

OmniHalluc-L: Counterfactual Benchmarking and Modality-Perturbation Calibration for Long-Form Omni Hallucination

Zixuan Dong^{1,*} Jiafu Tang^{2,*} Zhide Lei² Zhe Cao² Zijie Zhang²
Yanghai Wang² Shihao Li² Xiaodong Wang¹ Baoyun Peng³ Jiaheng Liu^{2,†}

¹ National University of Defense Technology, Changsha, China

² Nanjing University, Nanjing, China ³ Academy of Military Sciences, Beijing, China

Abstract

Long-video Omni assistants often fail not by inventing content, but by misbinding real evidence: they hear the right utterance and see the right event, yet attach it to the wrong speaker, moment, or modality. These *almost-true* errors evade standard video QA because local evidence remains valid, so item-level scoring can reward both a supported claim and its near-counterfactual. We introduce a counterfactual event-binding protocol that constructs paired supported/counterfactual claims from the same audio-visual event evidence and evaluates them by strict-pair accuracy. We instantiate it as OmniHalluc-L, to our knowledge the first benchmark for long-video Omni hallucination, with 3,600 single-claim QA items from 638 long-form videos averaging 24.16 minutes and covering 256.87 hours. Under this protocol, open-weight Omni models remain weak at pair-level binding: Qwen2.5-Omni-7B reaches 32.06% and Qwen3-Omni-Instruct reaches 41.55%, versus 76.54% for a closed-source reference. To narrow this gap, we propose MPRC (Modality-Perturbation Reliability Calibration), a frozen-backbone calibrator that combines native confidence with response shifts from selected audio-negative probes. MPRC raises these two models to 36.22% and 51.09%, respectively, and improves target-adapted MCQ accuracy on OmniVideoBench and WorldSense. The code is available at <https://github.com/ZexDong/omnihallucL-mprc>.

1 Introduction

Omni assistants are moving from curated short clips to minutes-long audio-visual streams such as lectures, podcasts, and livestreams. At this horizon, hallucination often means misbinding real evidence rather than inventing content: a model may

* Equal contribution.

† Corresponding author: liujiaheng@nju.edu.cn.

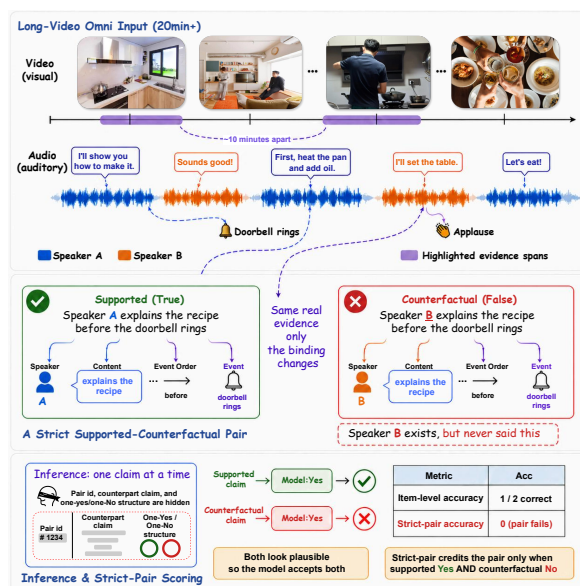


Figure 1: Motivation for strict-pair scoring. Supported and counterfactual claims share the same real evidence but differ in binding; OmniHalluc-L evaluates each claim independently and credits a pair only when the supported claim is accepted and the counterfactual is rejected.

hear the right utterance and see the right event, yet assign them to the wrong speaker, moment, scene, or modality. The evidence is present, but the claimed relation is false. Such *almost-true* errors evade standard video QA because item-level scoring can reward both a supported claim and a minimally edited near-counterfactual.

Current evaluations ask adjacent questions. Long-video benchmarks measure comprehension and reasoning (Fu et al., 2025; Wu et al., 2024; Zhou et al., 2025), hallucination benchmarks have moved from images to vision-only video settings (Li et al., 2023; Wang et al., 2024; Lu et al., 2025), and Omni evaluations target broad audio-visual understanding and real-world omnimodal reasoning (Li et al., 2026; Hong et al., 2026; Tao et al., 2026; Goel et al., 2026). What remains missing is

whether recurring evidence is attached to the right source, time, scene, and modality in a long stream.

We address this gap with a counterfactual event-binding protocol. Starting from real audio-visual event evidence, we author a supported claim and a minimally edited counterfactual that preserves the evidence but changes its binding relation. Pair identity is hidden: models verify one claim at a time, and strict-pair accuracy credits only accepting the supported claim and rejecting the counterfactual. This turns hallucination evaluation from local plausibility checking into relation-consistency testing.

We instantiate this protocol as OmniHalluc-L, to our knowledge the first benchmark for long-video Omni hallucination. It contains 3,600 claims from 638 long-form videos, organized into 1,800 hidden supported/counterfactual pairs across three evidence-binding families. We further propose MPRC (Modality-Perturbation Reliability Calibration), a frozen-backbone calibrator that combines native confidence with response shifts from selected audio-negative probes.

Experiments reveal a large reliability gap and an actionable path to reduce it. Qwen2.5-Omni-7B reaches 32.06% SPA and Qwen3-Omni-Instruct reaches 41.55%, while the strongest closed-source reference reaches 76.54%. Without updating the Omni backbone, MPRC lifts them to 36.22% and 51.09%, outperforming the MAD contrastive decoder (Chung et al., 2026). The same audio-counterfactual signal improves target-adapted MCQ accuracy on OmniVideoBench (+2.20) and WorldSense (+1.51) with Qwen3-Omni-Instruct.

Our contributions are: (i) a counterfactual audio-visual event-binding construction and strict-pair evaluation protocol; (ii) OmniHalluc-L, a 3,600-item long-video Omni hallucination benchmark built under this protocol; (iii) MPRC, a frozen-backbone reliability calibration framework based on structured audio-negative probes; and (iv) a systematic study of open-weight, frozen-decoding, and closed-source Omni systems under strict evidence-binding evaluation.

2 Related Work

Long-video understanding and Omni evaluation. Long-video benchmarks test retrieval, summarization, event localization, and reasoning over extended temporal context (Mangalam et al., 2023;

Wu et al., 2024; Zhou et al., 2025; Fu et al., 2025; Geng et al., 2025). Omni, speech-centric, and inconsistency benchmarks evaluate audio-visual understanding, real-world omnimodal reasoning, speaker-level audiovisual grounding, and injected conflicts (Nguyen et al., 2025; Tao et al., 2026; Goel et al., 2026; Chen et al., 2026b). They cover comprehension, perception, and inconsistency handling broadly, but not the stricter question of whether recurring real evidence is bound to the correct speaker, moment, scene, or modality under hidden supported/counterfactual pairs.

Multimodal hallucination benchmarks. Hallucination benchmarks began with image-level object and language-vision conflicts (Li et al., 2023; Guan et al., 2024; Wang et al., 2023; Sun et al., 2024), then expanded to video event, temporal, semantic-aggregation, and compositional failures (Wang et al., 2024; Zhang et al., 2024; Li et al., 2025; Lu et al., 2025; Xing et al., 2026). AVHBench and AV-ConfuseBench test audio-visual confusion under altered or missing sounds (Kim et al., 2025; Ye et al., 2026), while AHA mines shared-evidence counterfactuals for DPO-based short-clip audio reasoning (Chen et al., 2026a). Related work covers omni-modal captioning and generation alignment (Wang et al., 2026; Cao et al., 2026). Unlike these efforts, OmniHalluc-L tests long-form claims that share real evidence but differ in binding.

Frozen-model hallucination mitigation.

Training-free mitigation methods use decoding-time calibration or contrastive signals without updating the backbone (Shi et al., 2024; Manevich and Tsarfaty, 2024; Leng et al., 2024; Huang et al., 2024). Audio-visual decoders such as MAD and AVCD compare native responses with perturbed modality views (Chung et al., 2026; Jung et al., 2025). These methods are useful frozen-decoding references; however, long-form binding errors are heterogeneous, so MPRC estimates reliability over structured probes rather than relying on one universal perturbation rule.

