Senses Wide Shut: A Representation–Action Gap in Omnimodal LLMs

Nguyen Quang Trung[1,3]*, **Yiming Gao**[1,2,3]*, **Fanyi Pu**[1,3], **Kaichen Zhang**[1,3], **Shuo Sun**[4], **Ziwei Liu**[1,3]†
[1]Nanyang Technological University [2]Jinan-NTU Green Technology Research Institute, China
[3]LMMs-Lab Team [4]Johns Hopkins University
Project Page · **Code** · **Dataset**

Abstract

When an omnimodal large language model accepts a question whose textual premise contradicts what it actually sees or hears, does the failure lie in perception or in action? Recent omnimodal models are positioned as perception-grounded agents that jointly process video, audio, and text, yet a basic form of grounding remains untested: catching a textual claim that conflicts with the model's own sensory input. We introduce IMAVB, a curated benchmark of 500 multi-minute movie clips with a 2×2 design crossing target modality (vision, audio) and premise condition (standard, misleading), which lets us measure conflict detection separately from ordinary multimodal comprehension. Across eight open-source omnimodal LLMs and Gemini 3.1 Pro, we document a *Representation–Action Gap*: hidden states encode premise–perception mismatches reliably for vision and more weakly for audio, even when the same models almost never reject the false claim in their outputs. Behaviorally, models fall into two failure modes: *under-rejection*, in which they answer misleading questions as if the false premise were true; and *over-rejection*, in which they reject more often but also reject standard questions, sacrificing ordinary comprehension accuracy. The gap is modality-asymmetric (models reject false audio premises far less often than false visual ones) and prompt-resistant across seven variants. As an initial diagnostic intervention, a probe-guided logit adjustment (PGLA) re-injects the encoded mismatch signal into decoding and consistently improves rejection behavior at a bounded cost to standard accuracy. Together, these results suggest the bottleneck for omnimodal grounding lies in acting on the encoded mismatch signal rather than in encoding it.

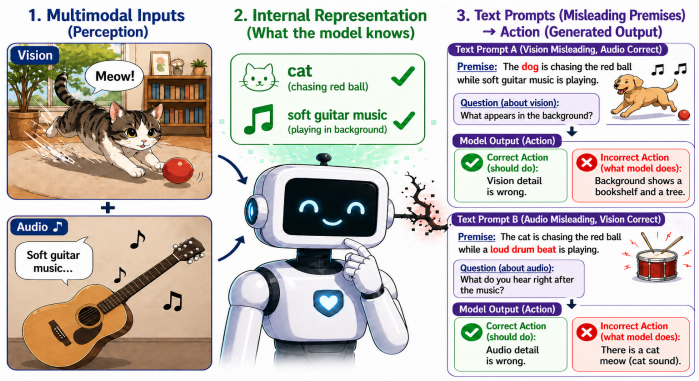


Figure 1: Overview of the Representation–Action Gap on IMAVB.

*Equal contribution. †Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Omnimodal large language models (LLMs) [1–3] are increasingly positioned as perception-grounded agents: they jointly process video, audio, and text, and are expected to understand textual claims against their own sensory input. Yet a basic form of grounding remains untested. *If a question embeds an incorrect claim about what the model sees or hears (for example, the video shows a character in a maroon shirt but the question refers to a "blue" one), can the model catch the mismatch, or does it blindly trust the text?* Silent compliance with false claims about the environment is a failure mode that cooperative benchmarks cannot surface, and it is central to any deployment that requires grounding against reality. The question behind this failure is sharper: when a model accepts such a false premise, is the failure one of **perception** (the mismatch was never detected) or one of **action** (a detected mismatch failed to propagate to the output)?

Recent interpretability work on text-only LLMs has shown that models routinely encode information they do not express: internal states distinguish truthful from untruthful statements [4–6], linearly encode refusal and honesty directions [7, 8], and predict answer correctness before generation [9]. Concurrent work extends this internal–external dissociation to vision-language models (VLMs): VLMs linearly encode visual evidence even when their answers ignore it [10–12]. Whether this dissociation extends to **omnimodal** models that jointly process video, audio, and text is unknown; whether it is **modality-asymmetric** across vision and audio is unknown; and whether it resists prompt-level interventions and remains actionable under a lightweight inference-time correction has not been tested.

A parallel literature on false-premise questions [13, 14], sycophancy [15–19], and honesty [20, 21] shows that models often comply with user-asserted claims even when they have evidence to refuse, but this line of work has not yet connected behavioral sycophancy to its internal-state signature in the cross-modal perceptual setting. Existing cross-modal benchmarks cannot bridge the two: they either remove one modality [22–25], assume cooperative premises [26–29], or prompt the model to verify explicitly [30, 31]. A detailed review is in Appendix K.

To ask the perception-vs-action question, we construct IMAVB, an evaluation harness of 500 multi-minute movie clips (60–300 s) arranged in a 2×2 design over modality (vision, audio) and premise (standard, misleading). The video and audio content remain unaltered; only the textual question varies. Misleading variants copy the standard query but swap exactly one premise detail (e.g., "maroon shirt" $\rightarrow$ "blue shirt"), expecting the model to reject via option E or F. Across eight open-source omnimodal LLMs and Gemini 3.1 Pro, we find a *Representation–Action Gap*. Behaviorally, seven of eight open-source models *under-reject*: they handle 40–75% of standard questions but reject misleading ones in only $\leq$16.2% of vision and $\leq$6.6% of audio cases, with four scoring 0% on audio under fixed option order; Qwen3-Omni and Gemini 3.1 Pro instead *over-reject* (72.8% and 94.0% on vision), trading 15–25pp of standard accuracy. Internally, however, linear probes recover the misleading premise from hidden states at up to 86%, with a 3–7pp residualized margin above text-only baselines on vision and a weaker margin on audio: the signal is linearly separable yet does not propagate to the output. The gap survives seven prompt variants and stratification by video length, and a probe-guided logit adjustment (PGLA) that re-injects the encoded signal at the output improves balanced accuracy by +9.9pp at a 3pp standard-accuracy budget, and by +15.0pp unconstrained, across all eight open-source models. The diagnostic claim, that the signal is sufficient to shift behavior, already holds at the conservative point.

