used in this work.

Our analysis builds on these results and asks a narrower question: what happens when a temporal annotation collected as valid evidence is reused as the target of a deletion-based claim. The deletion response is well defined for the information actually removed; a broader interpretation also depends on how that information corresponds to the support named by the claim. Redundancy and target mismatch are distinct mechanisms. Under redundancy, the intervention withholds the support a claim concerns, yet other predictive information allows the prediction to persist. Under target mismatch, part of that support was never withheld. Both can produce weak deletion responses, and they can co-occur, but only the second is diagnosed by measuring what the intervention covers. Our experiments measure this coverage directly rather than inferring it from prediction persistence.

We do so in two complementary ways. Observation-bounded support audits measure how much independently identified support falls within the source target and whether this coverage tracks the difference between source and audited-support responses. Matched randomization asks whether the larger audited-support response can instead be explained by properties such as the amount of observation content removed. These analyses characterize the scope of the removal-based conclusion without assuming that the audited support sets recover the model's complete causal evidence.

Table 3: **What the intervention responses support.** Each row states the behavioral quantity and the interpretation warranted by the corresponding target construction.

Intervention	Behavioral quantity	Supported interpretation	Not established by this intervention
Source evidence only	R_{pres}	The source target can sustain the prediction under the restricted observation	The source target is necessary, uniquely relied upon, or more sustaining than same-count random content (Appendix B.4)
Source removal	R_{shift}	Sensitivity to the frozen source-removal intervention	Dependence on all task-relevant or semantic support
Same-count random control	Matched-random difference	A larger audited-target response is not explained by matched deletion count alone	All other properties of the removed content are controlled, or target identity is the unique cause of the response
Audited-support removal	Audited-support shift	Sensitivity to the protocol-defined audited target within the frozen observation	Dependence on all support in the raw video or on the model's complete causal evidence

B Intervention Construction and Evaluation Details

B.1 Source Temporal Annotations and Extent

The source target E_i is inherited from the temporal annotations released with each source dataset rather than produced by the evaluated model. HateClipSeg (Wang et al., 2025) assigns labels to predefined semantic segments; we retain all segments assigned a qualifying non-normal category

(`hateful`, `insulting`, `sexual`, `violence`, or `harm`). HateMM (Das et al., 2023) provides annotator-selected `hate_snippet` spans used as rationales, while MultiHateClip (Wang et al., 2024) provides start–end intervals localizing hateful or offensive content. We retain all valid qualifying intervals from each source.

Across sources, overlapping or adjacent intervals are normalized into their temporal union. These annotations define source-specific grounding targets rather than exhaustive support sets for the video-level judgment. Table 4 summarizes the annotation target inherited from each source.

Table 4: **Source annotations used to construct the source target E_i.**

Dataset	Source annotation	Annotation target	HARMFULFAITH E_i
HateClipSeg	Segment labels and timestamps	Harmful/offensive categories over predefined semantic segments	Union of all qualifying segments
HateMM	`hate_snippet`	Annotator-selected hateful rationale spans	Union of all valid spans
MultiHateClip	`Duration`	Interval(s) containing hateful or offensive content	Union of all valid intervals

Table 5 characterizes the temporal extent of the resulting E_i in the frozen intervention cohort.

Table 5: **Temporal extent of the source target.** Continuous quantities are medians with interquartile ranges in brackets; duration is video duration and coverage is the temporal coverage of E_i. Multi-interval is the percentage of cases whose final source target contains more than one disjoint interval.

Dataset	n	# intervals	$\lvert E_i \rvert$ (s)	Duration (s)	Coverage (%)	Multi-interval (%)
HateClipSeg	321	3 [2, 5]	126.0 [60.0, 181.4]	239.7 [206.2, 265.5]	54.4 [26.6, 77.4]	77.9
HateMM	431	1 [1, 2]	94.0 [36.5, 163.5]	150.1 [84.8, 212.4]	80.2 [50.4, 92.5]	26.2
MultiHateClip	211	1 [1, 1]	25.0 [13.0, 46.0]	38.5 [21.1, 54.4]	95.2 [74.6, 98.4]	4.7

HateClipSeg reports segment-level Offensive/Normal agreement of $\alpha = 0.715$ before discussion and 0.757 after discussion. HateMM reports $\kappa = 0.625$ for the video-level Hate/Non-Hate decision. MultiHateClip reports $\kappa = 0.62$ for the English three-class video label and $\kappa = 0.72$ for the binary Hateful/Offensive-versus-Normal grouping. Neither HateMM nor MultiHateClip reports temporal-boundary agreement, and none of the three source protocols evaluates exhaustive coverage of all semantic support for the video-level judgment.

B.2 DIRECT-VISUAL INTERVENTION CONSTRUCTION

For each video, the source intervals and their complement are each partitioned into eight cumulative-duration bins, with one midpoint anchor selected from each bin. This yields

$$E_8 = \{E_{8,0}, \ldots, E_{8,7}\}, \qquad R_8 = \{R_{8,0}, \ldots, R_{8,7}\}.$$

The primary evidence-only and removed worlds use E_8 and R_8, respectively. The original uses

$$E_4 = \{E_{8,0}, E_{8,2}, E_{8,5}, E_{8,7}\}, \qquad R_4 = \{R_{8,0}, R_{8,2}, R_{8,5}, R_{8,7}\},$$

presented under the deterministic frozen ordering protocol.

The strict nested control filters the exact frozen original payload into its existing E_4 and R_4 subsequences. No frame is added, resampled, or reordered. Table 6 reports the matched equal-count and nested results.

The paired nested-minus-equal-count changes in PR gap are -10.9 points (95% CI $[-15.1, -6.7]$) for Qwen and -16.6 points ($[-23.3, -9.9]$) for GLM.

The final primary direct-visual cohort contains 672 videos: 292 HateClipSeg and 380 HateMM videos. This cohort was frozen before MultiHateClip was incorporated into the subsequently constructed text-observation cohort; the modality difference therefore reflects the frozen evaluation design rather than a lack of MultiHateClip visual material. For the GPT visual replication, Ordering A

Table 6: **Strict common-base visual control.** Equal-count and nested rows use the same common-valid set within each model. The last column is the paired nested-minus-equal-count change in PR gap.

Construction	n	Pres. (%)	Rem. shift (%)	PR gap (pp)	Change in gap (pp)
Qwen3.6-27B					
Equal-count E_8/R_8	357	90.2	20.7	69.5	–
Nested E_4/R_4	357	86.8	28.3	58.5	−10.9 [−15.1, −6.7]
GLM-4.5V					
Equal-count E_8/R_8	313	90.7	21.4	69.3	–
Nested E_4/R_4	313	83.1	30.4	52.7	−16.6 [−23.3, −9.9]

Table 7: U_{16} **structural accounting.**

Source	Cohort	Materialized	Empty U^E	Empty U^R	Structural triads
HateClipSeg	292	292	3	9	280
HateMM	380	380	6	32	342
Overall	**672**	**672**	**9**	**41**	**622**

Table 8: **Evidence-agnostic U_{16} common-base results.** Rates (%) condition on strict-valid cases predicted `HARMFUL` in the corresponding U_{16} original observation; the PR gap and its 95% CI are in pp.

Source	n	Pres.	Rem. shift	H→NH	H→U	PR gap	95% CI
Qwen3.6-27B							
Overall	357	**86.55**	**26.33**	26.33	0.00	**60.22**	[54.06, 65.83]
HateClipSeg	157	77.07	37.58	37.58	0.00	39.49	[29.94, 49.04]
HateMM	200	94.00	17.50	17.50	0.00	76.50	[70.00, 82.50]
GLM-4.5V							
Overall	355	**91.55**	**20.00**	19.44	0.56	**71.55**	[66.48, 76.34]
HateClipSeg	156	87.82	25.64	24.36	1.28	62.18	[53.85, 69.87]
HateMM	199	94.47	15.58	15.58	0.00	78.89	[72.86, 84.92]

inherits the frozen request-bound within-input frame ordering from the Qwen track and Ordering B that from the GLM track. The two conditions use the same videos and unordered frame sets, with otherwise identical prompts and inference configuration.

B.3 Evidence-Agnostic Uniform Common-Base Visual Control

The U_{16} control constructs the original observation without access to E_i. Each video timeline is divided into 16 equal-duration bins and one midpoint anchor is materialized per bin under the frozen decode policy. The resulting frames are stored in chronological order.

Only after the 16-frame observation is frozen do we join E_i and partition frames by their realized timestamps:

$$U_i^E = \{f \in U_i^{16} : t_f \in E_i\}, \qquad U_i^R = U_i^{16} \setminus U_i^E.$$

No frame is added, resampled, duplicated, or reordered. Empty evidence or complement subsets are excluded rather than repaired. Table 7 summarizes the resulting structural accounting.

For both Qwen and GLM, three HateClipSeg videos lack a complete valid prediction triad and are excluded, leaving 619 complete triads per model. The visual inputs themselves pass the materialization, membership, common-base, and ordering audits.

Among structurally valid cases, the median numbers of evidence and complement frames are 11 [5, 14] and 5 [2, 11] overall. For HateClipSeg they are 9 [4, 12] and 7 [4, 12]; for HateMM they are 12 [7, 15] and 4 [1, 9]. Table 8 reports the overall and source-specific behavioral results.

Table 9: **Multi-draw same-count randomization on the evidence-agnostic U_{16} observation.** Expected random rates are case-weighted averages over 16 or 20 unique same-count random draws.

Source	n	Source shift (%)	Exp. random (%)	Δ_{shift} (pp)	95% CI	Empirical p
Qwen3.6-27B						
Overall	356	26.40	19.05	**+7.35**	[3.55, 11.26]	.00010
HateClipSeg	157	37.58	26.26	+11.32	[4.90, 17.75]	.00010
HateMM	199	17.59	13.37	+4.22	[−0.15, 8.89]	.0146
GLM-4.5V						
Overall	355	20.00	13.95	**+6.05**	[2.75, 9.40]	.00010
HateClipSeg	156	25.64	17.88	+7.76	[1.91, 13.61]	.00070
HateMM	199	15.58	10.86	+4.72	[0.93, 8.71]	.00210

Across the 3,732 model–world requests, 3,724 yield valid outputs under the frozen evaluation contract; incomplete triads are excluded from the paired analysis.

B.4 Multi-Draw Same-Count Randomization Control

For each structurally valid U_{16} observation, random subsets contain the same number of frames as U_i^E and their complements therefore match the source removal frame budget. Sampling is independent of model predictions and outcomes, and the same frozen draw manifest is used for Qwen and GLM. When only 16 unique same-count subsets are feasible, all 16 are enumerated; otherwise we use 20 frozen unique draws. Included cases have complete paired predictions for every planned draw. One Qwen case is excluded because one of its 16 paired draws remains invalid under the frozen output contract.

For each case we average the random outcomes over its valid draws and then average across cases, so every video receives equal weight. Table 9 reports the resulting comparison. Confidence intervals use 5,000 paired case-level bootstrap replicates with seed 20260903. A separate 10,000-replicate Monte Carlo reference distribution selects one completed random draw per case in each pseudo-experiment; empirical p-values use the plus-one correction.

The preservation contrast is smaller. Qwen source-only preservation is 86.52% versus a multi-draw random expectation of 84.68%, a +1.83-point difference (95% CI $[-0.73, 4.36]$). GLM shows 91.55% versus 89.00%, a +2.55-point difference ($[-0.07, 5.25]$). For removal, both source-target statistics lie at the upper edge of the corresponding Monte Carlo reference distributions. The removal difference is positive in all four model–source cells; the Qwen–HateMM bootstrap interval slightly crosses zero even though its randomization p is below .05, so we do not interpret individual source cells as stand-alone significance claims.

An earlier single-draw control gave +7.84 and +7.32 point removal differences for Qwen and GLM. Multi-draw averaging yields +7.35 and +6.05 points, respectively. We use the multi-draw analysis throughout because it reduces dependence on any particular random subset.