3 The OmniHalluc-L Benchmark

3.1 Counterfactual Claim Verification

OmniHalluc-L evaluates whether an Omni model binds real audio-visual evidence to the correct relation. Given a long video and a natural-language claim, the model answers YES or NO. Each

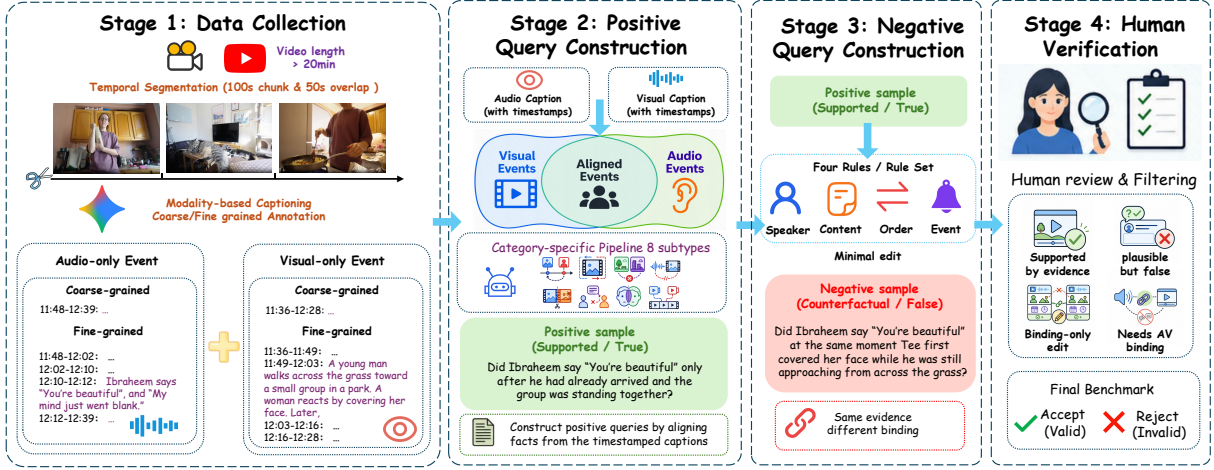


Figure 2: OmniHalluc-L construction pipeline. Long-form videos are temporally segmented, annotated through isolated audio and visual views, authored into supported/counterfactual claim pairs, and validated by human verification so that each pair differs only in the evidence-binding relation.

Dataset	Long	Omni	Hall.	SPA
VideoHallucur	×	×	✓	×
EventHallusion	×	×	✓	×
VidHalluc	×	×	✓	×
AVHBench	×	✓	✓	×
AV-ConfuseBench	×	✓	✓	×
ELV-Halluc	✓	×	✓	×
OmniVCHall	×	×	✓	×
AVID	×	✓	✓	×
OmniHalluc-L	✓	✓	✓	✓

Table 1: Benchmark coverage comparison. Long denotes average duration above 5 minutes; Omni denotes joint audio-visual input; Hall. denotes hallucination, AV-confusion, or cross-modal inconsistency evaluation; SPA denotes our strict hidden supported/counterfactual pair scoring.

instance belongs to a hidden strict pair $p = (v_p, q_p^+, q_p^-)$: q_p^+ is supported by the video, while q_p^- preserves the same evidence but rebinds it to the wrong speaker, scene, moment, or modality. At test time, the model receives only one claim; the pair identifier, paired counterfactual, and one-positive-per-pair structure are hidden. The task therefore tests whether a model can reject a false relation over real evidence, not merely whether it can detect that relevant evidence exists.

A pair is credited only when both sides are answered correctly. The primary metric is strict-pair accuracy:

$$\text{SPA} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbb{I}[g(v_p, q_p^+) = 1] \cdot \mathbb{I}[g(v_p, q_p^-) = 0]. \quad (1)$$

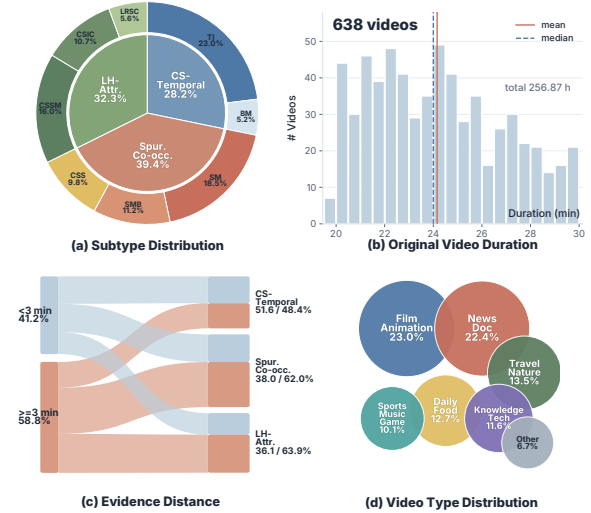


Figure 3: OmniHalluc-L dataset statistics over 3,600 verified QA items from 638 long-form videos. Panels show subtype and family distributions, source-video duration, evidence distance, and video type coverage.

SPA is 0% for constant YES/NO predictors and 25% in expectation for random guessing, making answer-prior shortcuts visible rather than hiding them inside item-level accuracy. Table 1 situates this protocol against vision-only hallucination, audio-visual hallucination or confusion, and injected-inconsistency benchmarks.

3.2 Benchmark Construction

OmniHalluc-L is built from 638 long-form source videos averaging 24.16 minutes and totaling 256.87 hours, a regime where speakers, sounds, scenes, and events recur across many minutes. We segment each video into overlapping 100-second

Model	In.	CS-Temporal		Spur. Co-occ.			LH-Attr.			All
		TI	BM	SM	SMB	CSS	CSSM	CSIC	LRSC	
Open-weight Omni models										
Qwen2.5-Omni-3B	AV	14.98	18.09	29.43	34.83	31.82	25.00	23.32	38.61	25.50
Qwen2.5-Omni-7B	AV	27.05	14.89	39.04	38.31	41.48	22.57	34.72	38.61	32.06
OmniVinci	AV	28.99	21.28	32.13	37.81	42.05	25.35	31.09	30.69	31.17
MiniCPM-o 4.5	AV	35.75	26.60	36.04	41.29	47.73	24.31	37.82	48.51	36.22
Qwen3-Omni-Instruct	AV	37.92	37.23	43.07	38.19	44.83	40.28	48.44	47.00	41.55
Open-weight video-only models										
Qwen2.5-VL-3B	V	14.49	18.09	28.23	32.34	35.80	21.18	26.94	28.71	24.50
Qwen2.5-VL-7B	V	19.57	19.15	28.53	31.34	30.68	19.10	38.34	33.66	26.33
Qwen3-VL-32B	V	28.02	20.21	29.82	33.17	41.95	18.06	46.88	37.00	30.79
Closed-source Omni models										
Gemini 3.1 Flash-Lite	AV	50.00	38.30	63.83	62.44	55.43	61.97	69.63	50.00	57.86
Gemini 3.0 Flash	AV	65.45	55.32	74.09	61.31	69.14	65.61	81.25	60.00	67.83
Gemini 3.1 Pro	AV	81.31	68.09	78.55	74.75	73.71	73.59	82.29	64.00	76.54

Table 2: Strict-pair accuracy (%) on the full OmniHalluc-L benchmark. All models follow the same single-query protocol; pair identity is hidden at test time.

windows with a 50-second stride, limiting caption drift while preserving boundary-crossing speech and audio-visual cues.

Construction follows the same principle as the evaluation protocol: keep the evidence real and vary only the binding relation. We first use modality-isolated annotation, describing audio without visual context and visual content without audio context, so the pipeline does not pre-solve cross-modal grounding. Category-specific authoring protocols then produce a supported claim and a near-counterfactual that changes only the relation being tested. Human verification checks entailment, false-but-plausible counterfactuality, binding-only edits, and shortcut leakage; full instructions and filtering statistics are provided in Appendix A.

The verified benchmark contains 3,600 single-claim QA items across three relation-defined families. **Cross-Stream Temporal** tests timing and boundary alignment between audio and visual streams. **Spurious Co-occurrence** tests false association among real scene, sound, and action evidence. **Long-Horizon Attribution** tests speaker attribution, identity continuity, and source tracking across distant moments. Together, they cover the timing, co-occurrence, and attribution relations that long-video Omni models must preserve. This relation-centric taxonomy makes near-counterfactual evaluation possible: the local evidence remains real, while the tested claim changes only how that evidence is connected.

Evidence intervals are retained as diagnostic

metadata but are not provided to models. The evaluated task remains single-query verification over the model’s normalized video input; subtype definitions, evidence metadata, and verification details are reported in Appendix A.