Contributions. Our contributions are as follows:

- We introduce IMAVB, a 500-clip benchmark of multi-minute movie scenes in a 2×2 design (§2) that combines intact video+audio stimuli, implicit (rather than telegraphed) false premises, and surgical modality targeting; this allows vision-grounded and audio-grounded failures to be measured separately while every modality remains present and competing, a configuration no prior omni benchmark provides.
- We document a *Representation–Action Gap* that extends the text-only internal–external dissociation to *audio* and *joint video+audio (omnimodal)* grounding, and characterize it as *modality-asymmetric* (models reject false audio premises far less often than false visual ones, and the residualized probe signal is weaker for audio), *present across all eight open-source architectures and on Gemini 3.1 Pro with heterogeneous underlying mechanism* (§4.2, §4.3), and *prompt-resistant* across seven variants.

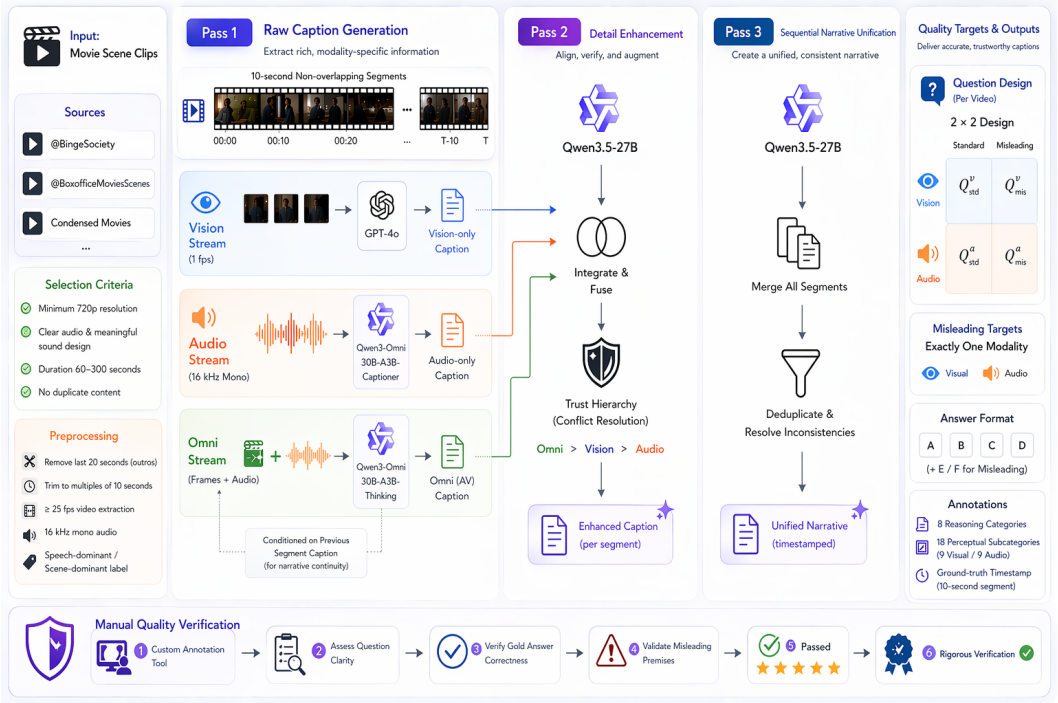


Figure 2: IMAVB three-pass annotation pipeline and 2×2 QA generation design. Full pipeline details and prompts in Appendix I.

- We propose a *probe-guided logit adjustment* (PGLA) that improves mean balanced accuracy by +9.9pp at a 3pp standard-accuracy budget, and by +15.0pp unconstrained, across eight models, providing diagnostic evidence that the encoded signal is sufficient to improve rejection when re-injected at the output.

2 IMAVB: An Omni Benchmark for Perception–Action Dissociation

Asking whether an omnimodal failure arises from perception or from action requires an evaluation harness with five properties that no existing benchmark provides simultaneously: (i) *all modalities must remain intact*, because removing a modality prevents us from measuring cross-modal competition [22–25]; (ii) *false premises must be implicit*, embedded as background assumptions rather than posed as explicit verification prompts that telegraph the test [30]; (iii) *modality targeting must be surgical*, so vision-grounded and audio-grounded failures can be measured separately; (iv) *stimuli must extend over time*, supporting cross-modal tracking rather than instantaneous perception [26–29]; and (v) *symmetric standard controls must exist*, permitting internal-gap measurement against the same stimuli. IMAVB is an omni benchmark design that provides all five.

2.1 Dataset Construction

Why movies? Film pairs audio and vision intentionally: filmmakers compose visual scenes with specific soundscapes, and both modalities carry complementary information that viewers interpret jointly. This makes movies a natural test domain for cross-modal grounding.

Source and Preprocessing. We curate 500 clips (60–300 s, median 140 s, 20.7 hours total) of movie cut-scenes from three sources: *@BingeSociety* [32], *@BoxofficeMoviesScenes* [33], and Condensed Movies [34]. Clips are selected for resolution, audio quality, duration, and uniqueness; video and audio content are preserved unedited. Each video is additionally labeled *speech-dominant* or *scene-dominant* to enable stratified analysis. Full technical preprocessing specifications are in Appendix J.

2.2 Three-Pass Annotation Pipeline

Generating reliable QA pairs for long-form video requires temporally grounded captions faithfully reflecting both what is seen and heard. Our pipeline (Figure 2) uses three sequential caption passes under a trust hierarchy that prevents audio-only captions from over-inferring visual events, followed by QA generation. Full per-pass model details, prompts, and design rationale are in Appendix I.

2.3 Question Design

From each unified narrative, Qwen3.5-27B generates four question variants per video in a 2×2 design: two modalities (vision, audio) $\times$ two conditions (standard, misleading). Each misleading variant targets exactly one modality, so that detection performance directly measures the model's ability to ground against that specific sensory channel when all modalities compete for attention.

Each question has a *premise* describing a specific moment and a *query* about one detail from that moment. The premise may reference both modalities for context, but the query targets exactly one: vision questions ask about something visible (color, gesture, spatial position); audio questions about something audible (dialogue, sound type, background music).