B.5 Text-Observation Intervention Construction

The text track uses project-generated timestamped ASR/OCR observations frozen before model evaluation rather than transcripts or OCR supplied by the source datasets. English ASR is extracted with `faster-whisper` using the `small.en` checkpoint (`vad_filter=True`, beam size 1). English OCR is extracted with EasyOCR at a 0.35 confidence threshold from sampled frames within the predefined observation segments; targets that cannot be decoded through the primary OpenCV path use the frozen FFmpeg fallback.

The source target E_i is projected onto these timestamped observations. An ASR word is retained only when its full temporal extent lies within the corresponding retained interval; boundary-crossing words are dropped rather than clipped. OCR observations are timestamped point events, with points exactly at either endpoint of a source interval assigned to the evidence interval.

Before inference, leading and trailing whitespace is removed and consecutive whitespace is collapsed to a single ASCII space; case, punctuation, and Unicode content are otherwise unchanged.

Table 10: **Model identifiers and explicitly specified inference settings.** Token limits are adjudicator/selector maxima; – marks a parameter that was not explicitly supplied.

Model	OpenRouter identifier	Temperature	Reasoning	Max tokens
Qwen3.6-27B	`qwen/qwen3.6-27b`	0	disabled	32/128
GLM-4.5V	`z-ai/glm-4.5v`	0	disabled	32/128
DeepSeek-V4-Pro	`deepseek/deepseek-v4-pro`	0	disabled	32/128
GPT-5.6-sol	`openai/gpt-5.6-sol`	–	effort medium	40,000 (output)

Retained ASR events are concatenated chronologically and split preferentially at sentence-final punctuation before being chunked into at most 48 whitespace-delimited tokens. A frozen fallback for text without whitespace caps chunks at 240 Unicode code points. OCR is normalized per event and deduplicated by exact case-sensitive string match within each case–world payload before the same chunking rule is applied. Final units receive opaque identifiers under the frozen presentation order; timestamps and ASR/OCR channel labels are not shown to the evaluated model.

The support-coverage audit uses a half-open OCR end-boundary convention, whereas source masking gives evidence priority at exact interval endpoints. This implementation difference is empirically inert in the frozen 120-case audit: no OCR unit lies exactly on a source end boundary (0/904 overall and 0/436 among SUPPORT units). Remapping with the source-masking convention reproduces every case-level coverage value and all coverage–response results. The extraction settings and frozen observations are retained with the evaluation artifacts; the Python package version of `faster-whisper` and the EasyOCR package/model revision were not separately pinned.

B.6 INFERENCE AND OUTPUT CONTRACT

All primary adjudication conditions use the same model-facing instruction:

> You will receive only the content items shown below. Based solely on those items, classify the content as HARMFUL, UNCERTAIN, or NOT_HARMFUL. Return only one JSON object with exactly one key, verdict. verdict must be exactly one of HARMFUL, UNCERTAIN, or NOT_HARMFUL. Use UNCERTAIN only when the supplied items are insufficient for a reliable harmfulness decision. Do not provide any explanation or text outside the JSON object.

The selected-evidence analyses additionally use the following selector instruction:

> You will receive content items with opaque identifiers. Select exactly the requested number of items that are most necessary for deciding whether the overall content is harmful. Return only one JSON object with exactly one key, selected_ids. selected_ids must be an array of exactly K distinct supplied item IDs. Do not provide any explanation, rationale, caption, ranking, or text outside the JSON object.

Each request contains a single user message, with no system prompt, few-shot examples, or prior conversation history. Visual items are rendered as an opaque item identifier followed by the corresponding image; text items are rendered as an opaque identifier followed by the normalized text content. The adjudication response must be a single JSON object whose only key is `verdict`, with value exactly one of `HARMFUL`, `UNCERTAIN`, or `NOT_HARMFUL`. The parser requires this exact object/key/enum contract; for GLM, the frozen protocol additionally accepts only the exact `<|begin_of_box|>...<|end_of_box|>` wrapper. Protocol-invalid outputs are excluded rather than repaired. Retries are limited to transport failures and HTTP 408/429/5xx responses under the frozen schedule, with at most three attempts.

All model inference is conducted through OpenRouter. Table 10 summarizes the model identifiers and explicitly specified inference settings. Adjudicator and selector token limits are 32 and 128, respectively, for Qwen, GLM, and DeepSeek.

Parameters not listed above were not explicitly supplied. No model-side generation seed is used; the request-level frozen seed controls presentation ordering rather than model generation. Models and tracks are analyzed separately, and primary rates condition on complete original-`HARMFUL` triads.

Table 11: **Primary evaluation accounting and original-world predictions.** Left: frozen cohorts and complete triads per evaluation condition. Right: original-world prediction distribution before original-HARMFUL conditioning.

	Accounting		Original-world predictions			
Model (track)	Cohort	Complete triads	Valid n	H (%)	NH (%)	U (%)
Visual track						
Qwen3.6-27B	672	672	672	53.12	46.43	0.45
GLM-4.5V	672	640	666	54.50	42.49	3.00
GPT-5.6-sol (A)	672	672	672	61.61	36.90	1.49
GPT-5.6-sol (B)	672	672	672	62.20	36.31	1.49
Text track						
DeepSeek-V4-Pro	963	963	963	74.35	23.99	1.66
GPT-5.6-sol	963	963	963	81.62	16.82	1.56

B.7 Primary Evaluation Accounting

Table 11 summarizes the frozen primary cohort and complete-triad accounting for each evaluation condition.

B.8 Primary Denominator and Original-World Predictions

All 672 visual and 963 text intervention videos are labeled harmful under the frozen source-to-task mapping:

$$292 + 380 = 672$$

for HateClipSeg and HateMM in visual, and

$$321 + 431 + 211 = 963$$

for HateClipSeg, HateMM, and MultiHateClip in text. No source-labeled non-harmful video enters either primary intervention cohort.

Consequently, cases entering the primary original-HARMFUL denominator are label-consistent true positives relative to the frozen dataset labels. The separate non-harmful negative-safety cohort does not contribute to this denominator. Table 11 reports the original-world predictions before conditioning on original-HARMFUL.

The primary analysis therefore characterizes intervention behavior among dataset-harmful videos for which the model initially makes a label-consistent HARMFUL prediction; it is not an unconditional estimate over all harmful videos. The frozen prediction protocol provides categorical outputs rather than calibrated probabilities, so probability calibration is not estimable under this protocol.

B.9 Statistical Conventions

Primary paired analyses use video-level bootstrap confidence intervals and two-sided exact McNemar tests. Residual-support stratification compares different case strata and instead uses within-stratum bootstrap resampling, with two-sided Fisher exact tests reported as secondary analyses.

Existing frozen analyses retain their original seeds. The original observation-bounded and U_{16} controls use 5,000 video-level bootstrap replicates with seed 20260901. The U_{16} multi-draw randomization control uses 5,000 paired case-level bootstrap replicates with seed 20260903 and 10,000 Monte Carlo pseudo-experiments. The coverage–response analysis uses 10,000 paired case-level bootstrap replicates for the fixed strata and continuous trends. The within-case random-draw coverage analysis uses a case fixed-effects linear probability model, 10,000 case-cluster bootstrap replicates, and 10,000 within-case outcome permutations with seed 20260905. The repeatability, same-duration, and original same-count text controls retain their frozen protocols.

The additional target-control sensitivities use 10,000 case-level bootstrap replicates with seed 20260909. Multi-draw random outcomes are averaged within case before aggregation so that each

Table 12: **Exact-input repeatability.** E/B denotes exact three-way agreement and binary HARMFUL-versus-other agreement. The no-op ceiling is the larger binary instability across the original and removed inputs; excess is the frozen removal shift minus this ceiling.

Model	Valid	Orig. E/B	Rem. E/B	Shift (%)	Ceiling (%)	Excess (pp)
Visual track						
Qwen3.6-27B	120/120	100.0/100.0	95.8/95.8	18.3	4.2	**+14.2** [7.5, 21.7]
GLM-4.5V	120/120	98.3/99.2	95.8/96.7	19.2	3.3	**+15.9** [9.2, 22.5]
GPT-5.6-sol (A)	68/120	97.1/97.1	94.1/94.1	23.5	5.9	**+17.6** [7.4, 27.9]
GPT-5.6-sol (B)	83/120	96.4/96.4	94.0/94.0	14.5	6.0	**+8.4** [0.0, 16.9]
Text track						
DeepSeek-V4-Pro	120/120	92.5/92.5	93.3/95.0	42.5	7.5	**+35.0** [25.8, 43.3]
GPT-5.6-sol	120/120	97.5/97.5	95.8/97.5	39.2	2.5	**+36.7** [26.7, 44.2]

case receives equal weight. Length-adjusted within-case analyses resample cases rather than individual draws. Models, tracks, target constructions, and validity denominators are analyzed separately throughout. All reported risk differences and preservation–removal gaps are computed from the underlying full-precision rates and rounded only for display; a displayed difference therefore need not equal the subtraction of two already rounded percentages.

B.10 Exact-Input Repeatability

To bound ordinary endpoint instability, we rerun the exact frozen ORIGINAL and EVIDENCE-REMOVED requests on deterministic source-stratified 120-case samples, without selecting on intervention outcome. The frozen response is the first run and the fresh request is the second. Table 12 reports the exact-input repeatability sentinel for Qwen, GLM, DeepSeek, and GPT. In every reported row, the no-op ceiling is below the corresponding frozen removal shift; the GPT visual orderings use reduced valid subsets but show the same direction.

We use this sentinel only as a repeatability check, not as a causal decomposition of intervention effects.

C Primary Results and Support-Target Analyses

C.1 Complete Primary Controlled-World Results

Table 13 gives the complete primary transition statistics underlying Figure 2.

C.2 Source-Stratified Primary Results

Table 14 decomposes each primary model–track result by source dataset.

C.3 Same-Duration Temporal Random Control

As a temporal control, we compare source text removal with three frozen, disjoint, in-bounds random intervals of exactly the same duration. Random intervals are sampled without using model outcomes, residual annotations, or text content, and the original deletion pipeline is otherwise unchanged. The disjoint same-duration constraint is feasible for 67 of 240 selected model-cases, yielding 32 DeepSeek and 35 GPT cases with complete random draws. Table 15 summarizes the resulting comparison.

Fourteen random intervals remove no realized ASR/OCR unit and remain in the primary randomization, as specified before inference. Every retained payload is nonempty. The point estimates are consistent with source intervals being more consequential than arbitrary same-duration intervals. The feasible cohort is small, however, and the DeepSeek confidence interval crosses zero. We therefore use this result only as a secondary temporal control.

Table 13: **Evidence-only preservation and prediction response under source removal.** Rates (%) condition on complete cases predicted HARMFUL in the original world; PR gaps are in pp. These values underlie Figure 2.

Model	Track	n	Pres.	Rem. shift	H→NH	H→U	PR gap
Qwen3.6-27B	Visual	357	90.2	20.7	20.2	0.6	69.5
GLM-4.5V	Visual	359	91.6	19.2	18.1	1.1	72.4
DeepSeek-V4-Pro	Text	716	92.0	42.6	38.5	4.1	49.4
GPT-5.6-sol	Visual (A)	414	93.0	19.1	19.1	0.0	73.9
GPT-5.6-sol	Visual (B)	418	90.0	21.5	20.1	1.4	68.4
GPT-5.6-sol	Text	786	93.1	40.1	30.0	10.1	53.1

Table 14: **Primary controlled-world results by source dataset.** Source-specific rows exactly reaggregate to the corresponding primary model–track totals. Rates are in %, PR gaps in pp; shading groups the rows of each model–track condition.

Model	Track	Source	n	Pres.	Rem. shift	PR gap
Qwen3.6-27B	Visual	HateClipSeg	140	88.6	32.1	56.4
		HateMM	217	91.2	13.4	77.9
GLM-4.5V	Visual	HateClipSeg	153	90.8	28.1	62.7
		HateMM	206	92.2	12.6	79.6
GPT-5.6-sol	Visual (A)	HateClipSeg	165	91.5	28.5	63.0
		HateMM	249	94.0	12.9	81.1
GPT-5.6-sol	Visual (B)	HateClipSeg	169	86.4	34.3	52.1
		HateMM	249	92.4	12.9	79.5
DeepSeek-V4-Pro	Text	HateClipSeg	254	89.4	31.5	57.9
		HateMM	378	95.5	41.0	54.5
		MultiHateClip	84	84.5	83.3	1.2
GPT-5.6-sol	Text	HateClipSeg	272	90.8	22.1	68.8
		HateMM	388	94.3	40.5	53.9
		MultiHateClip	126	94.4	77.8	16.7

Table 15: **Contiguous same-duration temporal random control.** Random shift is the case-weighted mean over three frozen random intervals.