3.3 Evaluation Protocol

Open-weight Omni models are scored from canonical YES/NO logits using a binary margin and softmax-normalized support score, while closed-source systems are parsed from generated text with bounded retries; token choices, API identifiers, decoding settings, and failure-handling rules are given in Appendix A.

The primary metric is SPA (Eq. 1); item accuracy, balanced item accuracy, and predicted-YES rate are reported only as diagnostics because item-level metrics can mask pair collapse. All held-out evaluations use the same single-query setting. For models with duration constraints, we apply fixed query-agnostic temporal preprocessing and 256-frame sampling, with model-specific sidecar or union-clip formatting where required; no evidence span, retrieval window, or pair-level information is selected per query.

3.4 Benchmark Results

Table 2 reports SPA on the full OmniHalluc-L benchmark. Current open-weight Omni systems remain far below the strongest closed-source reference: Qwen2.5-Omni-7B reaches 32.06% SPA and Qwen3-Omni-Instruct reaches 41.55%, while Gemini 3.1 Pro reaches 76.54%. The gap leaves

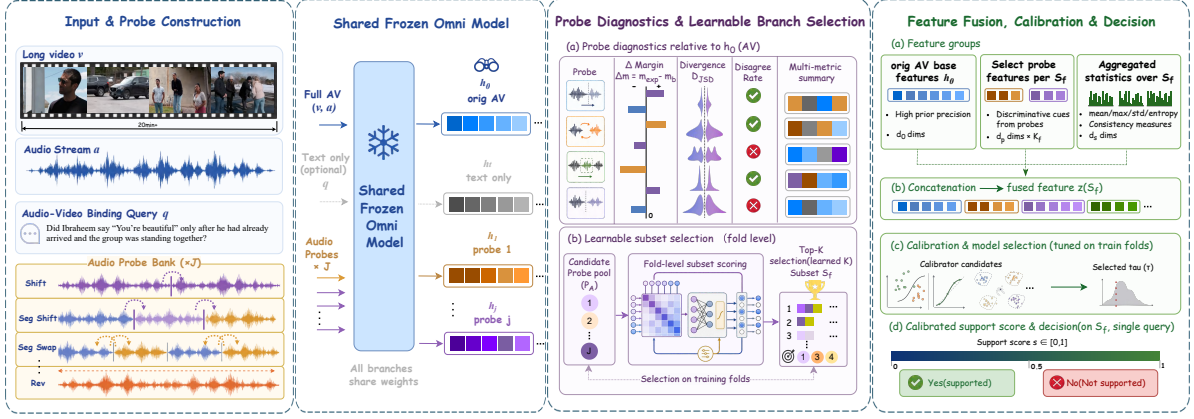


Figure 4: MPRC overview. A frozen Omni backbone is queried under native audio-visual input and structured audio-negative probes. Branch subsets and a lightweight calibrator are selected within video-level folds; held-out claims are predicted independently from native confidence and branch-induced response shifts.

substantial headroom, while nontrivial video-only baselines and stronger matched Omni results show that OmniHalluc-L is not reducible to language plausibility, visual recognition, or transcript-level text alone (Appendix C). Errors span all three binding families, making OmniHalluc-L a targeted diagnostic for long-form evidence binding rather than a generic video QA leaderboard.

4 MPRC: Modality-Perturbation Reliability Calibration

Long-video hallucination is difficult to mitigate because the relevant evidence may already be present in the model’s response pathway. The failure is often relational: a real utterance is assigned to the wrong speaker, a real sound is attached to the wrong scene, or a true event is carried beyond the moment that supports it. A single audio perturbation cannot diagnose all such failures. Temporal shifts stress cross-stream alignment, segment swaps stress local co-occurrence, and long-range dislocations stress attribution continuity. The same perturbation, however, can be informative for one claim and irrelevant for another. MPRC therefore treats audio perturbations not as a universal contrastive recipe, but as reliability probes for estimating whether the model’s native binding judgment should be trusted. The probe is not meant to replace the original evidence; it asks whether the model’s support decision remains coherent when the audio relation most relevant to binding is disturbed.

MPRC is a frozen-backbone calibration framework that keeps the visual stream fixed, queries native audio-visual and structured audio-negative

views, and converts response shifts into support features. A lightweight reliability layer, selected within the training folds, combines native confidence with selected audio-negative signals without updating the backbone. This reframes mitigation as post-hoc evidence-binding calibration: the model supplies perceptual and sensitivity signals, while MPRC learns when those signals justify per-claim support.

Reliability features. Let F be the frozen Omni backbone and let (v, q) denote a video-claim instance. The native audio-visual query produces $h_0 = F(v, q)$. From h_0 , we extract $\phi_{av}(h_0)$, including the parsed answer, support confidence, Yes/No margin, and response uncertainty when available. For an audio-negative branch b , let $T_b(v)$ denote the same visual stream paired with a perturbed audio stream, and let $h_b = F(T_b(v), q)$. The branch feature $\phi_b(h_b, h_0)$ contains the branch answer and confidence as well as deltas in support score, margin, and uncertainty relative to the native AV view. An optional text-prior block $\phi_{text}(F(q), h_0)$ measures whether the claim is answerable from language priors alone; this block is included only when selected by validation.

For a candidate branch subset $\mathcal{B}$, the feature representation is

$$\mathbf{x}_{\mathcal{B}}(v, q) = \left[\phi_{av}(h_0), \phi_{text}(F(q), h_0), \{\phi_b(h_b, h_0)\}_{b \in \mathcal{B}} \right], \quad (2)$$

with the text-prior block omitted unless selected by validation. The representation remains claim-sensitive through (v, q) -dependent response shifts, while the branch subset is fixed by training-video

selection and never chosen from test-time pair information.

Audio-negative probe bank. The probe bank is designed to disturb audio-visual binding while preserving the visual evidence. Short global shifts and segment shifts probe **Cross-Stream Temporal** errors by changing the relative timing between speech, sounds, and visual events. Adjacent segment swaps probe **Spurious Co-occurrence** errors by keeping plausible local audio content while attaching it to a neighboring visual context. Long-range shifts and reversal probes target **Long-Horizon Attribution** errors by breaking source continuity and sequential integration. No branch is assumed to be universally helpful. The candidate bank includes these shift, swap, long-range, and reversal probes; the exact subset used by a model and fold is selected on training videos. Pilot comparisons with visual and audio+video pools did not consistently improve over the audio-only bank (Table 10).

Calibration and model selection. For each video-level outer fold f , MPRC searches over branch subsets, thresholds, and lightweight heads including logistic-regression variants, extra-trees, and gradient-boosted trees, using only training videos. Each head maps $\mathbf{x}_B(v_i, q_i)$ to an item-level support estimate $y_i \in \{0, 1\}$; pair identifiers are never input features. Model selection then uses out-of-fold training predictions to optimize strict-pair reliability:

$$(\mathcal{B}_f^*, h_f^*, \tau_f^*) = \arg \max_{\mathcal{B}, h, \tau}^{\text{lex}} \mathbf{m}_{\mathcal{B}, h, \tau}, \quad (3)$$

$$\mathbf{m}_{\mathcal{B}, h, \tau} = \left[\text{SPA}^{\text{OOF}}, \text{Acc}_{\text{item}}^{\text{OOF}}, -|\hat{p}_Y^{\text{OOF}} - 0.5|, -|\mathcal{B}| \right].$$

Here SPA^{OOF} uses training-video pair identifiers, $\text{Acc}_{\text{item}}^{\text{OOF}}$ is item accuracy, and $\hat{p}_Y^{\text{OOF}}$ is the predicted-YES rate. The lexicographic rule prioritizes SPA, item accuracy, balanced predicted-YES rate, and compact branch subsets; ties prefer the full feature mode. Because selection is video-level, held-out videos cannot influence branch choice, calibration, or threshold tuning.

Pair-blind test-time inference. After selection, a held-out claim (v, q) in fold f is scored by

$$s_f(v, q) = h_f^* \left(\mathbf{x}_{\mathcal{B}_f^*}(v, q) \right), \quad (4)$$

$$g_f(v, q) = \mathbb{I}[s_f(v, q) \geq \tau_f^*].$$

At test time, MPRC predicts each claim independently. It receives no pair identifier, paired counterfactual, evidence span, or query-specific retrieval window, and imposes no one-positive-per-pair constraint. Pair identifiers are used only for training-validation model selection and final evaluation, so deployment remains single-query and pair-blind.