Standard and misleading variants. Standard variants (Q^v_{std}, Q^a_{std}) use correct premises; the answer is one of A–D. Misleading variants (Q^v_{mis}, Q^a_{mis}) copy the corresponding standard question and swap exactly one premise detail (a color, object, speaker identity, or sound type) to something incorrect but plausible while keeping the query identical; the correct answer becomes E ("The visual detail is incorrect") or F ("The audio detail is incorrect"). Figure 3 shows one sample item.

For diversity, the QA generator (Qwen3.5-27B) is prompted to span eight categories (existence, time order, emotional, scene description, cross modality, plot, causal, temporal) and, for misleading variants, to target one of 18 subcategories (nine vision, nine audio) specifying which sensory element the false premise should swap (full taxonomy in Appendix H). Each standard–misleading pair shares the same A–D answer options; options E and F are appended for misleading variants. Each question records a ground-truth *answer timestamp* indicating the 10 s segment containing the evidence, enabling temporal analysis (§4). Variants are evaluated independently.

2.4 Manual Quality Verification

We verified all 2,000 items manually by watching each clip with audio in a custom annotation tool. We give the rubric, the quality statistics, the relabel protocol, and the tool screenshots in Appendix B.

3 Experimental Setup

Models. We evaluate eight open-source omnimodal LLMs: OLA [35], OmniVinci [36], Qwen2.5-Omni [37], MiniCPM-o 2.6 [38], Uni-MoE-2.0-Omni [39], Baichuan-Omni-1.5 [40], Video-SALMONN-2 [41], and Qwen3-Omni [3]. Selection criteria require support for long video understanding and general audio comprehension (not speech-only, but also sound effects, music, and ambient audio). To test whether the observed patterns extend to proprietary models, we additionally evaluate Gemini 3.1 Pro [2] on the baseline experiment (A1 with fixed and shuffled options). Due to API access and cost constraints, we limit the proprietary evaluation to this single experiment.

Evaluation Protocol. Our primary evaluation is a 6-choice MCQ where A–D are content answers, E is "The visual detail in the question is incorrect," F is "The audio detail in the question is incorrect." We evaluate each sample under both fixed option order and $K{=}3$ random shuffles to quantify position bias effects (Table 1). To validate that the observed failures reflect model limitations rather than benchmark artifacts, we conduct six diagnostic ablations covering prompt interventions and temporal controls (full results in Appendix C). None resolves the core misleading detection failure, confirming that the results are not driven by question format or video length. Rejection is not the only acceptable response to a false premise. However, options E and F appear in every prompt, and three variants ask the model directly whether the premise is accurate (A2, A5) or warn that it may be false (A4), so a model that registers the conflict always has a way to report it.



Audio cue: two men talk normally to each other, over a gentle, melancholic string underscore that fits a sad-scene moment.

	Standard premise	**Misleading premise**
Vision	Q^v_{std} *(D):* When the man in the dark suit faces the younger man wearing a maroon polo shirt, what colour is the dress of the woman in the doorway? A. Black B. Blue C. White **D. Red** E. The visual detail in the question is incorrect. F. The audio detail in the question is incorrect.	Q^v_{mis} *(E):* When the man in the dark suit faces the younger man wearing a blue polo shirt, what colour is the dress of the woman in the doorway? A. Black B. Blue C. White D. Red **E. The visual detail in the question is incorrect.** F. The audio detail in the question is incorrect.
Audio	Q^a_{std} *(D):* As the suited man pleads while a delicate string melody swells, what does he say after "Who cried for the little boy?" A. I will cry for him B. He is lost forever C. It hurts so much **D. He cries inside me** E. The visual detail in the question is incorrect. F. The audio detail in the question is incorrect.	Q^a_{mis} *(F):* As the suited man pleads while a loud drum beat swells, what does he say after "Who cried for the little boy?" A. I will cry for him B. He is lost forever C. It hurts so much D. He cries inside me E. The visual detail in the question is incorrect. **F. The audio detail in the question is incorrect.**

Figure 3: A 2×2 IMAVB sample. Rows target vision and audio; columns hold the standard or misleading premise; the same stimulus (filmstrip and audio cue above) drives both columns of a row. Misleading variants swap exactly one premise detail (red). Red boxes in the filmstrip mark the woman in the red dress (frames 6–8). More examples in Appendix P.

All experiments are conducted using lmms-eval [42], an open-source multimodal evaluation framework (Appendix J). Beyond behavioral evaluation, we conduct representational analysis via linear probing of hidden states and logit lens projection through the unembedding matrix. All experiments use temperature=0 with top-p=top-k=1; we report 95% bootstrap CIs (B=10,000).

Metrics. We report per-split accuracy: **std_v** (standard vision), **std_a** (standard audio), **mis_v** (misleading vision), **mis_a** (misleading audio), and balanced accuracy:

$$\text{Bal} = \frac{1}{2}\left(\frac{\text{std_v} + \text{std_a}}{2} + \frac{\text{mis_v} + \text{mis_a}}{2}\right) \tag{1}$$

4 Results and Analysis

4.1 Baseline Performance

Table 1 surfaces the *under-rejection* mode as a stark asymmetry: models that handle standard questions competently collapse when a single premise detail is wrong. The gap routinely exceeds 60pp. For instance, OmniVinci answers 75.4% of standard vision questions correctly but catches only 6.6% of visual misleads, a failure that cooperative benchmarks never expose. A second gap emerges between modalities: under fixed option order, four of the eight open-source models never once select "audio detail incorrect." Shuffling recovers some latent audio-rejection capacity, but the vision–audio gap persists across every architecture, showing that the failure is systematic rather than caused by position bias.

Two models occupy the *over-rejection* mode instead, but their behavior supports rather than contradicts this pattern. Qwen3-Omni attains the highest balanced accuracy under both option orders by aggressively selecting E and F, at the cost of standard accuracy that falls well below every other open-source model. Crucially, the LLM-as-Judge analysis of its binary-with-explanation outputs (Appendix F) shows the tightest prediction–explanation alignment among all eight open-source

Table 1: Baseline accuracy (%). **Fixed Order**: options A–F in canonical position (E/F always last). **Shuffled**: all six options randomly permuted per sample (K=3 independent shuffles). Bal = balanced accuracy (Eq. 1). ‡Proprietary model; evaluated via API only (no representational analysis).