Model	n	Source (%)	Random (%)	RD (pp)	95% CI
DeepSeek-V4-Pro	32	28.12	13.54	+14.58	[−3.12, 32.29]
GPT-5.6-sol	35	22.86	6.67	+16.19	[0.00, 33.33]

C.4 Residual-Support Sampling and Annotation

The residual-support audit contains 190 visual and 230 text cases sampled independently of model predictions and removal outcomes. Two annotators independently assign *clear residual cue*, *ambiguous*, or *no clear residual cue* before consensus adjudication.

For visual observations, three-way raw agreement is 83.2% ($\kappa = 0.736$), and clear-versus-other agreement is 88.9% ($\kappa = 0.776$). For text, the corresponding values are 87.0% ($\kappa = 0.774$) and 92.2% ($\kappa = 0.843$). Table 16 reports the pre- and post-adjudication label distributions.

Table 17 further breaks the adjudicated residual-support prevalence down by source dataset.

C.5 Mechanical Verification of Source Removal

None of the 1,520 frames in the 190 audited R_8 payloads lies inside a source temporal interval. Across all 963 text videos, the removed observations retain 103,134 units (99,617 ASR and 3,517 OCR) with zero temporal-mask violations. All 2,073 ASR words crossing a mask boundary are dropped. Expanding from the four complement frames already present in the visual original to the full R_8 introduces a new clear residual cue in only 6 of 190 cases (3.2%).

Table 16: **Residual-support annotation distributions.**

Label	Annotator 1	Annotator 2	Adjudicated
Visual ($n = 190$)			
Clear residual cue	86	83	**82**
Ambiguous	37	41	**40**
No clear residual cue	67	66	**68**
Text ($n = 230$)			
Clear residual cue	124	124	**125**
Ambiguous	24	29	**25**
No clear residual cue	82	77	**80**

Table 17: **Adjudicated residual-support status by source dataset.**

Source	n	Clear (%)	Ambiguous (%)	No clear (%)
Visual				
HateClipSeg	89	41.6	15.7	42.7
HateMM	101	48.5	22.8	28.7
Text				
HateClipSeg	83	72.3	8.4	19.3
HateMM	95	57.9	13.7	28.4
MultiHateClip	52	19.2	9.6	71.2

Table 18: **Removal-sensitive behavior by adjudicated residual-support status.** The last column gives the no-clear-minus-clear difference in shift with its 95% within-stratum bootstrap CI.

Status	n	H→H	H→NH	H→U	Shift (%)	No clear − clear (pp)
Qwen3.6-27B, visual						
Clear	62	56	6	0	9.68	
No clear	19	7	10	2	63.16	**+53.48** [29.63, 76.15]
GLM-4.5V, visual						
Clear	59	56	3	0	5.08	
No clear	18	9	7	2	50.00	**+44.92** [21.00, 68.83]
DeepSeek-V4-Pro, text						
Clear	113	91	19	3	19.47	
No clear	39	0	36	3	100.00	**+80.53** [73.45, 87.61]
GPT-5.6-sol, text						
Clear	120	104	15	1	13.33	
No clear	46	2	28	16	95.65	**+82.32** [73.30, 90.00]

These checks verify source temporal exclusion; support coverage is assessed separately below.

C.6 Residual Support and Removal-Sensitive Behavior

Table 18 reports the full transition decomposition underlying the residual-support contrasts in the main text.

The no-clear-minus-clear risk differences are +53.48 points for Qwen (95% CI [29.63,76.15]), +44.92 for GLM ([21.00,68.83]), +80.53 for DeepSeek ([73.45,87.61]), and +82.32 for GPT ([73.30,90.00]).

C.7 Sensitivity to Ambiguous Residual-Support Labels

Table 19 shows that the direction of the residual-support contrast is unchanged for all four models under either assignment of ambiguous cases.

Table 19: **Sensitivity to alternative treatment of ambiguous residual-support labels.** Each block reports the risk difference in shift with its 95% CI and Fisher exact p.

	Clear + ambiguous vs. no clear			Clear vs. ambiguous + no clear		
Model	RD (pp)	95% CI	Fisher p	RD (pp)	95% CI	Fisher p
Qwen3.6-27B	+51.11	[27.65, 72.16]	1.10×10^{-5}	+30.32	[13.55, 46.94]	4.42×10^{-4}
GLM-4.5V	+42.50	[19.03, 67.22]	8.60×10^{-5}	+25.68	[11.13, 41.07]	9.95×10^{-4}
DeepSeek-V4-Pro	+75.19	[67.67, 81.95]	5.69×10^{-17}	+65.28	[53.19, 76.56]	1.06×10^{-15}
GPT-5.6-sol	+74.38	[65.02, 82.93]	4.31×10^{-17}	+73.23	[62.57, 82.89]	1.77×10^{-16}

Table 20: **Coverage of observation-bounded audited support by the source target.** Coverage and outside-target fractions are defined only for cases with at least one adjudicated support unit; values are medians with interquartile ranges.

Source	Defined n	Target coverage (%)	Outside target (%)	Fully covered
HateClipSeg	33	44.90 [12.50, 69.35]	55.10 [30.65, 87.50]	1/33 (3.03%)
HateMM	37	75.00 [38.89, 90.91]	25.00 [9.09, 61.11]	7/37 (18.92%)
MultiHateClip	29	80.00 [60.00, 100.00]	20.00 [0.00, 40.00]	10/29 (34.48%)
Overall	**99**	**69.23** [36.11, 90.40]	**30.77** [9.60, 63.89]	**18/99 (18.18%)**

Table 21: **Source versus audited-support removal.**

Model	n	Source (%)	Audited (%)	RD (pp)	95% CI	McNemar p
DeepSeek-V4-Pro	81	40.74	**93.83**	**+53.09**	[41.98, 64.20]	2.61×10^{-12}
GPT-5.6-sol	91	38.46	**89.01**	**+50.55**	[39.56, 61.54]	2.27×10^{-12}

C.8 Observation-Bounded Support Audit

The observation-bounded audit contains 120 cases, with 40 each from HateClipSeg, HateMM, and MultiHateClip. Two annotators label every ASR/OCR unit in the full frozen original text observation as Support, Not Support, or Uncertain. The annotation interface hides E_i, model predictions, and intervention outcomes. Let S_i^{obs} denote the resulting adjudicated support set.

For the text audit, audited-support removal deletes exactly S_i^{obs} from the frozen original observation. All Not Support and Uncertain units are retained. In particular, the audited target is not constructed as $E_i^O \cup S_i^{\text{obs}}$. The source-removal comparator is the separately frozen temporal evidence-removed world produced by the primary masking pipeline. The two conditions therefore instantiate distinct removal targets rather than a nested source-versus-union construction.

These support sets are defined only with respect to the frozen text observation; they are not intended as exhaustive accounts of support in the underlying video or of the model's causal features.

Across 2,983 observation units, three-way agreement is 88.94% with $\kappa = 0.7910$. Support-versus-other agreement is 90.78% with $\kappa = 0.8157$. The mean support-set Jaccard is 0.8376; the median is 0.9751 [0.7778,1.0000].

Twenty-one cases contain no adjudicated support unit in the frozen text observation. Among the remaining 99 cases, the source target covers a median 69.23% [IQR 36.11,90.40] of the audited support. Only 18/99 (18.18%) have all identified support contained within the source target. Table 20 reports the corresponding source-specific coverage statistics.

Full source-target coverage means $S_i^{\text{obs}} \subseteq E_i^O$; it does not require $S_i^{\text{obs}} = E_i^O$. Among the 18 support-defined cases with 100% coverage, the source target strictly contains the audited support set in 15 cases. Source and audited-support removal can therefore remain different interventions even under full coverage. Moreover, the source comparator is rematerialized through the frozen temporal masking pipeline rather than constructed as literal item deletion from the frozen original payload.

Across all 1,611 adjudicated support units, 914 lie within the source temporal target and 697 lie outside it. Table 21 compares source removal with audited-support removal.

Table 22: **Audited-support removal results by source.** Shifts are in %; RD is audited minus source removal in percentage points.

	DeepSeek-V4-Pro				GPT-5.6-sol			
Source	n	Source	Audited	RD	n	Source	Audited	RD
HateClipSeg	30	26.67	90.00	+63.33	32	12.50	84.38	+71.88
HateMM	34	38.24	94.12	+55.88	36	38.89	88.89	+50.00
MultiHateClip	17	70.59	100.00	+29.41	23	73.91	95.65	+21.74

Table 23: **Same-cohort ASR/OCR-count-matched random text control.** Random targets match the ASR/OCR unit counts of S_i^{obs} exactly. All empirical $p < .001$ under a 10,000-replicate two-sided centered randomization test with plus-one correction.

		Removal response (%)			Audited − random (pp)	
Model	n	Source	Audited	Random	Difference	95% CI
Primary cohort						
DeepSeek-V4-Pro	81	40.74	**93.83**	34.19	**+59.63**	[51.09, 68.18]
GPT-5.6-sol	90	38.89	**88.89**	31.66	**+57.23**	[48.34, 66.09]
Excluding empty audited-removal inputs						
DeepSeek-V4-Pro	75	38.67	**93.33**	28.93	**+64.41**	[55.88, 72.62]
GPT-5.6-sol	82	36.59	**87.80**	25.00	**+62.81**	[53.59, 71.80]

Table 22 shows the same directional increase within each of the three source datasets for both models.

Same-cohort matched-count random control. Source and audited-support removal differ in both target identity and the amount of realized text removed. We therefore construct a random control on the same audit cohort. For each case, every random target removes exactly the same numbers of ASR and OCR units as S_i^{obs} from the same frozen original observation. Random targets are not forced to avoid audited support and may therefore partially or fully overlap it. Cases with at most 20 feasible unique subsets enumerate all of them; otherwise 20 unique draws are frozen before inference. Random outcomes are averaged within case before aggregation, giving each case equal weight. Table 23 reports the primary and non-empty comparisons.

All DeepSeek calls are valid. One GPT random request violates the frozen request contract; its entire 20-draw case is excluded, changing the GPT cohort from 91 to 90 and the audited-support shift from 89.01% to 88.89%. The non-empty sensitivity gives the same direction and a slightly larger difference.

Across all prespecified draws, mean random–audited target overlap is 65.30% for DeepSeek and 65.97% for GPT (medians 68.75% and 70.59%). ASR overlap means are 67.52% and 67.99%, and OCR overlap means are 76.67% and 77.69%. Thus the random baseline frequently removes some audited support rather than serving as a deliberately disjoint negative target. Even under this overlap, audited-support removal remains substantially more disruptive. The comparison therefore shows that matched ASR/OCR deletion count alone does not account for the larger audited-support response.

A stricter disjoint ASR/OCR-type- and count-matched random target is feasible for only a small subset. On the resulting $n = 15$ common-valid sets, random versus audited-support shift is 6.67% versus 93.33% for DeepSeek and 0.00% versus 73.33% for GPT. Because feasibility sharply reduces the cohort, we retain this comparison as a secondary sensitivity and use the same-cohort multi-draw analysis above as the primary deletion-count control.

Count- and length-matched random control. Matching ASR/OCR unit counts does not guarantee that the removed text has the same realized length. On the original same-count cohorts, audited targets are somewhat longer: the median numbers of deleted words are 220.0 versus 175.7 under the random control for DeepSeek and 210.5 versus 172.8 for GPT. We therefore construct a stricter

Table 24: **Random text deletion matched in both ASR/OCR unit count and deleted word length.** Random targets match the audited ASR and OCR unit counts exactly and the total deleted word count within 5%, with a minimum tolerance of one word. All $p_{\text{emp}} < .001$.