Frozen-backbone reliability layer. MPRC is deliberately post-hoc. No gradients flow into the audio encoder, visual encoder, or language model, and no video evidence is relabeled at test time. Following standard post-hoc calibration practice, the method calibrates whether the backbone’s own evidence-binding judgment is reliable under native confidence and structured audio-negative sensitivity. The calibration target is relational rather than merely probabilistic: branch-induced shifts test whether support remains coherent when binding-sensitive audio structure is disturbed. This separation is useful for long-form Omni models: as the backbone improves, its confidence margins and perturbation responses can become cleaner reliability signals, while MPRC converts those signals into stricter support decisions without changing the model itself.

5 Experiments

5.1 Experimental Protocol

All experiments use video-level five-fold splits: every query from the same source video is assigned to the same fold, and trainable components are fit only on training videos before evaluation on held-out videos. We report held-out fold means, with confidence intervals computed by video-level bootstrap over out-of-fold predictions.

The central comparison is backbone-training-free reliability improvement. Raw AV uses the frozen model, MAD is the closest frozen-decoding reference with contrastive strength selected within training, and MPRC fits a lightweight reliability layer on labeled training videos without updating the Omni backbone. Because existing mitigation methods are not designed for paired long-form Omni verification, the comparison asks whether a fixed Omni model can be made more reliable through post-hoc evidence-binding calibration. Additional details on MAD tuning and comparison scope are given in Appendix D.

The primary metric is strict-pair accuracy (SPA). Item accuracy, balanced item accuracy,

Model	Method	C/S/L	All	Δ	Model	Method	C/S/L	All	Δ
Qwen2.5-O-7B	Raw AV	24.80/39.44/29.38	32.06 _[29.76,34.47]	–	Qwen3-O-Inst.	Raw AV	37.80/42.13/44.14	41.55 _[39.19,44.03]	–
	MAD	25.98/41.27/32.30	34.06 _[31.80,36.42]	+2.00		MAD	40.75/46.38/45.86	44.62 _[42.17,47.07]	+3.07
	MPRC	27.36/42.68/36.08	36.22 _[33.82,38.57]	+4.16		MPRC	43.31/52.34/56.38	51.09 _[48.66,53.48]	+9.54
OmniVinci	Raw AV	27.56/36.20/28.18	31.17 _[28.94,33.41]	–	MiniCPM-o 4.5	Raw AV	34.06/40.42/32.99	36.22 _[33.84,38.52]	–
	MAD	25.79/38.31/29.55	31.94 _[29.81,34.11]	+0.77		MAD	37.20/41.69/35.22	38.33 _[35.90,40.84]	+2.11
	MPRC	27.95/38.56/30.29	32.89 _[30.64,35.17]	+1.72		MPRC	38.19/45.49/39.35	41.44 _[39.10,43.83]	+5.22

Table 3: OmniHalluc-L main results on the full benchmark (%). C/S/L denotes Cross-Stream Temporal, Spurious Co-occurrence, and Long-Horizon Attribution. All entries report SPA with video-level bootstrap 95% confidence intervals as small subscripts, and Δ is over Raw AV.

and predicted-YES rate are tracked as diagnostics because item-level metrics can hide answer priors and pair collapse. We evaluate open-weight Omni models including Qwen2.5-Omni-7B (Xu et al., 2025a), Qwen3-Omni-Instruct (Xu et al., 2025b), OmniVinci (Ye et al., 2025), and MiniCPM-o 4.5 (Cui et al., 2026). We also include Qwen2.5-VL and Qwen3-VL as video-only references (Bai et al., 2025b,a). Closed-source Omni models are reported only as Raw AV references because their logits and confidence features are not exposed.

5.2 Main Results

Table 3 shows that fixed contrastive decoding is useful but limited. MAD improves every open-weight backbone (+0.77 to +3.07 SPA), showing that modality perturbation exposes reliability signal even without backbone updates. Appendix D further shows that some training-free fixed branches are already competitive with MAD, but the best branch is not stable across data distributions and output formats. MPRC therefore learns a post-hoc reliability layer on training videos. This raises Qwen2.5-Omni from 32.06% to 36.22%, Qwen3-Omni-Instruct from 41.55% to 51.09%, OmniVinci from 31.17% to 32.89%, and MiniCPM-o 4.5 from 36.22% to 41.44%.

For Qwen2.5-Omni, the largest gain appears in Long-Horizon Attribution, rising from 29.38% to 36.08% (+6.70). For Qwen3, gains are largest in Long-Horizon Attribution, from 44.14% to 56.38% (+12.24), and similarly large in Spurious Co-occurrence, from 42.13% to 52.34% (+10.21). These are the cases where fixed item-level confidence tends to blur who did what, when, and in which context. Thus, MPRC is not merely correcting low-confidence items; it is strongest when real evidence must be assigned to the correct event, source, or distant context.

Model	Variant	All	Drop	K
Qwen2.5-Omni-7B	Full	36.22	0.00	2.60
	w/o Geo.	34.67	1.55	1.40
	w/o Unc.	35.05	1.17	2.80
	w/o Aud.	35.10	1.12	–
Qwen3-Omni-Instruct	Full	51.09	0.00	4.00
	w/o Geo.	50.98	0.11	4.00
	w/o Unc.	50.77	0.32	3.80
	w/o Aud.	49.70	1.39	–

Table 4: Component ablation under the refined component-isolation protocol. Drop is relative to Full within this ablation setting; Geo., Unc., and Aud. denote score geometry, uncertainty, and audio-counterfactual features, and K is the mean selected branch count.

The improvement also scales with backbone quality. Qwen3 starts from a stronger Raw AV baseline than Qwen2.5, yet receives the larger MPRC gain. This suggests that stronger backbones can expose cleaner confidence margins and more separable audio-counterfactual responses. Rather than making calibration unnecessary, stronger perception supplies better reliability signals for stricter evidence-binding decisions.

5.3 Ablation Study

Table 4 isolates components under a refined ablation protocol; the Full rows are diagnostic, not replacements for Table 3. The w/o Aud. variant is a strong native-feature control: it uses the same reliability layer but removes audio-counterfactual branch features. Adding those features improves SPA by 1.12 points for Qwen2.5 and 1.39 points for Qwen3, isolating the incremental value of binding-sensitive response shifts beyond native confidence calibration. Score geometry, uncertainty, and compact selected branch counts support the design: MPRC benefits from structured probes without using all probes uniformly.

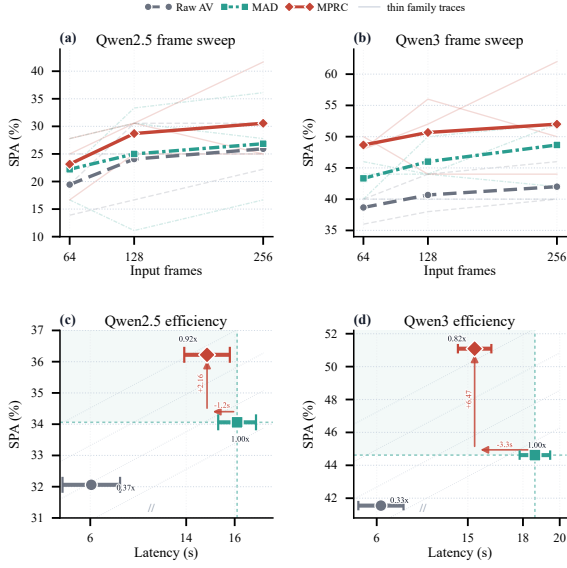


Figure 5: Frame-count and efficiency diagnostics on per-model frame-count subsets. Panels (a) and (b) compare Raw AV, MAD, and MPRC under 64, 128, and 256 input frames; panels (c) and (d) summarize latency versus SPA. Thin traces show family-level trends.

5.4 Frame Count Analysis

Figure 5 tests whether denser visual sampling or higher decoding cost can replace reliability calibration on per-model, subtype-balanced subsets. Increasing frames improves Raw AV, but MPRC remains higher at every density. Panels (c) and (d) add the efficiency view: selected-branch MPRC is faster than MAD while achieving higher SPA, because it evaluates only the training-selected probes. Thus, broader visual coverage and heavier fixed decoding help, but neither resolves which recurring evidence should support the claim; evidence-binding reliability still requires calibration.