	Fixed Option Order					Shuffled (K=3, mean ± 95% CI)				
Model	**std_v**	**std_a**	**mis_v**	**mis_a**	**Bal**	**std_v**	**std_a**	**mis_v**	**mis_a**	**Bal**
OLA	71.0	71.6	6.8	0.0	37.4	70.5±1.2	70.3±3.5	2.9±1.5	1.7±0.5	36.4
OmniVinci	75.4	71.4	6.6	0.0	38.4	71.0±2.2	66.1±2.6	4.5±2.2	3.6±0.4	36.3
Qwen2.5-Omni	64.4	69.0	16.0	0.6	37.5	60.1±1.3	64.7±1.6	13.3±3.7	3.4±1.6	35.4
MiniCPM-o 2.6	56.6	54.2	9.0	6.6	31.6	56.4±1.8	54.0±1.8	6.9±1.2	4.9±1.7	30.6
Uni-MoE-2.0-Omni	74.8	69.0	9.0	0.0	38.2	71.0±1.4	67.8±0.4	4.9±2.6	3.3±0.9	36.8
Baichuan-Omni-1.5	66.0	66.8	13.8	0.6	36.8	61.5±2.8	62.6±1.7	9.3±0.9	9.3±1.0	35.7
Video-SALMONN-2	69.8	66.6	16.2	0.0	38.2	64.5±2.6	59.1±1.5	11.1±0.3	15.9±1.8	37.7
Qwen3-Omni	40.6	46.6	**72.8**	**23.6**	**45.9**	52.6±2.5	58.9±2.5	**43.5**±2.5	**30.5**±2.4	**46.4**
Gemini 3.1 Pro‡	50.2	53.8	94.0	48.6	61.6	50.0±3.9	48.0±3.9	91.9±2.1	56.1±3.8	61.5

models, indicating that these rejections are not a random E/F bias. Because the judge shares the model family of the QA generator, however, we treat this result as partial evidence of grounding (Appendix N). The same behavioral pattern appears at proprietary scale: Gemini 3.1 Pro over-selects the rejection options, trades standard accuracy for misleading detection, and shows the same 45.4pp vision–audio gap. The behavioral pattern therefore generalizes beyond the open-source sample.

4.2 Cross-Modal Interference

Since prompt modifications cannot improve misleading detection (Appendix C), we test whether cross-modal interference contributes by running each open-source model in vision-only and audio-only modes. Table 2 reports the change in misleading accuracy when the other modality is removed.

Table 2: Cross-modal interference (pp). A→V = V-only − AV on mis_v; V→A = A-only − AV on mis_a. Positive values indicate interference.

Model	**A→V**	**V→A**	**Interpretation**
Qwen2.5-Omni	**+6.2**	+1.0	Audio interferes
Video-SALMONN-2	**+6.4**	+0.0	Audio interferes
Baichuan-Omni-1.5	**+5.4**	+0.6	Audio interferes
OmniVinci	+1.2	+0.0	Minimal
OLA	−0.6	+0.0	No interference
Uni-MoE-2.0-Omni	−1.8	+0.4	No interference
MiniCPM-o 2.6	−4.4	−5.8	AV synergistic
Qwen3-Omni	+2.0	**+14.4**	Video interferes

The effect is architecture-dependent rather than universal. Three models show clear audio interference on visual detection, with accuracy improving by roughly 5–6pp when audio is removed; two others show only small effects; and OLA and Uni-MoE-2.0-Omni show no interference at all. Two outliers are more informative. MiniCPM-o 2.6 is the only model where joint audio-visual input actively helps, because removing either modality hurts both splits. Qwen3-Omni shows the reverse pattern of the audio-interfering group: removing video substantially improves audio detection, consistent with its being the only model with non-trivial baseline audio rejection. Taken together, no single cross-modal explanation fits all architectures.

4.3 The Representation–Action Gap

Models fail to detect misleading premises regardless of prompt format, modality configuration, or option order. Does a linearly decodable signal for the premise–perception mismatch form in the hidden states at all, or is it absent entirely? We investigate using two complementary interpretability tools: linear probing [6] and logit lens projection [43]. High probe accuracy shows that the stan-

Table 3: The Representation–Action Gap. **HS Probe**: modality-specific linear probe accuracy at peak layer (Eq. 2). **Residualized**: probe accuracy after projecting out text-predictive features (Appendix L). **Behavioral**: misleading detection rate. **Logit Lens**: peak P_l(correct) per split (Eqs. 3–4). †Baichuan-Omni-1.5 is excluded from aggregate probing statistics where noted; see §4.3.

	HS Probe (%)		**Residualized (%)**		**Behavioral (%)**		**Logit Lens P_l(correct)**			
Model	Vis	Aud	Vis	Aud	mis_v	mis_a	std_v	std_a	mis_v	mis_a
OLA	84.0	77.8	79.0	67.7	6.8	0.0	.704	.693	.120	.007
OmniVinci	84.4	78.8	78.9	66.1	6.6	0.0	.005	.007	.001	.001
Qwen2.5-Omni	86.0	75.6	77.7	67.0	16.0	0.6	.685	.687	.122	.021
MiniCPM-o 2.6	83.2	78.6	78.7	69.5	9.0	6.6	.035	.040	.000	.000
Uni-MoE-2.0-Omni	84.4	76.3	76.7	65.6	9.0	0.0	.732	.660	.207	.009
Baichuan-Omni-1.5†	99.3	98.7	99.8	100.0	13.8	0.6	.002	.004	.001	.000
Video-SALMONN-2	83.5	77.1	80.7	65.8	16.2	0.0	.635	.592	.195	.055
Qwen3-Omni	76.5	64.9	79.8	65.0	72.8	23.6	.255	.212	**.833**	**.338**
Text (TF-IDF)	*73.4*	*71.4*	—	—	—	—	—	—	—	—
Text (SBERT)	*66.8*	*57.3*	—	—	—	—	—	—	—	—

dard/misleading distinction is linearly decodable from hidden states; the diagnostic intervention in §5 provides a complementary test of whether this signal is practically useful.