Model	n	Audited (%)	Random (%)	RD (pp)	95% CI
Primary cohort					
DeepSeek-V4-Pro	78	**94.87**	47.67	**+47.20**	[38.12, 56.43]
GPT-5.6-sol	89	**89.89**	42.98	**+46.91**	[37.23, 56.80]
Excluding empty audited-removal inputs					
DeepSeek-V4-Pro	72	**94.44**	43.31	**+51.13**	[41.77, 60.37]
GPT-5.6-sol	81	**88.89**	37.35	**+51.54**	[41.46, 61.64]

Table 25: **Within-case association between random-target support coverage and removal response.** Estimates are changes in removal-shift probability per +10 percentage points of audited-support coverage. "q varies" counts cases with within-case variation in q_{ij}; "both vary" counts cases that also vary in removal outcome.

Model	Cases	q varies	Both vary	RD per +10 pp	p_{emp}
DeepSeek-V4-Pro	81	70	43	**+8.68** [6.72, 10.45]	.00010
GPT-5.6-sol	90	75	36	**+6.72** [4.42, 8.93]	.00010

random control that matches the ASR and OCR unit counts exactly while requiring the total deleted word count to fall within 5% of the audited target, with a minimum tolerance of one word.

Table 24 reports the resulting comparison. The stricter matching reduces the complete-valid cohorts to 78 DeepSeek and 89 GPT cases. Audited-support removal remains substantially more disruptive: the paired differences are +47.20 points for DeepSeek and +46.91 points for GPT. The corresponding non-empty sensitivities are +51.13 and +51.54 points. Thus, realized deletion length explains part of the original same-count contrast, but not most of it.

Thirty-five cases have only one feasible target under the stricter matching rule, and two candidate searches are search-limited. Across the 1,299 frozen nonzero random targets, audited-support overlap averages 71.30%, with a median of 75.76%. The random baseline therefore continues to remove substantial audited support rather than being constructed to avoid it.

Within-case support coverage among random targets. The frozen multi-draw control also varies which audited-support units are removed while holding the case and the ASR/OCR deletion counts fixed. For random draw j of case i, let

$$q_{ij} = \frac{|R_{ij} \cap S_i^{\text{obs}}|}{|S_i^{\text{obs}}|},$$

where R_{ij} is the realized random removal target. All 1,412 unique random targets match the case-specific audited ASR/OCR deletion counts exactly. Mean q_{ij} is 0.653 for DeepSeek and 0.660 for GPT; the corresponding medians are 0.688 [0.481, 0.864] and 0.714 [0.500, 0.864]. Within-case variation in q_{ij} is available for 70/81 DeepSeek cases and 75/90 GPT cases; 43 and 36 cases, respectively, vary in both q_{ij} and removal outcome.

Using a case fixed-effects linear probability model, each additional 10 percentage points of audited-support coverage is associated with an 8.68-point increase in removal-shift probability for DeepSeek (95% CI [6.72,10.45]) and a 6.72-point increase for GPT ([4.42,8.93]). Confidence intervals use 10,000 case-cluster bootstrap replicates, and 10,000 within-case outcome permutations give $p_{\text{emp}} = 0.00010$ for both models. Table 25 summarizes the analysis. These estimates describe within-case association under a matched ASR/OCR unit-count budget; they do not identify a causal effect of audited-support coverage.

Adjustment for realized deletion length. Because same-count random targets can still differ in realized text length, we additionally include the deleted-word fraction as a within-case covariate. Table 26 reports the resulting estimates. The association with audited-support coverage remains

Table 26: **Within-case coverage–response association after adjusting for deleted text length.** Effects are reported per +10 percentage points of the corresponding predictor.

Model	Coverage effect (pp)	Length effect (pp)
DeepSeek-V4-Pro	**+8.35** [5.96, 10.38]	+1.33 [−1.31, 3.98]
GPT-5.6-sol	**+5.65** [2.82, 7.94]	+4.73 [0.96, 8.64]

Table 27: **Removal response by source-target coverage of audited text support.** RD is audited-support minus source removal.

Coverage	n	Source shift (%)	Audited shift (%)	RD (pp)	95% CI
DeepSeek-V4-Pro					
$< 50\%$	27	22.22	92.59	**+70.37**	[51.85, 85.19]
50–$< 100\%$	41	39.02	92.68	+53.66	[36.59, 70.73]
100%	13	84.62	100.00	+15.38	[0.00, 38.46]
GPT-5.6-sol					
$< 50\%$	30	10.00	83.33	**+73.33**	[53.33, 90.00]
50–$< 100\%$	46	39.13	91.30	+52.17	[36.96, 67.39]
100%	15	93.33	93.33	0.00	[−20.00, 20.00]

positive after this adjustment. Per additional 10 percentage points of support coverage, the estimated response increases by +8.35 points for DeepSeek (95% CI [5.96,10.38]) and +5.65 points for GPT ([2.82,7.94]). The corresponding deleted-length coefficients are +1.33 points ([-1.31,3.98]) and +4.73 points ([0.96,8.64]) per 10 percentage points of deleted-word fraction.

This remains a within-case association rather than a causal estimate. The result shows that variation in realized deletion length does not account for the observed relation between support coverage and removal response.

Coverage–response association. We next ask whether the paired difference between source and audited-support removal varies with source-target coverage. The analysis uses the same frozen audit labels and model outcomes; no new inference or annotation is introduced. Table 27 reports the three fixed coverage strata used in the main text.

The fixed strata show the same ordering for both models: the paired gap is largest when less than half of audited support is covered and smallest under full coverage. Full coverage means that all audited support lies within the source target; it does not imply that the source and audited removal targets are identical.

We also fit a descriptive linear risk-difference trend using the uncovered-support fraction. Per additional 25 percentage points of uncovered support, the paired gap increases by +13.42 points for DeepSeek (95% CI [5.65,20.39]) and +16.98 points for GPT ([10.01,23.09]). Adding source-dataset fixed effects gives +12.47 points ([4.25,20.45]) and +13.64 points ([6.16,20.67]), respectively. Excluding empty audited-removal inputs leaves adjusted estimates of +12.66 points ([3.63,21.41], $n = 75$) for DeepSeek and +12.88 points ([4.62,20.44], $n = 83$) for GPT. Figure 5 summarizes the continuous estimates.

We interpret these associations descriptively because the audited support sets are observation-bounded human judgments rather than estimates of the model's causal features.

Eight of the 120 audited-support payloads are empty after deletion. On the model-specific primary denominators, excluding these cases yields 75 DeepSeek cases and 83 GPT cases. Source shift is 38.67% versus 93.33% audited shift for DeepSeek, a +54.67-point difference (95% CI [42.67,66.67]); for GPT the corresponding rates are 36.14% and 87.95%, a +51.81-point difference ([39.76,63.86]).

Word-weighted source-target coverage. The unit-level coverage measure gives equal weight to each ASR/OCR unit, although units differ in length. We therefore recompute source-target coverage by weighting each audited support unit by its word count. Among the same 99 support-defined

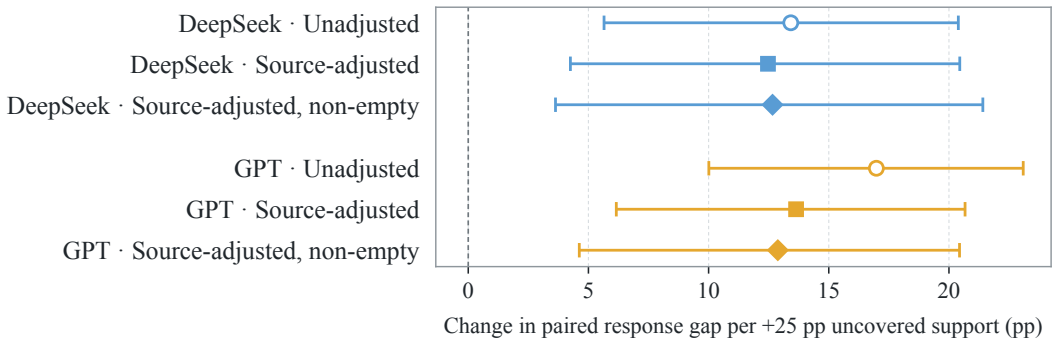


Figure 5: **Continuous association between uncovered support and the source-to-audited response gap.** Points show the estimated change in paired response gap for each additional 25 percentage points of audited support outside the source target. Error bars show 95% case-level bootstrap confidence intervals.

Table 28: **Consensus-only text support removal and matched random control.** Consensus targets contain only units independently labeled SUPPORT by both annotators before adjudication. Random targets match the consensus ASR and OCR unit counts exactly.

		Removal response (%)			Differences (pp)	
Setting	n	Source	Consensus	Random	Consensus − source	Consensus − random
DeepSeek-V4-Pro						
Primary	80	40.00	**88.75**	29.85	+48.75 [37.50, 60.00]	**+58.90** [49.77, 67.78]
Non-empty	75	38.67	**88.00**	25.17	+49.33 [37.33, 61.33]	**+62.83** [54.00, 71.56]
GPT-5.6-sol						
Primary	89	38.20	**87.64**	26.68	+49.44 [37.08, 60.67]	**+60.97** [52.39, 69.46]
Non-empty	84	38.10	**86.90**	22.31	+48.81 [36.90, 60.71]	**+64.59** [56.10, 72.79]

cases, the median word-weighted coverage is 63.74% [IQR 27.68,90.97], compared with 69.23% [36.11,90.40] under unit counting; 18/99 cases remain fully covered. The coverage–response relation is also preserved under this weighting. Per additional 25 percentage points of uncovered word-weighted support, the source-adjusted paired gap increases by +12.17 points for DeepSeek (95% CI [4.48,19.85]) and +14.96 points for GPT ([7.63,21.72]). Excluding empty audited-removal observations gives +12.30 points ([4.06,20.49]) and +14.28 points ([6.30,22.00]), respectively. Thus, the coverage pattern is not specific to treating differently sized text units equally.

Consensus-only support sensitivity. As an annotation-rule sensitivity, we restrict the audited text target to the strict intersection of the two independent pre-adjudication annotations. Specifically, a unit enters $S_i^{\cap}$ only when both annotators label it SUPPORT. The resulting target contains 1,404 units, including 1,017 ASR and 387 OCR units, across 95 of the 120 audit cases. The source target covers a median 66.67% [38.75,92.26] of this consensus support, and 19/95 cases are fully covered. These values closely track the adjudicated-support result of 69.23% [36.11,90.40] median coverage and 18/99 fully covered cases. Across the consensus target, 783 support units lie inside the source target and 621 lie outside it. We then repeat the support-removal analysis using $S_i^{\cap}$ while keeping the frozen original observations, source-removal comparator, model prompts, and inference settings unchanged. Random targets remove exactly the same numbers of ASR and OCR units as the consensus target for each case and are frozen before inference. Five of the 120 consensus-support-removal inputs are empty; the primary analysis retains them under the same protocol, while the non-empty sensitivity excludes them. Table 28 summarizes the resulting comparisons. For the primary source-to-consensus comparisons, exact McNemar tests give $p = 3.82 \times 10^{-11}$ for DeepSeek and $p = 3.71 \times 10^{-11}$ for GPT. The corresponding non-empty values are 1.46×10^{-10} and 2.46×10^{-10}. For all four consensus-versus-random comparisons in Table 28, the 10,000-replicate centered randomization test gives $p_{\text{emp}} = 0.00010$. Thus, the source-to-target and target-to-random orderings persist when support is restricted to units independently labeled SUPPORT by both annotators.

Table 29: **Literal common-base text sensitivity.** Both projected source removal and audited-support removal are constructed by deleting items from the same frozen original observation. "Original" is the masking-based source removal used in the main analysis.