5.5 Cross-Benchmark Generalization

Table 5 evaluates whether the reliability principle extends beyond paired Yes/No verification on OmniVideoBench and WorldSense. These are not zero-shot transfer experiments. For MCQ benchmarks, we do not verbalize each option as a binary claim; instead, MPRC calibrates choice-logit and audio-counterfactual option features under video-level target folds. The Omni backbone remains frozen throughout. This format shift is useful: it tests whether audio-counterfactual reliability remains informative when the decision is expressed through competing answer options rather than hidden Yes/No pairs.

The pure MPRC variant improves all four tar-

Model	Raw	MPRC	Cal.	Cal.+Opt.
<i>OmniVideoBench</i>				
Qwen2.5-Omni-7B	34.30	35.50 _{+1.20}	34.80	34.80 _{+0.50}
Qwen3-Omni-Instruct	39.50	41.70 _{+2.20}	40.80	43.10 _{+3.60}
<i>WorldSense</i>				
Qwen2.5-Omni-7B	43.00	45.24 _{+2.24}	44.07	48.36 _{+5.36}
Qwen3-Omni-Instruct	49.24	50.76 _{+1.51}	50.32	54.16 _{+4.92}

Table 5: Cross-benchmark MCQ adaptation accuracy (%). Subscripts show gains over Raw. MPRC uses audio-counterfactual MCQ decoding; Cal. adds video-level target calibration; Cal.+Opt. further adds option-position normalization.

get settings, with gains of +1.20 and +2.20 on OmniVideoBench and +2.24 and +1.51 on WorldSense. Target calibration and option-position normalization further improve Qwen3 on OmniVideoBench (+3.60) and both models on WorldSense (+5.36 and +4.92).

Although these gains are smaller than in-domain OmniHalluc-L improvements, their consistent direction across benchmarks, backbones, and output formats suggests that MPRC captures a reusable reliability signal rather than a shortcut tied to OmniHalluc-L’s pair format. This is notable because the MCQ setting removes the hidden-pair structure entirely: the model must express reliability through competing option scores rather than a supported/counterfactual decision. The positive direction therefore indicates that audio-counterfactual sensitivity is not merely a scoring artifact of OmniHalluc-L, but a portable signal for long-form Omni reliability.

6 Conclusion

We presented OmniHalluc-L, a long-form Omni hallucination benchmark testing correct audio-visual evidence binding across speaker, moment, scene, and modality. With 3,600 verified QA items from 1,800 pairs and 638 roughly 24-minute videos, OmniHalluc-L recasts hallucination evaluation as strict supported/counterfactual testing and reveals a reliability gap in open-weight Omni models. We introduced MPRC, a frozen-backbone reliability layer calibrating AV confidence with audio-negative response shifts. Gains on OmniHalluc-L and target-adapted MCQ benchmarks suggest a route to more reliable long-form Omni assistants. We release the MPRC implementation and will release the benchmark annotations and evaluation resources at the same public repository.

Limitations

OmniHalluc-L evaluates long-form evidence binding through single-claim verification. This makes supported/counterfactual consistency directly measurable but does not cover multi-turn dialogue, open-ended generation, or interactive evidence seeking. Extending strict binding evaluation to these settings remains future work. MPRC is a post-hoc frozen-backbone calibration framework, most applicable to models exposing confidence signals or option scores; closed-source systems are therefore reported as Raw AV references when these features are unavailable.

Ethical Considerations

The video clips comprising OmniHalluc-L are gathered from publicly accessible online platforms. In order to respect intellectual property and align with standard copyright regulations, the benchmark will be released under a restrictive license that limits its application exclusively to academic research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62506161). This work was supported in part by the Jiangsu Science and Technology Project (BK20251199).

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-vl technical report](#). *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *arXiv preprint arXiv:2502.13923*.
- Zhe Cao, Tao Wang, Jiaming Wang, Yanghai Wang, Yuanxing Zhang, Jialu Chen, Miao Deng, Jiahao Wang, Yubin Guo, Chenxi Liao, Yize Zhang, Zhaoxiang Zhang, and Jiaheng Liu. 2026. [T2AV-Compass: Towards unified evaluation for text-to-audio-video generation](#). In *International Conference on Machine Learning*.
- Yanxi Chen, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Xin Li, Peijie Qiu, Hao Wang, Xuanzhao Dong, Yujian Xiong, Anderson Schneider, Yuriy Nevmyvaka, and Yalin Wang. 2026a. [AHA: Aligning large audio-language models for reasoning hallucinations via counterfactual hard negatives](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 29294–29306, San Diego, California, United States. Association for Computational Linguistics.
- Zixuan Chen, Depeng Wang, Hao Lin, Li Luo, Ke Xu, Ya Guo, Huijia Zhu, Tanfeng Sun, and Xinghao Jiang. 2026b. [Avid: A benchmark for omni-modal audio-visual inconsistency understanding via agent-driven construction](#). *arXiv preprint arXiv:2604.13593*.
- Sangyun Chung, Se Yeon Kim, Youngchae Chee, and Yong Man Ro. 2026. Mad: Modality-adaptive decoding for mitigating cross-modal hallucinations in multimodal large language models. *arXiv preprint arXiv:2601.21181*.
- Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, and 17 others. 2026. [Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction](#). *arXiv preprint arXiv:2604.27393*.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiaowu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multimodal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. 2025. Long-VALE: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Arushi Goel, Sreyan Ghosh, Vatsal Agarwal, Nishit Anand, Kaousheik Jayakumar, Lasha Koroshinadze, Yao Xu, Katie Lyons, James Case, Karan Sapra, Kevin J. Shih, Siddharth Gururani, Abhinav Shrivastava, Ramani Duraiswami, Dinesh Manocha, Andrew Tao, Bryan Catanzaro, Mohammad Shoeybi, and Wei Ping. 2026. [Mmou: A massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos](#). *arXiv preprint arXiv:2603.14145*.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced

- diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2026. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *The Fourteenth International Conference on Learning Representations*.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. 2025. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In *Advances in Neural Information Processing Systems*.
- Sung-Bin Kim, Hyun-Bin Oh, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. 2025. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *The Thirteenth International Conference on Learning Representations*.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, Ying He, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, and 23 others. 2026. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *The Fourteenth International Conference on Learning Representations*.
- Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2025. Vid-halluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Hao Lu, Jiahao Wang, Yaolun Zhang, Ruohui Wang, Xuanyu Zheng, Yepeng Tang, Dahua Lin, and Lewei Lu. 2025. Elv-halluc: Benchmarking semantic aggregation hallucinations in long video understanding. *arXiv preprint arXiv:2508.21496*.
- Avshalom Manevich and Reut Tsarfaty. 2024. [Mitigating hallucinations in large vision-language models via language-contrastive decoding](#). *arXiv preprint arXiv:2408.04664*.
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.
- Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, and Yong Jae Lee. 2025. [See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models](#). *arXiv preprint arXiv:2512.02231*.
- Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2024. [Trusting your evidence: Hallucinate less with context-aware decoding](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 783–791.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL*.
- Keda Tao, Yuhua Zheng, Jia Xu, Wenjie Du, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, Junhan Zhu, Bohan Yu, Weiqiang Wang, Jian Liu, Can Qin, Yulun Zhang, Ming-Hsuan Yang, and Huan Wang. 2026. [Lvomnibench: Pioneering long audio-video understanding evaluation for omnimodal llms](#). *arXiv preprint arXiv:2603.19217*.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2023. [Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation](#). *arXiv preprint arXiv:2311.07397*.
- Yanghai Wang, Jiahao Wang, Jiafu Tang, Yuanxing Zhang, Zhe Cao, Hanyan Bian, Zijie Zhang, Weiliang Luo, Zhiyu Pan, Zixuan Dong, Jiaheng Liu, and Zhaoxiang Zhang. 2026. [AVSCap: Orchestrating audio-visual synergy for omni-modal video captioning](#). *arXiv preprint arXiv:2607.12820*.
- Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024. [Videohalluciner: Evaluating intrinsic and extrinsic hallucinations in large video-language models](#). *arXiv preprint arXiv:2406.16338*.

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

Wenbin Xing, Quanxing Zha, Lizheng Zu, Mengran Li, Ming Li, and Junchi Yan. 2026. [Learning to decode against compositional hallucination in video multimodal large language models](#). *arXiv preprint arXiv:2602.00559*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. [Qwen2.5-omni technical report](#). *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. [Qwen3-omni technical report](#). *arXiv preprint arXiv:2509.17765*.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, Yuming Lou, Dong Yang, Zhijian Liu, Yukang Chen, Amrith Dantrey, Ehsan Jahangiri, Sreyan Ghosh, Daguang Xu, Ehsan Hosseini-Asl, and 13 others. 2025. [Omnivinci: Enhancing architecture and data for omni-modal understanding llm](#). *arXiv preprint arXiv:2510.15870*.