Linear Probing. At each transformer layer $l \in \{1, \ldots, L\}$, we extract the hidden state $\mathbf{h}_l \in \mathbb{R}^d$ at the last token position for each input and train a logistic regression probe:

$$P(y = \text{mis} \mid \mathbf{h}_l) = \sigma(\mathbf{w}_l^\top \mathbf{h}_l + b_l) \tag{2}$$

where y is 1 for misleading splits and 0 for standard splits. Probes are trained via 4-fold stratified group cross-validation grouped by video, so all four query variants of a given video stay in the same fold and identical visual/audio content does not span training and test folds; each fold contains $\sim$125 disjoint videos $\times$ 4 variants = 500 samples. The layer achieving the highest mean CV accuracy is designated l^*. Modality-specific probes (Table 3, "HS Probe" columns) train separate vision and audio classifiers, each using the target misleading split as positives and all standard samples as negatives at a 1:2 class ratio with stratified splits. Full details are in Appendix L. Table 3 presents a large and systematic gap between what models represent and what they output.

Text confound control and residualization. Because misleading and standard questions differ by a single premise swap, lexical features alone may be predictive. Text-only baselines establish this ceiling: TF-IDF reaches 73.4% on vision and 71.4% on audio, and sentence-BERT reaches 66.8% and 57.3%. To isolate the multimodal component, we project out text-predictive features from hidden states via Ridge regression with orthogonal projection (nested 4-fold group CV grouped by video; Appendix L) and retrain modality-specific probes on the residual. Residualized probe accuracy remains above the SBERT baseline for every model and modality and above the TF-IDF baseline for vision (paired bootstrap on per-sample held-out predictions, $p < 1/B$, B=10,000), losing only 1–4pp to text removal under within-modality probing; for audio, residualized accuracy falls below the TF-IDF ceiling for several 7B models, indicating that a portion of the audio signal may be captured by lexical features and that the audio-side evidence for genuinely multimodal encoding is weaker than the vision-side. For the vision modality, this persistence after removing text-predictive variance addresses the concern that linear probes may detect surface features rather than task-relevant encoded information [44, 45].

Modality-specific probing. With the text confound resolved, we turn to the vision–audio asymmetry. We train separate probes for vision and audio questions; Table 3 presents the results.

Residualized probes show a 9–15pp internal gap across the six standard 7B models (Baichuan-Omni-1.5 is excluded because its layer-2 signal is near ceiling and uninformative for asymmetry analysis). This gap exceeds the 9.5pp SBERT text-only asymmetry, showing that it arises from multimodal processing and that models under-weight audio-misleading representations rather than simply reflecting text difficulty. However, audio premises inherit the more abstract register of the

audio captions (e.g., a music style rather than a color), so part of this difference may reflect that abstract premises are harder to verify, and not only weaker audio grounding. The gap within each modality, in contrast, does not depend on this comparison, and behavioral misleading detection for these same models still does not exceed 16.2%. The difficulty therefore lies less in *forming* a decodable mismatch signal than in *acting* on it at the output. Here, the probes show that mismatch-related information is linearly decodable after residualization, but they do not show that the model has localized the exact contradictory detail.

Logit Lens. To trace how this signal propagates toward the output, we project hidden states at layer l through the model's final pre-LM-head RMSNorm and unembedding matrix, following the standard logit-lens convention [43]:

$$\mathbf{z}_l = W_{\text{unembed}}\,\mathrm{RMSNorm}(\mathbf{h}_l) \tag{3}$$

and compute the probability assigned to the correct answer token:

$$P_l(\text{correct}) = \mathrm{softmax}(\mathbf{z}_l)[\text{correct token}] \tag{4}$$

The rightmost columns of Table 3 report the logit lens results, revealing two distinct regimes. On the one hand, models like Uni-MoE-2.0-Omni, OLA, and Qwen2.5-Omni produce readable signals on standard splits ($>$0.6) under the logit lens that degrade sharply on misleading ones; the correct answer is present in the representation, aligned with the unembedding mid-stack, but fails to survive to the output. On the other hand, Baichuan-Omni-1.5, MiniCPM-o 2.6, and OmniVinci show near-zero signal at all layers under the vanilla logit lens, indicating that, under this lens, the correct token is not aligned with the model's pre-trained unembedding matrix at any intermediate layer; learned linear probes still decode the misleading signal in these models with 83–99% accuracy (Table 3), so the model's native readout direction is misaligned with the subspace where the signal lives. Notably, Qwen3-Omni reaches 0.833 on mis_v, far above any other model. Taken together, these patterns lead us to describe what the vanilla logit lens shows, under the normalization procedure of Appendix J, as spanning two regimes: a *translation-bottleneck regime* (OLA, Qwen2.5-Omni, Uni-MoE-2.0-Omni, Video-SALMONN-2, Qwen3-Omni), where the correct answer is readable through the unembedding matrix mid-stack but decays before the output; and an *unembedding-misaligned regime* (OmniVinci, MiniCPM-o 2.6, Baichuan-Omni-1.5), where the misleading signal is decodable by a learned probe but never aligns with the unembedding's column space. Probes succeed in both regimes, and under the vanilla lens the difference lies in whether the model's native readout direction picks up that signal. A tuned lens [43] could show more than the vanilla lens does here. The representational claim therefore rests on the probes, not on the lens.

Runner-up analysis. Even a model that gives its best guess may still rank the reject option second, so we also examine the ranking of the output options. To do this, we project the stored final-layer state of each item through the final normalization and unembedding, which recovers its six option-letter logits exactly (Appendix M). Under fixed order, the correct reject option is the runner-up on 23.0% of misleading items for the seven under-rejecting models (25.8% vision, 20.2% audio), which is above the 16.7% uniform-ranking reference, whereas it is the runner-up on only 15.2% of the standard twins. Six of the seven models lie above chance, and the exception is Baichuan-Omni-1.5, whose signal peaks at layer 2 and decays the most. Therefore, a residue of the signal reaches the output logits, but the runner-up does not systematically point to the reject option.