		Source shift (%)			
Model	n	Original	Projected	Audited (%)	Audited − projected (pp)
Contained projection					
DeepSeek-V4-Pro	80	40.00	25.00	93.75	**+68.75** [58.75, 78.75]
GPT-5.6-sol	90	38.89	28.89	88.89	**+60.00** [48.89, 70.00]
Overlap projection					
DeepSeek-V4-Pro	81	40.74	56.79	93.83	**+37.04** [25.93, 48.15]
GPT-5.6-sol	91	38.46	45.05	89.01	**+43.96** [32.97, 54.95]

The within-case association also persists under this stricter target. Among matched random draws, each additional 10 percentage points of consensus-support coverage is associated with a +9.32-point increase in removal-shift probability for DeepSeek (95% CI [7.08,11.30]) and a +5.92-point increase for GPT ([3.16,9.67]), with $p_{\text{emp}} = 0.00010$ for both. The fixed coverage strata show the same descriptive ordering: for DeepSeek, the source-to-consensus gap decreases from +69.23 points below 50% coverage to +13.33 points at full coverage; for GPT it decreases from +79.31 to +5.88 points. In the continuous source-adjusted analysis, each additional 25 percentage points of uncovered consensus support is associated with a +13.94-point larger paired gap for DeepSeek (95% CI [5.67,21.84]) and +19.14 points for GPT ([12.30,26.31]). The corresponding non-empty estimates are +13.98 and +18.63 points.

On cases valid under both annotation rules, adjudicated-support removal changes 93.75% of DeepSeek predictions compared with 88.75% under the consensus target, and 89.89% of GPT predictions compared with 87.64%. The consensus targets are also similar in realized text length to their matched random targets: median deleted word counts are 166.0 [IQR 68.0,382.5] and 170.0 [71.0,394.0], respectively. Taken together, these analyses show that the coverage and removal-response patterns persist when the support target excludes units introduced only through adjudication. As elsewhere, this sensitivity remains observation-bounded and does not treat annotator agreement as complete or causal model evidence.

Literal common-base text sensitivity. The primary source-removal comparator is produced by temporal masking followed by normalization and rechunking, whereas audited-support removal deletes frozen items directly from the original observation. We therefore construct two supplementary source projections in which both source and audited targets are deleted literally from the same frozen item payload.

A *contained* projection removes only items fully contained in the source interval, while an *overlap* projection removes any item whose temporal extent overlaps that interval. These projections differ only in their treatment of boundary-crossing items and are not exact substitutes for the original word-level temporal mask. Across the 120 audit cases, 237 item incidences overlap a source boundary.

Table 29 shows that the source-to-audited gap persists under both constructions. Under the contained projection, the paired differences are +68.75 points for DeepSeek and +60.00 points for GPT; under the overlap projection, they remain +37.04 and +43.96 points, respectively. Thus, the source-to-audited difference is not attributable solely to the use of different input-materialization pipelines.

Most projected source targets still omit at least one audited-support unit, so these controls do not imply equivalence between the source and audited targets. Their purpose is narrower: the response gap remains when the two conditions share the same literal item base.

C.9 Incremental Residual-Support Removal

Starting from the source removal payload, we additionally delete the harmfulness-relevant residual ASR/OCR units identified in the residual-support audit. This tests the additional behavioral change obtained when identified support remaining after source removal is also withheld.

Table 30: **Targeted residual-support removal with exact matched-random control.** "Source" is temporal source removal; random and targeted deletion additionally remove the same numbers and types of units. Shifts are in %; RD is targeted minus random.

Model	n	Source	Random	Targeted	RD (pp)	95% CI	McNemar p
DeepSeek-V4-Pro	69	20.29	21.74	**73.91**	**+52.17**	[40.58, 63.77]	2.84×10^{-10}
GPT-5.6-sol	75	13.33	13.33	**58.67**	**+45.33**	[33.33, 57.33]	1.08×10^{-9}

Table 31: **Coverage of observation-bounded visual support by the source target.** Coverage is defined only for cases with at least one adjudicated support frame; values are medians with interquartile ranges.

Source	n	Target coverage (%)	Outside target (%)	Fully covered
HateClipSeg	43	68.75 [31.25, 80.63]	31.25 [19.38, 68.75]	2/43 (4.65%)
HateMM	40	79.51 [55.82, 87.50]	20.49 [12.50, 44.18]	3/40 (7.50%)
Overall	**83**	**75.00** [41.88, 81.25]	**25.00** [18.75, 58.13]	**5/83 (6.02%)**

On all eligible clear-residual cases, removal-sensitive shift rises from 19.47% to 78.76% for DeepSeek and from 13.33% to 69.17% for GPT. Table 30 reports the exact matched-random comparison.

Exact source-type- and count-matched random deletion is feasible for 69 of 113 DeepSeek and 75 of 120 GPT clear-residual cases. The remaining cases are structurally unmatchable because the removal payload does not contain enough non-target ASR and/or OCR units to construct a disjoint matched deletion; matchability is determined without reference to model outcomes. Because the matched random condition removes the same numbers and source types of text units, matched deletion count and ASR/OCR composition alone do not account for the larger targeted response on these matched subsets.

Matchability is structurally associated with observation density. For DeepSeek, the median total text-unit count is 16 in matchable cases versus 3.5 in unmatchable cases; for GPT the corresponding medians are 19 and 3. The targeted residual units therefore occupy a smaller median fraction of the available text in matchable cases (16.13% versus 56.35% for DeepSeek and 15.38% versus 57.14% for GPT). Despite these structural differences, the source-removal shift is similar across the two groups: 20.29% versus 18.18% for DeepSeek and 13.33% versus 13.33% for GPT. We therefore treat the matched-random result as a subset analysis: feasibility selects for denser observations, while the observed baseline source-removal rates are close across matchability groups. No separate inferential test was prespecified for this descriptive comparison.

C.10 Visual Observation-Bounded Support Audit

A separate visual audit contains 120 videos, balanced between HateClipSeg and HateMM. Each case uses the frozen evidence-agnostic U_{16} observation. Two annotators independently label every frame as SUPPORT, NOT SUPPORT, or UNCERTAIN while blinded to E_i, model predictions, intervention outcomes, and the earlier residual-support labels. Disagreements are adjudicated after the independent labels are frozen. As in the text audit, these labels characterize support within the frozen sampled observation rather than exhaustive support in the underlying raw video.

Across 1,920 frame units, three-way raw agreement is 74.06% with $\kappa = 0.4205$; support-versus-other agreement is 74.11% with $\kappa = 0.4099$. The mean support-set Jaccard is 0.7106 and the median is 1.000 [0.1875,1.000]. The adjudicated labels contain 1,136 SUPPORT, 443 NOT SUPPORT, and 341 UNCERTAIN frames. Of 498 initial disagreement frames, 340 remain unresolved after adjudication and are retained as UNCERTAIN; one additional frame was independently labeled UNCERTAIN by both annotators.

Eighty-three of 120 cases contain at least one adjudicated support frame. Table 31 reports how much of that support lies inside the source temporal target.

The low rate of full source-target coverage is stable across the two independent annotations. Using Annotator 1 alone gives a median coverage of 75.00%, with 3/89 support-defined cases fully

Table 32: **Source temporal versus broader audited visual-support removal.** The raw broader condition frequently exhausts U_{16} and is therefore accompanied by the non-empty sensitivity below.

Model	n	Source (%)	Broader (%)	RD (pp)	95% CI
Qwen3.6-27B	67	25.37	97.01	+71.64	[59.70, 82.09]
GLM-4.5V	66	19.70	95.45	+75.76	[65.15, 86.36]

Table 33: **Non-empty visual matched-random sensitivity.** Starting from source removal, targeted deletion removes the audited residual support; random deletion removes the same number of complement frames. Random rates are case-weighted expectations over all feasible combinations or up to 20 frozen unique draws.

Source	n	Targeted (%)	Exp. random (%)	RD (pp)	95% CI	Empirical p
Qwen3.6-27B						
Overall	11	**81.82**	28.81	**+53.01**	[25.82, 76.78]	.000200
HateClipSeg	7	71.43	28.23	+43.20	[7.24, 75.20]	.00430
HateMM	4	100.00	29.82	+70.18	[47.50, 85.36]	.00400
GLM-4.5V						
Overall	10	**70.00**	22.76	**+47.24**	[20.83, 71.55]	.000200
HateClipSeg	6	66.67	15.44	+51.23	[15.12, 84.01]	.00420
HateMM	4	75.00	33.75	+41.25	[12.50, 70.00]	.05289

covered; Annotator 2 gives 68.75%, with 3/92 fully covered. Restricting support to frames labeled SUPPORT by both annotators gives a median of 75.00%, with 5/78 fully covered. The corresponding adjudicated values are 75.00% and 5/83. Thus, the conclusion that full visual support coverage is uncommon does not depend on the adjudication rule.

Audited support extends outside the source temporal target in 78/83 (93.98%) support-defined cases. Unlike the text construction above, the visual broader-removal target is explicitly defined as the union of the source-covered U_{16} frames and all adjudicated support frames. On model-specific original-HARMFUL, support-defined common-valid cases, source removal and broader removal yield the results in Table 32.

The broader-removal payload is empty for 53/67 Qwen cases and 53/66 GLM cases. All 53 empty Qwen payloads yield UNCERTAIN, while all 53 empty GLM payloads yield NOT_HARMFUL. Because empty inputs dominate this comparison, we focus on the non-empty sensitivity below. Excluding empty broader-removal inputs leaves 14 Qwen and 13 GLM cases: source versus broader shift is 35.71% versus 85.71% for Qwen (RD +50.00 points, 95% CI [21.43,78.57], McNemar $p = .015625$) and 30.77% versus 76.92% for GLM (RD +46.15 points, [23.08,76.92], $p = .03125$).

We further control for the amount of additional deletion. Restricting to cases with support outside the source target and a feasible same-count random additional deletion gives 64 Qwen and 63 GLM cases overall. Because the empty cases contribute identical 100% shift to both targeted and random conditions, we report the non-empty matched-random sensitivity separately in Table 33.

All broader transitions in the overall non-empty matched-random sensitivity are $H \rightarrow$ NOT_HARMFUL; none are $H \rightarrow$ UNCERTAIN. Source cells are very small and are reported only as sensitivity checks. The overall confidence intervals use 5,000 paired video-level bootstrap replicates and the empirical p-values use 10,000 randomization pseudo-experiments with seed 20260904.

Exact-support-only visual feasibility. Because the broader visual target above also includes all source-covered frames, we additionally examine whether deleting only the adjudicated SUPPORT frames would provide a larger non-empty comparison set. This check uses the existing frozen annotation and requires no additional human labeling. Among model-specific original-HARMFUL, support-defined cases, exact-support-only deletion leaves a non-empty payload in 19/67 Qwen cases and 18/66 GLM cases, compared with 14 and 13 under the broader union target. Only 14 cases for each model retain at least four frames.

These cohorts remain below the frozen minimum of 30 non-empty cases set for running the additional behavioral comparison. We therefore do not execute this model intervention and report

the feasibility result only. The visual audit remains corroborating evidence rather than the primary controlled support-removal analysis.

Consensus-only visual feasibility. As a higher-agreement sensitivity, we further restrict visual support to frames independently labeled SUPPORT by both annotators. This consensus definition identifies support in 78/120 cases, with median source-target coverage of 75.00% [39.06,81.25] and 5/78 cases fully covered. Among model-specific original-HARMFUL cases, 61 cases per model contain consensus support outside the source target. After additionally removing this residual support, however, only 11 Qwen and 10 GLM cases retain non-empty observations, with 10 and 9 retaining at least four frames. Because both cohorts remain below the frozen minimum of 30 non-empty cases, we do not run an additional behavioral intervention. This feasibility check shows that the limited non-empty visual cohort persists even under a higher-agreement support definition.

D ADDITIONAL VALIDITY CHECKS FOR REMOVAL FIDELITY

This section examines validity questions surrounding removal-based interpretation: the relation of removal fidelity to rationale completeness and comprehensiveness, both directions of target correspondence in the frozen text audit, the cases without audited support, reweighting of the source-balanced audit to its sampling frame, the effect of changing the removal target, and whether the preservation–removal pattern persists under joint visual and textual input. No additional human annotation is introduced.