Qilang Ye, Wei Zeng, Meng Liu, Jie Zhang, Yupeng Hu, Zitong Yu, and Yu Zhou. 2026. [When eyes and ears disagree: Can mllms discern audio-visual confusion?](#) In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11955–11963.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, and Jingjing Chen. 2024. [Eventhallusion: Diagnosing event hallucinations in video llms](#). *arXiv preprint arXiv:2409.16597*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

A Benchmark Construction and Evaluation Details

Input normalization. OmniHalluc-L is sourced from public long-form YouTube videos, while all benchmark statistics and model inputs are reported at the source-video level. All evaluations use a fixed, query-agnostic video preprocessing policy.

For models with explicit duration or context constraints, the source video is compressed into the longest input representation accepted by the model interface. For models without such constraints, we sample 256 frames uniformly from the original source video. This normalization is performed once per video and shared by all claims from that video. It never uses evidence spans, claim-specific retrieval windows, pair identifiers, or the paired counterfactual claim.

This protocol separates benchmark scale from model-interface constraints. The dataset statistics describe the original long-form videos, while inference operates on a normalized input chosen only by the model’s context budget. The frame-count analysis uses per-model, subtype-balanced diagnostic subsets rather than the full benchmark, and should be read as a controlled trend study of visual coverage rather than as a replacement for the main full-benchmark results.

Answer normalization. All benchmark items are evaluated as binary YES/NO claim verification. For open-weight Omni models, we do not rely on free-form answer parsing. Instead, under the same verification prompt, we read the logits of the canonical YES and NO answer alternatives and compute the binary margin

$$m_{\text{YES}} = \ell_{\text{YES}} - \ell_{\text{NO}},$$

together with the softmax-normalized support score

$$p_{\text{YES}} = \frac{\exp(\ell_{\text{YES}})}{\exp(\ell_{\text{YES}}) + \exp(\ell_{\text{NO}})} = \sigma(m_{\text{YES}}).$$

Here ℓ_{YES} and ℓ_{NO} are the model logits for the two canonical answer alternatives. Raw AV uses this binary score for the native decision, while MAD and MPRC use the score, margin, and branch-induced changes as reliability features before applying their corresponding decoding or calibration rule.

For systems that expose only generated text, such as closed-source APIs, we use the same YES/NO verification prompt and parse the response deterministically. The parser normalizes case, whitespace, and punctuation, and accepts an answer only when the response contains an unambiguous YES or NO decision. If the response is verbose, contains both labels, refuses to answer, or lacks a parseable binary decision, we retry the same query up to three times. If all three attempts

Name	API model code	Status	Temp.
Gemini 3.1 FL	gemini-3.1	Stable	0
	-flash-lite		
Gemini 3.0 Flash	gemini-3	Preview	0
	-flash-preview		
Gemini 3.1 Pro	gemini-3.1	Preview	0
	-pro-preview		

Table 6: Closed-source Gemini API identifiers and decoding settings used for the Raw AV references. FL denotes Flash-Lite.

remain unparseable, the item is marked incorrect. Because these systems do not expose logits or confidence features, they are reported only as Raw AV references in the main tables. For reproducibility, Table 6 reports the exact API model identifiers used for the closed-source references in Table 2. The model names in the main tables are display names, while all API calls are issued to the listed Gemini API model codes. All closed-source calls use the same YES/NO verification prompt as open-weight models, with temperature set to 0 when supported. We disable external tools, search grounding, code execution, and URL context, and request text-only outputs from audio-visual inputs. Since these APIs expose generated text rather than YES/NO logits or confidence scores, they are evaluated only as Raw AV references and are not used by MAD or MPRC.

This rule is conservative: unparseable outputs cannot accidentally receive credit, and a hidden pair can be scored correct only when both of its claims yield parseable and correct binary decisions. The same parser is applied uniformly within each closed-source system, and no parser decision uses the paired counterfactual claim, evidence spans, or pair identity.

Implementation and efficiency notes. All YES/NO verification runs use the same claim wrapper and score only the answer alternatives. For Qwen2.5-Omni-7B, we use 256-frame inputs under the model’s 32K-token context budget with greedy one-token decoding. Qwen3-Omni-Instruct uses 256-frame inputs with a 64K-token context budget and YES/NO logit scoring. MiniCPM-o 4.5 uses its validated 64-frame audio-video setting with forced-logit YES/NO scoring, while OmniVinci uses 256-frame video input with forced logits. These settings reflect model-interface constraints rather than claim-specific preprocessing.

Efficiency diagnostics are reported within

model backends rather than as cross-hardware speed comparisons. Qwen2.5-Omni timings are measured on an A100-80G GPU, while Qwen3-Omni timings use an H200 backend. Selected- K MPRC timing sums selected branch calls and should be interpreted as an unfused implementation cost; it does not assume a shared or optimized multi-branch runner.

Subtype taxonomy. Table 7 expands the three relation-defined families used in the main text. The taxonomy is organized by the relation that must be preserved rather than by surface video content. This design allows supported and counterfactual claims to share local evidence while differing only in how that evidence is connected.

Evidence metadata and verification. For each verified pair, we retain evidence intervals as metadata for diagnostic analysis. These intervals support post-hoc studies of evidence distance, family-level behavior, and long-range attribution effects. They are never provided to models during evaluation, calibration, or test-time inference.

Human verification checks that each supported claim is entailed by the source video and that the counterfactual remains plausible but false because of an incorrect binding relation. A valid pair must satisfy three conditions: the two claims share the same underlying audio-visual evidence, the counterfactual differs by the tested binding relation rather than by an unrelated factual change, and the pair does not reveal its label through formatting or answer-side artifacts. This verification step preserves the near-counterfactual structure of the benchmark rather than making the task easier through evidence localization. Each candidate pair was independently reviewed by two annotators against five checklist items: positive entailment, false-but-plausible counterfactuality, shared underlying evidence, binding-only edit, and absence of formatting, language-prior, or single-modality shortcuts. Disagreements (6.9% of pairs) were adjudicated by a third senior reviewer over the evidence interval and its surrounding context. Pre-adjudication agreement on the final accept/reject decision was 93.1% (Cohen’s $\kappa = 0.84$), with per-item agreement averaging 95.2%. Of authored candidate pairs, 31.8% were rejected (68.2% retained), led by shortcut leakage including transcript-answerable items (26.8%), implausible or too-obvious counterfactuals (23.2%), sensitive source content (16.6%), and incomplete entail-

Family	Subtype	Binding relation tested
Cross-Stream Temporal	TI: Temporal Inversion	Event order across audio and visual streams
Cross-Stream Temporal	BM: Boundary Misalignment	Boundary alignment between speech, sound, and visual events
Spurious Co-occurrence	SM: Scene Misassociation	False association between real evidence and the wrong scene
Spurious Co-occurrence	SMB: Synchronous Misbinding	Incorrect local binding among simultaneous cues
Spurious Co-occurrence	CSS: Cross-Segment Splicing	Spurious relation between evidence from different segments
Long-Horizon Attribution	CSSM: Speaker Misattribution	Speaker attribution across distant moments
Long-Horizon Attribution	CSIC: Identity Conflation	Identity consistency across recurring appearances
Long-Horizon Attribution	LRSC: Source Continuity	Continuity of source, speaker, or event across the video

Table 7: Subtype taxonomy used in OmniHalluc-L. Subtypes are grouped by the evidence-binding relation they test.

ment (14.1%), with a tail of ambiguous evidence, uncertain identity, and near-duplicates from overlapping segmentation. Candidate retention, rejection categories, and the agreement statistics ship with the benchmark release.

Annotators. Human workers are involved at two stages of the benchmark construction process. First, trained crowd workers source candidate long-form videos according to predefined collection guidelines, focusing on duration, audio-visual clarity, and potential evidence for relation-sensitive claim pairs. Second, an additional manual review stage inspects supported/counterfactual claim pairs together with the relevant video context. Reviewers remove samples with ambiguous evidence, incorrect labels, insufficient audio-visual grounding, or overly obvious counterfactual edits.