5 Diagnostic Intervention: Probe-Guided Logit Adjustment

The probing and logit lens analyses show that hidden states contain a linearly decodable misleading signal that does not appear in model outputs. However, is this signal merely an incidental correlate of processing misleading inputs, or does it carry actionable information about the premise–perception mismatch? To answer this, we extract the internal signal and feed it back to the output distribution. If the signal is actionable, this intervention should consistently improve misleading detection; if it is an artifact, no reliable improvement should occur. Our approach draws on Inference-Time Intervention [8] and Representation Engineering [7], which use probes to amplify truthful directions in LLM activations, and on contrastive decoding methods [46, 47] that modify logits to reduce hallucination. However, our intervention requires no input perturbation or double forward pass.

Table 4: PGLA results (5-fold cross-validated). **Baseline Bal** = balanced accuracy (%) on the 75% held-out split before intervention. **After PGLA**: **Std/Mis** = mean standard/misleading accuracy after the best configuration; **Bal** = resulting balanced accuracy. **ΔBal** = gain over baseline. Per-split breakdowns in Appendix G.

	Baseline	After PGLA			
Model	**Bal**	**Std**	**Mis**	**Bal**	**ΔBal**
Uni-MoE-2.0-Omni	38.2	54.7	59.3	**57.0**	**+18.8**
MiniCPM-o 2.6	31.6	45.5	52.8	49.2	+17.6
OLA	37.4	58.7	51.2	54.9	+17.5
Video-SALMONN-2	38.2	46.2	64.1	55.1	+16.9
Qwen2.5-Omni	37.5	55.5	51.3	53.4	+15.9
OmniVinci	38.4	53.8	52.2	53.0	+14.6
Qwen3-Omni	45.9	46.5	68.8	57.7	+11.8
Baichuan-Omni-1.5	36.8	52.0	35.0	43.5	+6.7
Mean	*38.0*	*51.6*	*54.3*	*53.0*	*+15.0*

Probe. For each model, we extract hidden states at the last input token of the prefill pass at the peak probe layer l^* (selected by cross-validated binary probe accuracy; see Appendix G for per-model layers). A two-layer MLP (256 hidden units, ReLU) is trained on a 25% split (500 samples) to predict $P_{\text{mis}} = P(y{=}\text{misleading} \mid \mathbf{h}_{l^*})$. The probe uses only a binary label (standard vs. misleading) and receives no information about which modality is affected.

Confidence-gated adjustment. The probe confidence gates a logit boost on the rejection options E and F. Let $g = P_{\text{mis}}^{p}$ be the confidence gate (with power p controlling sharpness), and $\Delta = \max(L_{A\text{-}D}) - \max(L_E, L_F)$ the gap between content and rejection logits. The adjusted logits are:

$$L'_E = L_E + \sigma\big(\gamma(g-\alpha)\big) \cdot (s \cdot \Delta + \delta) - \tfrac{\beta}{2} \tag{5}$$

$$L'_F = L_F + \sigma\big(\gamma(g-\alpha)\big) \cdot (s \cdot \Delta + \delta) + \tfrac{\beta}{2} \tag{6}$$

where σ is the sigmoid function, γ controls gate steepness, α is the confidence threshold, s scales the gap-adaptive component, δ is a fixed boost, and β corrects for E/F positional bias estimated from standard training samples (Appendix G). The sigmoid gate provides graded down-scaling of the boost as probe confidence drops: at the steepest setting in our grid the gate is $\sigma(-2) \approx 0.12$ when the probe is fully confident the input is standard, and at the softest setting it is $\sigma(-0.15) \approx 0.46$. The gate therefore reduces but does not zero out the adjustment on standard inputs, which contributes to the standard-accuracy decrease we report in Table 4. The gap-adaptive term scales the boost by how far the model's logits are from selecting E/F; the debiasing term corrects for systematic positional preference between E and F.

Validation. We evaluate via 5-fold cross-validation on the 75% evaluation set (1,500 samples). For each fold, hyperparameters $\{\gamma, p, \alpha, s, \delta\}$ are tuned on four folds and evaluated on the held-out fifth. Table 4 reports the mean test performance across folds; the tune–test gap averages 0.4pp across models, indicating minimal overfitting.

Table 4 shows that PGLA improves balanced accuracy for all eight open-source models, with a mean gain of +15.0pp at the unconstrained operating point. When we also bound the loss in standard accuracy, the mean gain is +9.9pp at a 3pp budget and +5.4pp at a 1pp budget (Table 11), so the diagnostic claim already holds at a conservative operating point. Moreover, the improvement holds even on models that resisted every prompt-level intervention in Appendix C. Qwen3-Omni is the only model where standard accuracy also improves rather than falls: on standard inputs it heavily over-selects the rejection options, and the debiasing term β corrects this by shifting probability mass back toward content options A–D. The other seven models do not share this bias, so the same gated boost pushes their standard accuracy down while lifting misleading detection. This asymmetry shows that the confidence gate adapts to each model's specific failure mode rather than applying a uniform shift. For the seven models that trade standard accuracy for misleading detection, PGLA is shifting the decision boundary toward rejection rather than improving perceptual grounding. Still, the consistent balanced-accuracy gain across all eight models provides diagnostic evidence that

the internal mismatch signal is actionable rather than an artifact of the probing methodology. This supports the *Representation–Action Gap* thesis: hidden states contain signal sufficient to improve premise rejection when fed back to the output, yet the default output distribution does not reflect this encoding. Appendix G reports more results on this.