D.1 REMOVAL FIDELITY, COMPLETENESS, AND COMPREHENSIVENESS

Removal fidelity is related to, but distinct from, rationale completeness and deletion comprehensiveness. Completeness asks whether an annotated rationale covers the support relevant to a judgment; comprehensiveness measures the behavioral response obtained after deleting a designated rationale (DeYoung et al., 2020). Removal fidelity instead concerns the correspondence between the information an intervention actually withholds and the support scope to which the resulting prediction change is interpreted. It is therefore a property of the pairing between an intervention and a claim rather than of the annotation alone.

This distinction matters when a temporal annotation collected to identify valid evidence is later reused as a deletion target. Source-target removal validly measures sensitivity to that source-defined target even when other support for the same judgment remains available elsewhere in the observation. The problem arises only when the response is interpreted more broadly, for example as evidence that the model is insensitive to harmful support in the observation as a whole; under that interpretation, the intervention must withhold information that corresponds to the broader claim target.

The controlled text interventions provide an empirical example. On the model-specific matched-random cohorts of Table 1, source-target removal changes the prediction in 40.74% of DeepSeek cases and 38.89% of GPT cases. Removing audited support on the same cases yields responses of 93.83% and 88.89%, whereas same-count random deletion remains at 34.19% and 31.66% (Table 37). The source-removal response thus characterizes sensitivity to the source target but does not by itself support the broader conclusion that the model is similarly insensitive to harmful support in general.

D.2 DIRECTIONAL TARGET CORRESPONDENCE

We next examine how the source target corresponds to the support identified in the frozen 120-case text audit. As in Section 3.2, let S_i^{obs} denote the audited support units of case i and E_i^O the frozen original audit units whose timestamps fall inside the source target. For cases with nonempty audited support,

$$\mathrm{Recall}_i = \frac{|S_i^{\mathrm{obs}} \cap E_i^O|}{|S_i^{\mathrm{obs}}|}$$

measures how much audited support falls inside the source target; this is the coverage C_i^{src} reported in the main paper. The complementary direction asks how much of the source-target membership

corresponds to audited support:

$$\text{Precision}_i^{\text{mem}} = \frac{|S_i^{\text{obs}} \cap E_i^O|}{|E_i^O|}, \qquad \text{Collateral}_i^{\text{mem}} = 1 - \text{Precision}_i^{\text{mem}},$$

$$\text{Jaccard}_i^{\text{mem}} = \frac{|S_i^{\text{obs}} \cap E_i^O|}{|S_i^{\text{obs}} \cup E_i^O|}.$$

The superscript marks a scope limitation. These quantities are computed from source-interval membership on the frozen original audit units; they are not exact realized-deletion precision or collateral-deletion metrics. The text intervention temporally masks the observation and then rechunks the remaining ASR/OCR, so the realized post-removal unit representation differs from literal deletion of the membership-tagged units in 90/120 cases, and no frozen alignment rule maps the rechunked units back to the original audit units. We therefore use membership precision, collateral deletion, and Jaccard only as descriptive proxies for this direction of correspondence.

Table 34 reports the resulting distributions. Audited support is present in 99/120 cases; across them, the median support recall is 69.23%, with substantial variation across sources. Precision, collateral deletion, and Jaccard additionally require a nonempty source target and are defined for 94 support-positive cases.

Table 34: **Directional correspondence between source targets and audited text support.** Entries are medians with interquartile ranges. Recall is the fraction of audited support inside the source target (coverage). Membership precision, collateral deletion, and Jaccard are descriptive proxies computed on frozen original audit units, not exact realized-deletion metrics; they additionally require a nonempty source target.

Source	Recall n	Recall (%)	Proxy n	Membership precision (%)	Collateral deletion (%)	Jaccard (%)
HateClipSeg	33	44.90 [12.50, 69.35]	30	51.32 [29.12, 67.98]	48.68 [32.02, 70.88]	25.00 [6.45, 50.00]
HateMM	37	75.00 [38.89, 90.91]	35	84.21 [46.26, 98.89]	15.79 [1.11, 53.74]	50.00 [25.00, 80.00]
MultiHateClip	29	80.00 [60.00, 100.00]	29	73.68 [50.00, 85.71]	26.32 [14.29, 50.00]	50.00 [40.00, 73.68]
Overall	**99**	**69.23** [36.11, 90.40]	**94**	**66.67** [43.51, 89.22]	**33.33** [10.78, 56.49]	**44.44** [21.24, 66.67]

The source-removal response is also associated with this coverage direction across cases. On the original-HARMFUL common-valid source-removal cohorts, shift rates increase across quartiles of source-target support recall: 19.05%, 35.00%, 40.00%, and 70.00% for DeepSeek-V4-Pro ($n = 81$) and 8.70%, 26.09%, 40.91%, and 78.26% for GPT-5.6-sol ($n = 91$). A descriptive linear risk-difference model adjusting for the number of whitespace-delimited words deleted by the source target estimates a change of +6.91 percentage points [3.81, 9.63] for DeepSeek and +9.27 points [7.13, 11.10] for GPT per +10 percentage points of support recall. These are observational associations, not causal effects, and associations involving membership precision inherit the proxy limitation above.

D.3 Cases Without Audited Observation-Bounded Support

The audit identified no observation-bounded SUPPORT unit in 21/120 cases. Coverage is undefined for these cases rather than zero, so they are kept separate from the support-recall analysis. As Table 35 shows, neither text model predicts HARMFUL on any of these 21 frozen original inputs: both produce 17 NOT_HARMFUL and 4 UNCERTAIN judgments. These cases therefore do not enter the original-HARMFUL removal-shift denominator and cannot account for the weak source-removal response measured on that endpoint.

All 21 cases receive valid original and source-removal predictions in the complete text-track triads (Table 11); what they lack is an original-HARMFUL prediction. They therefore enter neither the original-HARMFUL audited-support intervention cohort nor the corresponding exact-input repeatability cohort, and because the removal-shift endpoint is defined as a transition from an original

Table 35: **Original predictions for the 21 cases with no audited observation-bounded support.** Coverage is undefined for this subset.

Model	n	Original H	Original NH	Original U
DeepSeek-V4-Pro	21	0	17	4
GPT-5.6-sol	21	0	17	4

HARMFUL prediction to NOT_HARMFUL or UNCERTAIN, that endpoint is undefined for this subset rather than 0%. The serialized audited-removal input is identical to the original input in all 21 cases, since there is no audited support to remove; this is an input-construction check, not an additional set of audited-removal verdicts. These cases establish only that the audit identified no observation-bounded support under the frozen procedure, and do not distinguish among sparse observations, support missed by the audit, and stronger prior-driven prediction.

D.4 Audit Denominators and Population Reweighting

The source-balanced audit contains 120 cases, 40 from each source. Among them, 99 contain nonempty audited support and 18 have all audited support inside the source target. The fully covered rate is therefore $18/120 = 15.00\%$ among all audited cases and $18/99 = 18.18\%$ among support-positive cases; the main paper reports the latter. A larger count of 39/120 arises if empty-support cases are treated as fully covered by vacuous truth; we do not use that quantity.

Because the 120 cases were sampled equally across sources, the source-balanced summary is not a prevalence estimate for the text collection. The direct sampling frame is the frozen 230-case text residual-support audit (83 HateClipSeg, 95 HateMM, and 52 MultiHateClip cases; Table 17). We therefore also compute source-stratified weights $w_s = N_s/n_s$, where N_s is the number of cases from source s in the 230-case frame and $n_s = 40$ is its audited sample size, with uncertainty from 10,000 source-stratified bootstrap draws. Table 36 compares the source-balanced statistics with the reweighted estimates. Reweighting changes the median support recall from 69.23% to 66.67% and the fully covered rate among support-positive cases from 18.18% to 16.34%, leaving the qualitative conclusion unchanged. These estimates apply to the 230-case frame and are not estimates for all 963 text videos.

Table 36: **Source-balanced and population-reweighted text-audit statistics.** All values are percentages. Reweighted estimates target the frozen 230-case text-audit sampling frame (83 HateClipSeg, 95 HateMM, 52 MultiHateClip), not the full 963-video collection; the last column is the 95% source-stratified bootstrap interval of the reweighted estimate. Starred rows are membership-based proxies (Section D.2).

Quantity	Source-balanced audit	230-frame reweighted	95% CI
Support present rate	82.50	84.37	[77.79, 90.37]
Mean support recall	60.67	59.02	[52.32, 65.54]
Median support recall	69.23	66.67	[52.94, 76.19]
Fully covered, all cases	15.00	13.78	[8.09, 19.95]
Fully covered, support-positive	18.18	16.34	[9.60, 23.62]
Mean membership precision*	63.45	63.30	[57.00, 69.36]
Mean collateral deletion*	36.55	36.70	[30.64, 43.00]
Mean Jaccard*	44.74	43.92	[38.22, 49.66]

D.5 Changing the Removal Target Changes the Behavioral Conclusion

We next compare three interventions on the same model-specific matched-random cohorts: source-target removal, audited-support removal, and same-count random deletion. Table 37 reports the resulting shift rates. The source-target intervention yields a valid measurement of response to that source target, but a substantially different conclusion follows when the intervention is expanded to audited support. The difference is not explained by deletion quantity alone: same-count random deletion remains much closer to source-target removal than to audited-support removal. This

comparison motivates treating removal fidelity as a property of the correspondence between an intervention and the scope of the claim drawn from it, rather than as a synonym for the magnitude of a deletion response.

Table 37: **Source-target, audited-support, and same-count random removal on the matched-random cohorts.** Rates are computed on each model's common-valid cohort of Table 1; the random column averages the frozen draws within case before aggregating cases. p_{emp} compares audited-support removal with the same-count random control.

Model	n	Source (%)	Audited (%)	Random (%)	Audited − source (pp)	Audited − random (pp)	p_{emp}
DeepSeek-V4-Pro	81	40.74	**93.83**	34.19	+53.09	**+59.63**	.00010
GPT-5.6-sol	90	38.89	**88.89**	31.66	+50.00	**+57.23**	.00010

D.6 Joint Multimodal Robustness

The primary experiments analyze visual and textual observations separately so that each intervention is controlled within a fixed observation space. We additionally test whether the preservation–removal pattern persists when both modalities are supplied jointly to one multimodal model. We use the frozen 120-video visual-audit subset (60 HateClipSeg and 60 HateMM videos). For each case, Qwen3.6-27B receives the frozen evidence-agnostic U_{16} frames together with the frozen aligned ASR/OCR in a single input; all 120 input partitions pass literal-subset validation. The source target is applied synchronously across modalities: the source-only condition retains source-target visual and textual content, and the source-removal condition withholds content inside the source target from both modalities.

Qwen returns 119 strict-valid triads; one source-removal response is invalid JSON, and its original prediction is `NOT_HARMFUL`. Among the 106 complete original-`HARMFUL` cases, source-only preservation is 95.28% and source removal changes 19.81% of predictions, a PR gap of +75.47 points [66.98, 83.96] under 5,000 case-level bootstrap draws (Table 38). The separation is therefore not limited to experiments that evaluate the two channels separately. It remains a source-target result: without a joint multimodal support audit, it does not establish that source removal exhausts the support available across both modalities.

Table 38: **Joint multimodal preservation and removal.** Frozen U_{16} frames and aligned ASR/OCR are supplied jointly to Qwen3.6-27B; rates condition on complete original-`HARMFUL` cases. The result characterizes response to the source target under a joint observation and does not assume that all multimodal support is removed.

Model	n	Source-only pres. (%)	Source removal (%)	PR gap (pp) [95% CI]
Qwen3.6-27B	106	95.28	19.81	+75.47 [66.98, 83.96]

Exploratory synchronized temporal placebo. We also construct a same-duration temporal placebo in which one randomly shifted interval, fixed before inference, is removed synchronously from both modalities. The comparator is structurally feasible for 32/120 cases (20 HateClipSeg and 12 HateMM), of which 22 are common-valid and original-`HARMFUL`. On these 22 cases, source removal changes 0/22 predictions and the placebo changes 1/22, a paired difference of -4.55 points [-13.64, 0.00] (Table 39). The placebo matches temporal duration rather than realized deletion content: among the 32 feasible cases, the numbers of removed frames match in 18 and the numbers of removed text units in only 3, and 18 source targets contain multiple intervals whereas the placebo uses one. This restricted single-draw comparison provides no evidence of a stronger source-specific response and is not used to support that claim. No joint exact-input no-op repeat was run.