Annotator instructions. Annotators were given the following verification instructions. For each candidate pair, inspect the source video and the provided evidence intervals before making a decision. Accept a pair only if: (1) the supported claim is directly entailed by the audio-visual evidence; (2) the counterfactual claim is plausible but false; (3) both claims reuse the same underlying evidence; (4) the counterfactual differs only in the tested binding relation, such as speaker, moment, scene, source, or modality; and (5) the answer cannot be inferred from formatting artifacts, common sense alone, or a single modality when cross-modal binding is required. Reject the pair if the evidence is ambiguous, the speaker or event identity is uncertain, the positive claim is not fully supported, the negative claim changes an unrelated fact, the edit is too obvious, or the item contains sensitive/offensive content unsuitable for release. Annotators were also instructed not to add private personal information beyond what is visible or au-

Model	SPA	Item	pY	C	S	L
Qwen2.5-Omni-7B	7.56	53.08	5.75	4.72	9.44	7.73
Qwen3-Omni	6.30	52.26	95.62	6.10	7.09	5.52

Table 8: Text-only sanity checks. SPA is strict-pair accuracy, Item is item accuracy, pY is predicted-YES rate, and C/S/L denote Cross-Stream Temporal, Spurious Co-occurrence, and Long-Horizon Attribution.

dible in the public source material.

B Text-Only Sanity Checks

Table 8 reports text-only probes for the two primary Qwen Omni backbones. The probe removes video and audio evidence while keeping the same YES/NO verification format. Both models fall far below the 25% random-guessing SPA baseline, showing that the benchmark cannot be solved from language priors alone. The opposite predicted-YES rates also indicate that low text-only SPA is not caused by a single shared prior: Qwen2.5 is highly conservative, whereas Qwen3 is strongly YES-biased without audio-visual evidence.

C External Text-Side Baseline

To test whether OmniHalluc-L is solvable from transcripts, we run an ASR-based entailment baseline on the full benchmark (1,800 pairs) under the same single-query, pair-blind protocol: transcripts with speaker diarization are paired with each claim and judged by an LLM-based NLI scorer. Speech-only ASR+NLI reaches 3.78% SPA (item accuracy 51.64%), far below the 25% random-pairing chance level and the weakest evaluated Omni model (25.50%). The pair-level pattern exposes the mechanism: 95.7% of pairs receive the same answer on both sides (95.4% both-NO, 0.3% both-YES)—the transcript provides no signal separating

Branch	Operation	Primary diagnostic role
sh30	Global audio shift by 30s	Light temporal misalignment
sh60	Global audio shift by 60s	Medium temporal misalignment
seg60	Segment-level audio shift, 60s	Local boundary disruption
seg90	Segment-level audio shift, 90s	Longer segment disruption
swap60	Adjacent audio segment swap, 60s	False local co-occurrence
rev	Reverse audio stream	Source continuity disruption

Table 9: Audio-negative branch definitions. Exact branch subsets are selected on training videos within each video-level fold.

a supported claim from its counterfactual. Alternative scorers behave alike: a fixed logit-margin variant reaches 4.83%, and a lexical-retrieval scorer with the opposite answer prior (predicted-YES $\sim 41\%$ vs. 2.5%) reaches 5.39%. Per-family scores are uniformly low (C/S/L: 3.77%/3.23%/4.44%), including Long-Horizon Attribution, supporting the effectiveness of the shortcut-screening criterion in removing transcript-answerable items.

D Audio-Negative Probes and Diagnostic Baselines

Audio-negative probe bank. MPRC uses audio perturbations as reliability probes rather than as replacement evidence. Each branch keeps the visual stream fixed and perturbs only the audio stream, so changes in the model response expose sensitivity to audio-visual binding rather than to visual perception. Table 9 lists the candidate probes used by the reliability feature bank.

The branch bank intentionally remains compact. A larger perturbation pool is not automatically better: overly broad banks can introduce noisy or destructive views, especially when visual perturbations disturb perception rather than binding. This motivates the main design choice of selecting structured audio-negative probes instead of treating all possible modality corruptions as useful calibration features.

Why the main bank is audio-negative. We explored visual and audio-visual perturbation pools during development. Table 10 summarizes representative diagnostics. Adding visual perturbations did not consistently improve over the audio-only pool. This is expected: visual perturbations can remove or distort the evidence needed to judge the claim, whereas audio perturbations can stress cross-modal binding while preserving the visual scene.

These diagnostics support the final method design. MPRC focuses on audio-negative reliabil-

Model	Selection setting	Audio	Audio+Video	Observation
Qwen3	Best family in pool	44.26	43.37	Audio pool higher
Qwen3	Duration reliability top- K	43.27	42.77	Audio pool higher
Qwen2.5	Mean- z selection	34.22	32.45	Audio pool higher
Qwen2.5	Best family in pool	35.33	35.33	No added gain

Table 10: Pilot comparison of audio-only and audio+video perturbation pools. Results are diagnostic development runs, not headline benchmark numbers.

ity because it provides a stable intervention on binding while leaving visual evidence intact. The calibration layer can then learn when an audio-induced response shift is informative, without introducing a second source of perceptual corruption from visual negatives.

MAD comparison scope. MAD is a frozen-decoding method, while MPRC keeps the Omni backbone fixed and performs fold-level reliability selection and calibration. We therefore use MAD as the closest frozen-decoding reference and complement it with component controls that isolate native-confidence and audio-negative features. This comparison matches the question studied in the paper: given a fixed Omni backbone, can evidence-binding reliability be improved without updating the audio encoder, visual encoder, or language model?

For protocol consistency, the MAD row in the main table uses the contrastive strength selected within the training split. We also checked the $\gamma = 2.5$ setting used in the MAD paper as a diagnostic. Pilot checks showed that $\gamma = 2.5$ can improve on a small balanced subset but becomes less stable on the larger diagnostic set. We therefore report the training-selected MAD configuration in the main benchmark table and treat $\gamma = 2.5$ as a robustness diagnostic rather than a replacement baseline.

Training-free fixed-branch diagnostic. To isolate whether audio-negative views contain useful reliability signal on their own, we run a fixed-branch diagnostic. In this diagnostic, a single pre-specified audio-negative branch is used directly for scoring. No fold-level selection or reliability calibration is applied. Thus, the diagnostic tests a narrower question than the main MPRC results: whether a fixed audio-negative view can expose evidence-binding signal by itself.

Table 11 reports the diagnostic with Qwen2.5-Omni-7B. Several fixed audio-negative branches outperform MAD without learned calibration. For example, fixed segment swap at 60 seconds

Method / branch	SPA	Δ	Time
MAD	36.33	0.00	20.12s
Fixed audio shift 60s	36.88	+0.55	13.05s
Fixed segment shift 60s	36.90	+0.57	13.20s
Fixed segment shift 90s	36.83	+0.50	12.70s
Fixed segment swap 60s	36.98	+0.65	12.62s
Fixed audio shift 30s	36.26	-0.07	13.78s

Table 11: Fixed-branch diagnostic with Qwen2.5-Omni-7B. “Fixed” means that one pre-specified audio-negative branch is used directly, without fold-level selection or reliability calibration. The result is diagnostic, runs on a subset, and does not replace the full-benchmark main results.

reaches 36.98%, while fixed segment shift at 90 seconds reaches 36.83%, both above 36.33% for MAD. This indicates that audio-negative evidence-binding signal is already present before fold-level reliability calibration.

This diagnostic also explains why the final method does not hard-code a single audio-negative branch. Although several fixed branches outperform MAD, the best branch is not guaranteed to remain the same across data distributions, backbones, and output formats. A branch that is useful in one setting can become weaker or slightly harmful in another, as shown by the 30-second shift variant. MPRC therefore treats audio-negative branches as candidate reliability probes rather than as a universal decoding rule. The branch subset, calibration head, and decision threshold are selected only from training videos, while test-time inference remains pair-blind and uses no paired counterfactual or evidence span information.

Together, the fixed-branch diagnostic and the main results support a two-step interpretation. Audio-negative views expose evidence-binding sensitivity, and MPRC converts this sensitivity into a more stable reliability layer through fold-level selection and calibration.

E External MCQ Benchmark Adaptation

We evaluate whether audio-counterfactual reliability is useful beyond OmniHalluc-L’s paired YES/NO format on two MCQ benchmarks introduced in the main text. OmniVideoBench contains 1,000 multiple-choice questions across 628 videos and evaluates broad video-Omni understanding. WorldSense contains 3,172 multiple-choice questions emphasizing real-world audio-visual reasoning across diverse scenarios.