6 Discussion

We can now answer the question we opened with: when an omnimodal LLM accepts a premise that contradicts its sensory input, the failure in our experiments lies more in acting on an encoded mismatch signal than in encoding it. Hidden states carry a decodable mismatch signal, reliably for vision and more weakly for audio, but this signal does not reach the output. This finding pushes against a common intuition for this failure mode: that grounding failures of this kind will be solved by richer or larger encoders. For premise–perception verification specifically, the bottleneck has already moved downstream. Omnimodal LLMs in this mode encode the relevant signal but do not act on it, and the pattern holds up against every prompt change, length variation (Appendix E), and modality removal we tried. Even the few models that look better at rejection do so by becoming more cautious overall rather than more grounded, with rejection rates rising only because standard accuracy drops. The way the field currently evaluates these models cannot catch this failure: cooperative benchmarks assume the question describes the input faithfully, so they cannot tell a well-grounded model apart from one that quietly trusts the text. Second, closing the Representation–Action Gap likely calls for training objectives that tie what the model says to what it has already detected internally, rather than further scaling of representations. Candidates include supervised rejection labels, contrastive grounding losses, and honesty-targeted alignment. We view PGLA as a diagnostic step toward that work, not a ready-made solution; benchmark and intervention limitations are catalogued in Appendix N.

7 Conclusion

We documented a *Representation–Action Gap* in omnimodal LLMs: hidden states encode premise–perception mismatches, reliably for vision and more weakly for audio, that the same models fail to act on. These findings extend the internal–external dissociation documented for text-only LLMs into cross-modal perceptual grounding. We release the IMAVB dataset at `https://huggingface.co/datasets/ngqtrung/IMAVB` and the code at `https://github.com/ngquangtrung57/imavb`, and we keep a project page at `https://ngquangtrung57.github.io/imavb-page`.

Acknowledgments and Disclosure of Funding

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20223-0002). This research is also supported by cash and in-kind funding from NTU S-Lab and industry partner(s). This research is also supported by Jinan-NTU Green Technology Research Institute (GreenTRI). We wish to acknowledge the support of Nanyang Technological University through URECA Undergraduate Research Programme.

References

[1] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2026.

[2] Google Blog. A new era of intelligence with Gemini 3, 2026. URL `https://blog.google/products-and-platforms/products/gemini/gemini-3/`.

[3] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, et al. Qwen3-omni technical report, 2025.

[4] Amos Azaria and Tom M. Mitchell. The internal state of an llm knows when it's lying, 2023.

[5] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2022.

[6] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2023.

[7] Andy Zou, Long Phan, Sarah Chen, James Campbell, et al. Representation engineering: A top-down approach to ai transparency, 2023.

[8] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2023.

[9] Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations, 2024.

[10] Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts. *arXiv preprint arXiv:2604.09364*, 2026.

[11] Haruka Kawasaki, Ryota Tanaka, and Kyosuke Nishida. Responses fall short of understanding: Revealing the gap between internal representations and responses in visual document understanding. *arXiv preprint arXiv:2604.04411*, 2026.

[12] Jing Tang, Kun Wang, Haolang Lu, Hongjin Chen, KaiTao Chen, Zhongxiang Sun, Qiankun Li, Lingjuan Lyu, Guoshun Nan, and Zhigang Zeng. Diagnosing knowledge conflict in multimodal long-chain reasoning. *arXiv preprint arXiv:2602.14518*, 2026.

[13] Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. Crepe: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, 2023.

[14] Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. $(QA)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, 2023.

[15] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, 2023.

[16] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023.

[17] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models, 2023.

[18] Shuo Li, Tao Ji, Xiaoran Fan, Linsheng Lu, Leyi Yang, Yuming Yang, Zhiheng Xi, Rui Zheng, Yuran Wang, Xiaohui Zhao, Tao Gui, Qi Zhang, and Xuanjing Huang. Have the vlms lost confidence? a study of sycophancy in vlms, 2024.

[19] Yunpu Zhao, Rui Zhang, Junbin Xiao, et al. Sycophancy in vision-language models: A systematic analysis and an inference-time mitigation framework, 2024.

[20] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022.

[21] Yanxu Zhu, Shitong Duan, Xiangxu Zhang, Jitao Sang, Peng Zhang, Tun Lu, Xiao Zhou, Jing Yao, Xiaoyuan Yi, and Xing Xie. Mohobench: Assessing honesty of multimodal large language models via unanswerable visual questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(34), 2026.

[22] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2023.

[23] Munan Ning, Bin Zhu, Yujia Xie, et al. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models, 2023.

[24] S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TeVAZXr3yv.

[25] Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, 2024.

[26] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025.

[27] Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, et al. Omnibench: Towards the future of universal omni-language models, 2024.

[28] Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities, 2025.

[29] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms, 2025.

[30] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models, 2024.

[31] Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20626–20636, 2025.

[32] Binge Society. Binge society (youtube channel), 2026. URL https://www.youtube.com/@bingesociety.

[33] Boxoffice Movie Scenes. Boxoffice movie scenes (youtube channel), 2026. URL https://www.youtube.com/@BoxofficeMoviesScenes.

[34] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[35] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025.

[36] Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm, 2025.

[37] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, et al. Qwen2.5-omni technical report, 2025.

[38] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, et al. Minicpm-v: A gpt-4v level mllm on your phone, 2024.

[39] Yunxin Li, Xinyu Chen, Shenyuan Jiang, et al. Uni-moe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data, 2025.

[40] Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report, 2025.

[41] Changli Tang, Yixuan Li, Yudong Yang, et al. video-salmonn 2: Caption-enhanced audio-visual large language models, 2025.

[42] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025. doi: 10.18653/v1/2025.findings-naacl.51. URL https://aclanthology.org/2025.findings-naacl.51/.

[43] Nora Belrose, Igor Ostrovsky, Lev McKinney, et al. Eliciting latent predictions from transformers with the tuned lens, 2023.

[44] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks, 2019.

[45] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals, 2020.

[46] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, 2024.

[47] Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding, 2025.

[48] OpenAI. Gpt-4o system card, 2024.

[49] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

[50] Davide Caffagni, Federico Cocchi, Luca Barsellotti, et al. The revolution of multimodal large language models: A survey, 2024.

[51] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in Video-LLMs, 2024.

[52] Xun Zhu, Zheng Zhang, Xi Chen, Yiming Shi, Miao Li, and Ji Wu. Connector-s: A survey of connectors in multi-modal large language models. In *Proceedings of IJCAI-25*, 2025.

[53] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Yue Wang, Yang Yuan, and Hang Zhao. Modality laziness: Everybody's business is nobody's business, 2022. URL https://openreview.net/forum?id=1eGFH6yYAJn.