Summary. These analyses sharpen the scope of the removal-fidelity claim. Support coverage captures one directional failure mode, relevant support remaining outside the intervention, but does not exhaust the correspondence between intervention and claim target; the membership-based quantities describe potential over-removal but are not exact realized-deletion metrics under the text rechunking pipeline. Changing the removal target from the source annotation to audited support changes

Table 39: **Exploratory synchronized temporal placebo.** One randomly shifted interval with the same temporal duration as the source target is removed from both modalities. Feasibility is restricted and the placebo does not match realized visual and textual deletion counts, so the comparison is exploratory rather than a primary joint multimodal control.

Model	n	Source removal (%)	Placebo (%)	Source − placebo (pp) [95% CI]
Qwen3.6-27B	22	0.00	4.55	−4.55 [−13.64, 0.00]

the behavioral conclusion substantially, while same-count random deletion does not reproduce that response, and the preservation–removal separation persists when frozen visual and textual observations are supplied jointly. We therefore interpret weak source-removal responses as sensitivity to the specified intervention rather than as direct evidence of weak reliance on all harmful support in the observation.

E TRANSFER STUDY: TEMPORAL EVENT GROUNDING

We use 120 Charades-STA test query–video pairs (Gao et al., 2017) and the corresponding Charades v1 480p videos (Sigurdsson et al., 2016). The sample, source temporal intervals, frozen U_{16} observations, human support annotations, intervention targets, and random manifests are shared across models. Qwen3.6-27B and GLM-4.5V independently predict whether the queried event is `PRESENT`, `ABSENT`, or `UNCERTAIN` under the same transfer protocol. The 16-frame observation is constructed uniformly over the full video without access to the source temporal interval and frozen before frames are partitioned by interval membership.

Two annotators independently label each of the 1,920 query-conditioned frames as SUPPORT, NOT SUPPORT, or UNCERTAIN while blinded to interval membership and model outcomes. Three-way agreement is 73.13% with $\kappa = 0.4937$, and support-versus-other agreement is 78.18% with $\kappa = 0.5450$. Mean support-set Jaccard is 0.5672 and the median is 0.600 [0.2875,0.875]. Of 516 initial disagreements, 190 remain unresolved after adjudication and are retained as UNCERTAIN.

Support is identified in 110/120 cases. The source interval covers a median 66.67% [31.77,100.00] of the observation-bounded event support. Sixty-eight of 110 support-defined cases (61.82%) contain support outside the source interval, while 42/110 (38.18%) are fully covered. These coverage statistics are fixed by the shared observation and annotation protocol and therefore do not depend on the evaluated model.

For this transfer, the broader removal target is the union of source-covered U_{16} frames and the adjudicated observation-bounded support frames. This construction preserves the source target while additionally withholding identified support outside it. In 89/120 cases, the source component adds at least one frame beyond the adjudicated support set. In the remaining 31, the union coincides with the support set.

Source versus broader-target removal. Model-specific analyses condition on valid original-`PRESENT` cases under the same frozen protocol. For Qwen3.6-27B, all 360 original, evidence-only, and source-removed base predictions are valid. The original prediction is `PRESENT` for 84/120 cases and `ABSENT` for 36/120, with no `UNCERTAIN` predictions. Evidence-only preservation among the 84 original-`PRESENT` cases is 62/84 (73.81%). Source-interval removal changes 53.57% of predictions, whereas broader-target removal changes 84.52%, a paired +30.95-point difference.

For GLM-4.5V, the primary paired cohort contains 107 original-`PRESENT` cases, with evidence-only preservation of 94/107 (87.85%). Source-interval removal changes 37.38% of predictions, whereas broader-target removal changes 60.75%, a paired +23.36-point difference. The same directional source-to-broader increase therefore appears under both evaluated visual models. Table 40 summarizes the model-specific primary and non-empty comparisons.

The broader-removal input is structurally empty for 7/120 cases. Five of these fall within each model's primary original-`PRESENT` denominator, leaving 79 non-empty Qwen cases and 102 non-empty GLM cases. For GLM, evidence-only preservation remains 89/102 (87.25%) on the non-empty cohort and 85/97 (87.63%) when at least four frames remain.

Table 40: **Charades-STA source versus broader-target removal across two visual models.** RD is broader-target minus source-interval removal. Model-specific denominators condition on valid original-PRESENT cases.

Setting	n	Source (%)	Broader (%)	RD (pp)	95% CI	McNemar p
Qwen3.6-27B						
Primary	84	53.57	84.52	**+30.95**	[21.43, 41.67]	2.98×10^{-8}
Non-empty	79	55.70	83.54	+27.85	[17.72, 37.97]	4.77×10^{-7}
GLM-4.5V						
Primary	107	37.38	60.75	**+23.36**	[14.95, 31.78]	4.17×10^{-7}
Non-empty	102	38.24	59.80	+21.57	[13.73, 30.39]	2.98×10^{-6}
Retained ≥ 4 frames	97	40.21	59.79	+19.59	[11.34, 27.84]	2.10×10^{-5}

Table 41: **Same-frame-count random control for the Charades-STA transfer.** Random targets remove the same number of frozen U_{16} frames as the broader target. RD is broader-target minus random removal. Denominators are model-specific common-valid cohorts.

		Shift (%)					
Setting	n	Source	Broader	Random	RD (pp)	95% CI	p_{emp}
Qwen3.6-27B							
Primary	82	53.66	**84.15**	33.29	**+50.85**	[41.28, 60.12]	$< .001$
Non-empty	77	55.84	83.12	28.96	+54.16	[44.28, 63.70]	$< .001$
Retained ≥ 4 frames	72	58.33	81.94	27.22	+54.72	[44.51, 64.58]	$< .001$
GLM-4.5V							
Primary	107	37.38	**60.75**	14.14	**+46.61**	[37.66, 55.40]	.00010
Non-empty	102	38.24	59.80	10.91	+48.90	[39.80, 58.01]	.00010
Retained ≥ 4 frames	97	40.21	59.79	9.85	+49.95	[40.98, 58.97]	.00010

Same-frame-count random control. The source-versus-broader comparison also changes the number of removed frames. We therefore compare broader-target removal with random targets from the same frozen U_{16} observation that remove exactly the same number of frames. The frozen random manifests are shared across models, while behavioral denominators remain model-specific. Random outcomes are averaged within case before aggregation. Table 41 reports the primary, non-empty, and retained-frame comparisons for both models.

For Qwen3.6-27B, broader-target removal changes 84.15% of predictions on the common-valid random-control cohort, compared with a 33.29% same-frame-count random expectation, a +50.85-point difference (95% CI [41.28,60.12], $p_{\text{emp}} < .001$). For GLM-4.5V, broader-target removal changes 60.75% of predictions compared with a 14.14% random expectation, a +46.61-point difference (95% CI [37.66,55.40], $p_{\text{emp}} = 0.00010$). The same ordering therefore replicates across both models despite differences in their source-removal rates.

For Qwen, two original-PRESENT candidates are excluded from the random-control analysis because at least one prespecified random draw is invalid. GLM has no corresponding random-draw exclusion, so its primary source-versus-broader and random-control cohorts both contain 107 cases. No invalid random draw is replaced or resampled.

Transition direction. All Qwen source-removal and broader-target shifts in the primary analysis are PRESENT→ABSENT, with none transitioning to UNCERTAIN. GLM shows the same pattern for the two targeted conditions. Under source removal, 67/107 predictions remain PRESENT and 40/107 (37.38%) change to ABSENT. Under broader-target removal, 42/107 remain PRESENT and 65/107 (60.75%) change to ABSENT, with no transitions to UNCERTAIN in either condition.

For the GLM same-count random control, case-weighted averaging over frozen draws gives an expected 85.86% remaining PRESENT, 14.08% changing to ABSENT, and 0.06% changing to UNCERTAIN. Expected transition counts can be fractional because random outcomes are averaged within case before aggregation rather than pooled across draws.

The Qwen common-valid random-control cohort contains 82 distinct raw videos for 82 query–video pairs, so clustering by raw video does not change the uncertainty calculation. The GLM replication reuses the same frozen cohort, observations, source intervals, human support labels, broader targets, random manifests, and statistical protocol, changing only the evaluated model.

This two-model transfer remains a focused scope check rather than a claim of universality. It nevertheless shows that the source-to-broader target mismatch extends beyond harmful-video annotation and is not specific to a single evaluated model. Across both Qwen3.6-27B and GLM-4.5V, broader-target removal produces more prediction change than source-interval removal and substantially more than same-frame-count random removal. The larger broader-target response is therefore not explained by removed frame count alone.

F Selected-Evidence Analysis

F.1 Selected Evidence Bottleneck Protocol

For each controlled world, the evaluated model first acts as a selector over the full candidate observation and returns exactly two supplied item IDs. The adjudicator then receives only those two contents under fresh opaque IDs. For visual tracks, one unit is one frame; for text tracks, one unit is one normalized ASR/OCR-derived text unit. The matched-random baseline uses one deterministic two-unit draw from the same video and world. It matches unit count, not token count, duration, source composition, or semantic content, and it may overlap with the selector-preferred subset. The locked downstream endpoint is evaluated on a fixed original-`HARMFUL` roster; availability is tracked separately. We therefore interpret SEB as a secondary comparison between selector-preferred and equal-unit-count random evidence, not as a test of optimal or uniquely necessary evidence. Figure 6 and Table 42 summarize the two secondary comparisons.

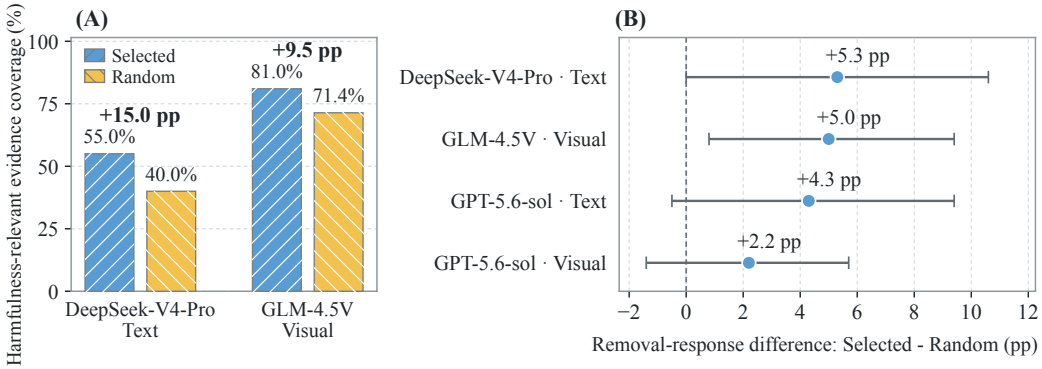


Figure 6: **Evidence selection has a larger effect on evidence presence than on downstream removal response. (A)** In the two settings covered by a blinded auxiliary analysis, selector-preferred subsets more often contain harmfulness-relevant evidence than matched random subsets. **(B)** Across four matched-$k = 2$ evaluations, selected-minus-random downstream response differences are modest. The two panels use separate evaluation subsets and are not pooled.

F.2 Coarse Harmfulness-Relevant Evidence Presence

As a secondary sensitivity analysis, a blinded auxiliary annotation examines whether selector-preferred and random $k = 2$ subsets contain harmfulness-relevant evidence. We use this annotation only to characterize the evidence content of the matched subsets. Table 43 reports the resulting coarse comparison.

F.3 Matched-$k = 2$ Removal Response

Table 44 gives the full paired removal-response statistics, including the exact tests omitted from the main presentation.

Table 42: **Selector-preferred versus random evidence under a matched unit-count budget.** Panel A reports the blinded auxiliary evidence-presence analysis; Panel B reports matched-$k = 2$ downstream removal response. The panels use separate subsets.