Field	Value
Pair ID	paper__v2_pair_001709
Video ID	rgpTnuqxfZA
Subtype	Cross-segment speaker misattribution
Evidence spans	00:29–00:37 and 18:48–21:20
Temporal gap	1102.90s
Text-only panel correct count	0

Table 12: Metadata for the long-range speaker-consistency case study.

GT	Claim	Gemini
Yes	The late subscription request is spoken by the same bearded desk host who introduced himself as David near the beginning.	Yes
No	The late subscription request is spoken by a different person than the bearded desk host who introduced himself as David near the beginning.	Yes

Table 13: Gemini 3.1 Pro accepts both sides of the hidden pair. MPRC predicts the pair correctly with Qwen2.5, Qwen3, and MiniCPM.

The MCQ adaptation does not convert each option into an OmniHalluc-L-style binary claim. Instead, each option is represented by choice-score features from the native audio-visual view and by option-score shifts under audio-negative views. The feature map includes the raw option score, centered score, margins against competing options, audio-counterfactual option scores, and branch-induced deltas. MPRC uses these features for audio-counterfactual MCQ decoding. MPRC-Cal adds video-level logistic calibration on the target benchmark, and Cal.+Opt. further adds option-position features.

This setup is intentionally a format adaptation rather than a zero-shot transfer claim. The Omni backbone remains frozen, and all target-benchmark adaptation is performed within video-level folds before evaluation on held-out videos.

F Qualitative Cases and Closed-Source Error Analysis

Long-range speaker-consistency case. This case illustrates why strict-pair scoring is necessary for long-form Omni hallucination. The pair is a Long-Horizon Attribution example with a temporal gap of 1102.90 seconds. The evidence appears near the beginning, where a bearded desk host introduces himself as David, and near the end, where the same on-camera host delivers the subscription and support outro.

The failure is not a lack of local evidence. The

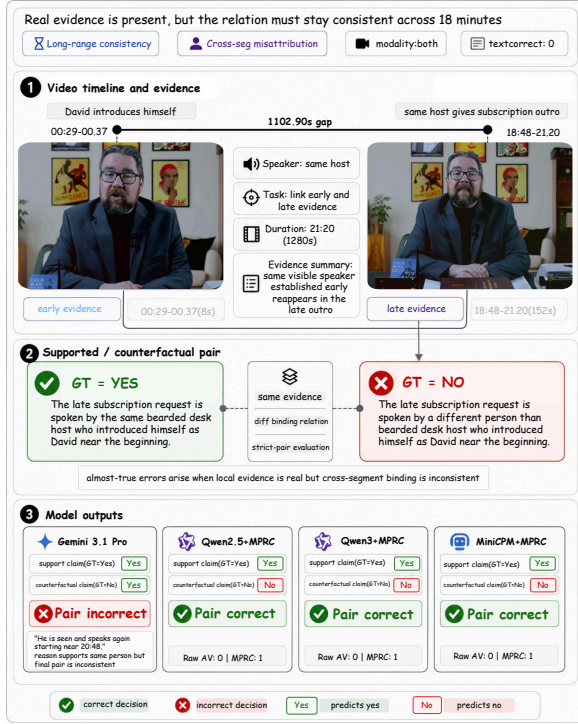


Figure 6: Long-range speaker-consistency case study. The same visible host appears in early and late evidence spans separated by 1102.90 seconds. Gemini 3.1 Pro accepts both the supported and counterfactual claims, producing an inconsistent pair, while MPRC with three open-weight backbones accepts the supported claim and rejects the counterfactual.

model can recognize relevant fragments, but the pair-level relation is inconsistent: accepting both claims means the same late utterance is treated as both spoken by David and not spoken by David. MPRC corrects this pair by using audio-negative reliability features that penalize unstable long-range speaker binding.

Closed-source error type analysis. Figure 7 analyzes residual errors from two closed-source references. The dominant output-only attribution category is temporal and boundary localization, accounting for 46.3% of Gemini 3.0 Flash errors and 37.7% of Gemini 3.1 Pro errors, followed by audio-visual alignment errors. Speaker, identity, source, scene, and event confusions are smaller in isolation, but they remain important because they often determine whether a locally plausible claim is bound to the correct entity or moment. The explanation-cue heatmap shows a similar pattern: both models frequently mention temporal, audio, and visual cues, but cover identity/source and scene/event cues much less often. Thus, even strong closed-source systems often retrieve rele-

vant evidence while under-specifying the relation that makes the claim true or false. This supports the central design of OmniHalluc-L: strict-pair scoring is needed because long-form hallucination is frequently a failure of relational attribution, not a failure to mention evidence.

G Prompt Templates

This section records the prompt templates used for claim verification, modality diagnostics, and MCQ adaptation.

YES/NO claim verification.

Is the following statement about the video correct: {query} Answer only ‘Yes’ or ‘No’. Do not include any explanation.

Modality diagnostic probe.

Question: {claim_question}
To answer this question, which modality is needed (audio, video, or both):

Qwen system message.

You are Qwen, a virtual human developed by the Qwen Team, Alibaba Group, capable of perceiving auditory and visual inputs, as well as generating text and speech.

MCQ direct decoding.

Watch and listen to the video carefully. Answer the multiple-choice question using the video and audio evidence. Select exactly one option. Do not explain.

Question: {question}
Options:
(A) ...
(B) ...
(C) ...
(D) ...
Final answer (A, B, C, D):

MCQ option-score prompt.

Watch and listen to the video carefully. Decide whether the candidate answer correctly answers the question using the video and audio evidence. Do not explain.

Question: {question}
Candidate answer: {option_text}
Final answer (Yes or No):

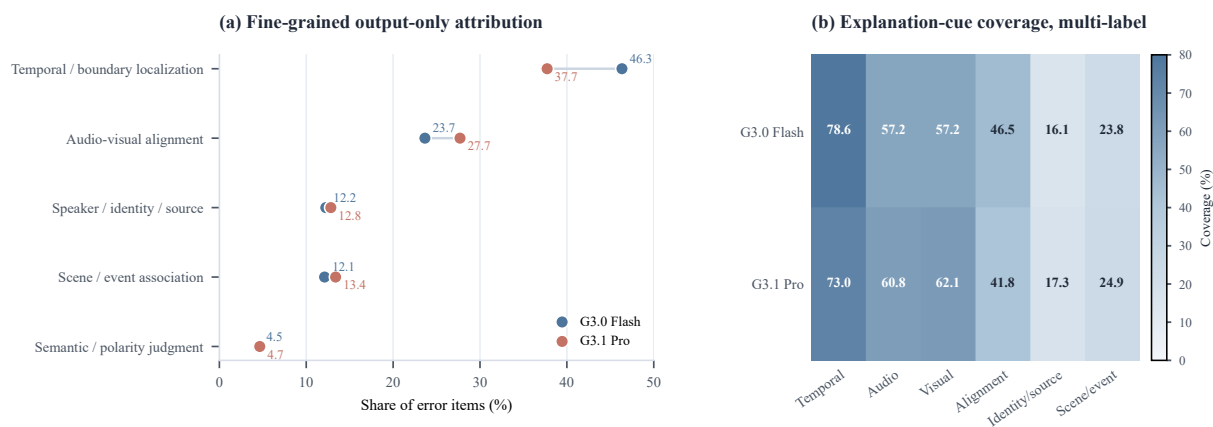


Figure 7: Closed-source error type analysis. Panel (a) attributes residual errors to fine-grained output-only categories, and panel (b) reports multi-label explanation-cue coverage. Errors concentrate on temporal/boundary localization and audio-visual alignment, while identity/source and scene/event cues are less consistently covered.

H Release, License, and Privacy

The MPRC code is available, and we will release the benchmark annotations, prompts, and evaluation scripts for research use. The released benchmark metadata includes source video identifiers, claim text, labels, pair membership, subtype annotations, and diagnostic metadata required for evaluation. Raw source videos remain governed by the terms of their original hosting platform, and the benchmark release does not redistribute raw video content.

The benchmark is constructed from publicly accessible long-form videos and does not add private personal metadata beyond what is already visible or audible in the public source material. Annotations are intended for model evaluation and reliability research. When source videos become unavailable, the corresponding identifiers can be excluded or updated in downstream evaluation following the released protocol.