Model	Track	n	Selected (%)	Random (%)	Δ (pp)	95% CI
A. Coarse harmfulness-relevant evidence presence						
DeepSeek-V4-Pro	Text	20	55.0	40.0	+15.0	[0.0, 30.0]
GLM-4.5V	Visual	42	81.0	71.4	+9.5	[2.4, 19.1]
B. Matched-$k = 2$ downstream removal response						
DeepSeek-V4-Pro	Text	358	20.1	14.8	+5.3	[0.0, 10.6]
GLM-4.5V	Visual	363	14.3	9.4	+5.0	[0.8, 9.4]
GPT-5.6-sol	Text	393	19.3	15.0	+4.3	[−0.5, 9.4]
GPT-5.6-sol	Visual	418	12.2	10.0	+2.2	[−1.4, 5.7]

Table 43: **Coarse harmfulness-relevant evidence presence under matched $k = 2$.**

Track	n	Selected (%)	Random (%)	RD (pp)
DeepSeek text	20	55.0	40.0	+15.0
GLM visual	42	81.0	71.4	+9.5

Table 44: **Matched-$k = 2$ selected-versus-random downstream removal response.**

Model	Track	n	Selected (%)	Random (%)	RD (pp)	95% CI	p
DeepSeek-V4-Pro	Text	358	20.1	14.8	+5.3	[0.0, 10.6]	.067
GLM-4.5V	Visual	363	14.3	9.4	+5.0	[0.8, 9.4]	.030
GPT-5.6-sol	Text	393	19.3	15.0	+4.3	[−0.5, 9.4]	.107
GPT-5.6-sol	Visual	418	12.2	10.0	+2.2	[−1.4, 5.7]	.298

Table 45: **Weighted coverage sensitivity for matched-$k = 2$ evidence.** Weighted results use the same $n = 147$ unit-role cohort within each track.

Weighting	n	Selected (%)	Random (%)	Δ (pp)
DeepSeek, text				
Unit-level	147	16.11	8.08	+8.03
Word-weighted	147	19.87	8.22	**+11.65** [8.22, 15.34]
GLM, visual				
Unit-level	147	37.78	22.81	+14.97
Duration-union	147	62.79	45.77	**+17.01** [9.24, 24.85]

All four point estimates favor selector-preferred evidence, but only the GLM visual confidence interval excludes zero.

F.4 Selected-Evidence Length- and Duration-Weighted Coverage Sensitivity

The matched-$k = 2$ selected-evidence coverage analysis above counts annotated support units equally. As a granularity sensitivity, we reweight text support by word length and visual support by the union duration of its annotated temporal extent, without changing selectors, random draws, support labels, or model outputs. Before reweighting, the frozen unit-level gate reproduces all four official coverage rows exactly. Table 45 reports the resulting sensitivity.

The selected-over-random direction therefore survives both weighting schemes and is somewhat larger under these size-aware definitions. This sensitivity uses no new model inference and does not change the interpretation of the selected-evidence analysis as secondary.

F.5 Fine-Grained Role-Annotation Reliability

The auxiliary role-annotation reliability study contains 90 examples covering 1,337 evidence units. Raw agreement on the key harmful-role decision is 73.3%, with $\kappa = 0.146$ and F1 $= 0.838$. Exact agreement on the complete role set is 14.4%, and mean role-set Jaccard similarity is 0.485. We therefore keep the selected-evidence interpretation coarse and treat the role analysis as supporting characterization.

F.6 Removal-Outcome Decomposition

Figure 7 decomposes removal-sensitive outcomes into transitions to `NOT_HARMFUL` and `UNCERTAIN`.

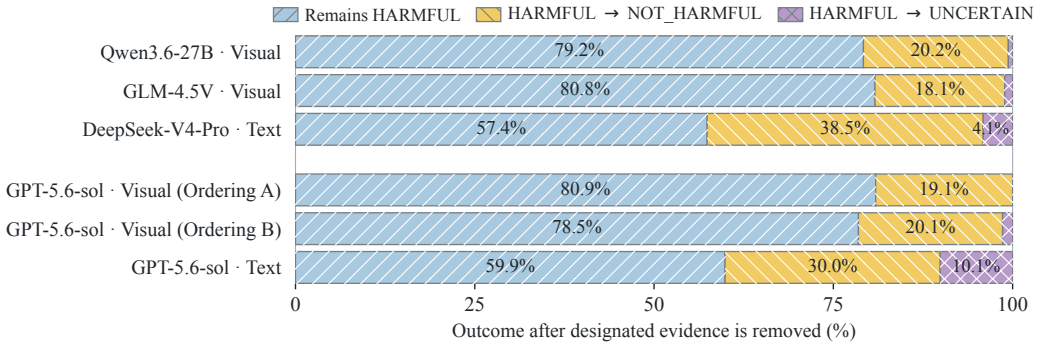


Figure 7: **Removal-sensitive shifts are dominated by transitions to `NOT_HARMFUL` rather than uncertainty.** Bars condition on complete original-`HARMFUL` cases.

Transitions to `UNCERTAIN` account for at most 1.4% of visual cases. The text tracks show more uncertainty, particularly for GPT-5.6-sol, but transitions to `NOT_HARMFUL` remain more frequent.

G Negative-Safety Evaluation

The negative-safety analysis uses separate non-harmful text-observation cohorts and is not part of the harmful-only primary intervention denominator.

G.1 GPT-5.6-sol Text Full-Cohort Results

Figure 8 visualizes the full-cohort negative-safety comparison, and Table 46 gives the exact paired rates and transitions.

Across the two roles, 132 of 1,097 cases predicted `NOT_HARMFUL` under full context transition to `HARMFUL` under selected-only input (12.03%).

G.2 DeepSeek-V4-Pro Text Sentinel

Table 47 reports the corresponding source-balanced 45-video sentinel.

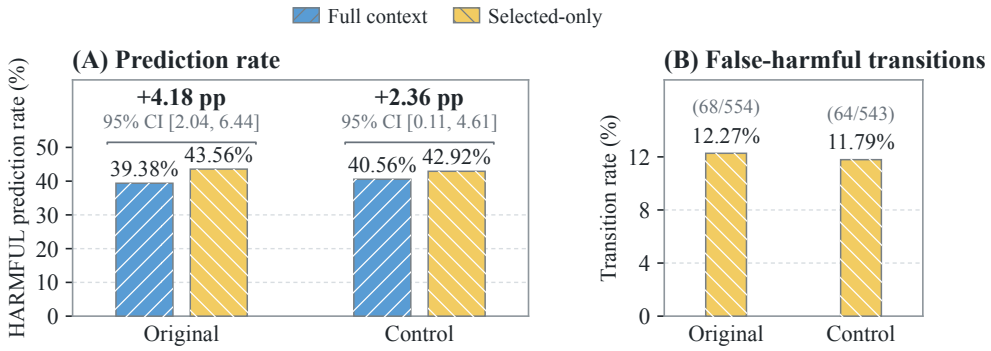


Figure 8: **Selected-only inputs can introduce false-harmful predictions.** Results use the full 932-video GPT-5.6-sol non-harmful text cohort.

Table 46: **GPT-5.6-sol text negative-safety results.** NH→H denotes cases changing from `NOT_HARMFUL` under full context to `HARMFUL` under selected-only input.

World role	`HARMFUL` prediction		RD (pp)	95% CI	McNemar p	NH→H
	Full context	Selected only				
Original	367/932 (39.38%)	406/932 (43.56%)	+4.18	[2.04, 6.44]	.000237	68/554 (12.27%)
Control	378/932 (40.56%)	400/932 (42.92%)	+2.36	[0.11, 4.61]	.0487	64/543 (11.79%)

Table 47: **DeepSeek-V4-Pro 45-video text negative-safety sentinel.** `HARMFUL` rates are in %.

Role	Full context	Selected only	RD (pp)	95% CI	p	NH→H
Original	26.67	44.44	+17.78	[4.44, 31.11]	.0386	10/31 (32.26%)
Control	26.67	31.11	+4.44	[−8.89, 17.78]	.7539	5/29 (17.24%)

H Ex-HateMM Semantic-Support Analysis

We retain the Ex-HateMM analysis as a secondary characterization of semantic support remaining in the visual removal observations. Among the 380 HateMM videos in the frozen visual cohort, 372 join exactly to Ex-HateMM; restricting to the official hate designation leaves 254 videos.

A blinded auxiliary annotation assigns each observed payload to Relation Complete, Surface Cue Only, or No Relevant Support. The last category means that no such support is identified in the observed payload; it does not establish complete removal from the underlying video. Figure 9 summarizes the resulting persistence pattern, with exact rates reported in Table 48.

Persistence is highest when the hate-defining relation remains recoverable, lower when only surface-level support remains, and lowest when no relevant support is identified. This analysis is descriptive and secondary to the observation-bounded support analyses above.

The pre-specified rationale-alignment analysis is not performed because comparable frozen rationales are not consistently available. The hard-negative analysis is also not estimated because the joined subset contains only three official non-hate cases, below the pre-specified minimum of $N \geq 30$.

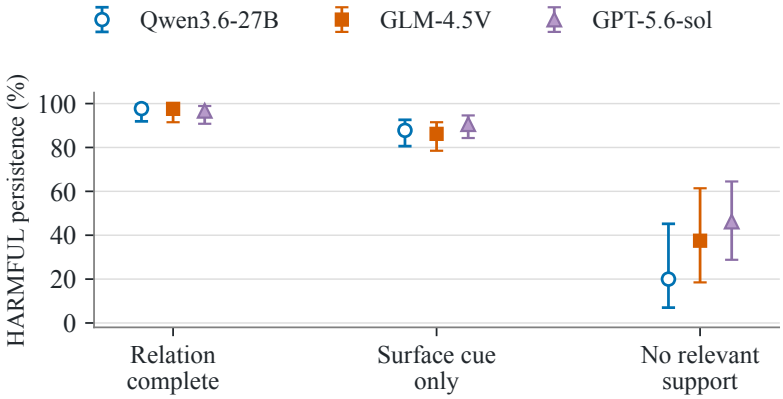


Figure 9: **Prediction persistence varies with the semantic support identified after source removal.**

Table 48: **HARMFUL persistence by semantic support remaining after source removal.**

Semantic state	n	HARMFUL	Persistence (%)	95% CI
Qwen3.6-27B				
Relation complete	86	84	97.7	[91.9, 99.4]
Surface cue only	115	101	87.8	[80.6, 92.6]
No relevant support	15	3	20.0	[7.0, 45.2]
GLM-4.5V				
Relation complete	82	80	97.6	[91.5, 99.3]
Surface cue only	109	94	86.2	[78.5, 91.5]
No relevant support	16	6	37.5	[18.5, 61.4]
GPT-5.6-sol				
Relation complete	91	88	96.7	[90.8, 98.9]
Surface cue only	128	116	90.6	[84.3, 94.6]
No relevant support	26	12	46.2	[28.8, 64.5]

I LIMITATIONS

Human-defined, observation-bounded support. Audited support is identified by human annotators within the model's observation; it is neither exhaustive support in the raw video nor the features the model uses, and a large response to removing content humans judge harmfulness-relevant is partly expected. Same-budget random controls, whose targets include about 65% of audited support, and the within-case coverage analysis show that deletion amount does not explain this response, not that the model relies on the cues humans mark; coverage–response relations are associations rather than causal effects.

Text-centric core evidence. The graded coverage–response and matched-random analyses rest on ASR/OCR observations extracted from the videos, a 120-case audit, and fully covered strata of only 13 and 15 cases. These channels carry harmful cues such as slurs and on-screen captions (Figure 1), but the visual evidence is weaker: the visual audit has lower agreement ($\kappa = 0.42$ versus 0.79), its matched-random comparison retains 11 and 10 non-empty cases, and its residual contrasts compare different cases. We treat it as corroborating; a joint frames-plus-text input reproduces the preservation–removal pattern (Appendix D.6) but has no support audit.

Observation density and scope. Inputs are 8 or 16 sampled frames or ASR/OCR units, so coverage estimates describe these observations rather than the underlying videos. All models are accessed through an API, and the two-model, 120-pair Charades-STA transfer shows that the mismatch can arise in general temporal grounding without establishing its prevalence.