[54] Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, et al. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio, 2024.

[55] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models, 2023.

[56] Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, et al. The troubling emergence of hallucination in large language models: An extensive definition, quantification, and prescriptive remediations. In *EMNLP*, 2023.

[57] Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models, 2024.

[58] Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models, 2025.

[59] Xu Zheng, Chenfei Liao, Yuqian Fu, et al. Mllms are deeply affected by modality bias, 2025.

[60] Jean Park, Kuk Jin Jang, Basam Alasaly, et al. Assessing modality bias in video question answering benchmarks with multimodal large language models, 2024.

[61] Bin Zhu, Yinxuan Gui, Huiyan Qi, Jingjing Chen, Chong-Wah Ngo, and Ee-Peng Lim. Benchmarking gaslighting negation attacks against multimodal large language models, 2025.

[62] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

[63] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of ACL*, 2024.

[64] Chaeyoung Jung, Youngjoon Jang, Jongmin Choi, and Joon Son Chung. Fork-merge decoding: Enhancing multimodal understanding in audio-visual large language models. *arXiv preprint arXiv:2505.20873*, 2025.

[65] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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*, pages 13418–13427, 2024.

[66] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models, 2024.

[67] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models, 2023.

[68] Mu Cai, Reuben Tan, Jianrui Zhang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models, 2024.

[69] Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S. Ryoo. Understanding long videos with multimodal language models, 2024.

[70] Heqing Zou, Tianze Luo, et al. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding, 2024.

[71] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning?, 2025.

[72] Ruohao Guo, Wei Xu, and Alan Ritter. How to protect yourself from 5g radiation? investigating llm responses to implicit misinformation. In *EMNLP*, 2025.

[73] Mohammadamin Shafiei, Hamidreza Saffari, and Nafise Sadat Moosavi. Multihoax: A dataset of multi-hop false-premise questions. In *Findings of ACL*, 2025.

[74] Yunkai Dang, Mengxi Gao, Yibo Yan, et al. Exploring response uncertainty in mllms: An empirical evaluation under misleading scenarios, 2024.

A Appendix Overview

The appendices provide full implementation details, extended results, and supporting analyses for all experiments reported in the main text. Table 5 lists each appendix with its scope and the main-text section it supports.

Table 5: Appendix overview. "Ref." indicates the main-text section where the appendix is primarily cited.

App.	Title	Contents	Ref.
B	Verification Tool & Statistics	Annotation tool screenshots and aggregated quality metrics over the full 500-video / 2,000-item benchmark (3.92/4 clarity, 3.91/4 answer correctness, 3.93/4 timestamp correctness, 100% valid misleading premises)	§2.4
C	Prompt Intervention Details	Full per-split results for all seven prompt variants (A1–A7)	§4
D	Position Bias Control	Option-shuffling protocol and per-sample consistency analysis	§4.1
E	Temporal Decay Analysis	Accuracy stratified by video duration and evidence position; logistic regression ruling out temporal decay as primary factor	§6
F	LLM-as-Judge Analysis	Prediction–explanation faithfulness results for A5 binary-with-explanation format	§4.1
G	PGLA Implementation	Hidden-state extraction, probe training, hyperparameter grid, cross-validation protocol, Pareto curve, per-split breakdown	§5
H	Category Breakdowns	Per-model accuracy by question type (8 categories) and misleading subcategory (18 categories)	§4.1
I	Prompt Templates	All single-turn evaluation prompts (A1–A7) and annotation pipeline prompts	§3
J	Evaluation Implementation	Per-model input configuration, hidden-state extraction, reproducibility details	§3
K	Related Work	Extended discussion of omnimodal LLMs, modality laziness, inference-time intervention, and evaluation benchmarks	§1
L	Residualized Probe Analysis	Ridge regression + orthogonal projection protocol; per-model residualized probe accuracy; layer-wise probe decay	§4.3
M	Runner-up Analysis	Rank of the correct reject option among the six output logits	§4.3
N	Limitations	Domain and clip length, generation–evaluation family overlap, text baselines, probe specificity, logit-lens scope, PGLA transfer	§6
O	Broader Impact	Positive (safety diagnostics) and negative (adversarial crafting, copyrighted content) societal impacts	—
P	Qualitative Examples	Two additional annotated samples extending the in-text Figure 3	§2.3

B Verification Tool and Statistics

To validate our annotation pipeline, the authors used a custom web-based tool to manually verify every item in the benchmark: all 500 videos $\times$ 4 splits, for a total of 2,000 items. The annotator watches each clip with audio, and for each item the tool presents the video alongside the question and its six options (A–F, where E and F are the misleading-escape options "The visual/audio detail in the question is incorrect"). The annotator rates three criteria on a 1–4 Likert scale, with the full rubric shown inline beneath each radio group so raters never have to recall the anchors.

Rubric.

- **Question Clarity (1–4)**. *Does the question describe a scene detail with a single reading and a single answer?*
  - **4, Clear:** the detail matches the ground-truth scene, there is only one way to read the question, and it leads to one answer.
  - **3, Mostly Clear:** a single intended reading, but phrasing is slightly awkward or verbose.
  - **2, Ambiguous:** admits multiple readings, or more than one of the options is plausible.
  - **1, Unclear:** the question cannot be understood or does not lead to a single answer.
- **Answer Correctness (1–4)**. *How well does the labeled correct option match the caption evidence?* The rubric branches by split:
  - *Standard items (gold A–D):*
    - **4, Exact Match:** the gold option matches the caption detail closely.
    - **3, Close Match:** matches the idea but wording differs.
    - **2, Off by Detail:** partial match that misses specifics (colour, count, object type).
    - **1, No Match:** the gold option does not describe the caption detail.
  - *Misleading items (gold E/F):*
    - **4, Clear Swap:** the premise swap clearly contradicts the caption, so E/F is the right escape.
    - **3, Close Swap:** the swap is wrong, but its wording stays near the ground-truth detail.
    - **2, Swap Off by Detail:** the swap only slightly differs from the ground-truth detail.
    - **1, Swap Accidentally True:** the swap is consistent with the caption, so E/F is inappropriate.
- **Timestamp Correctness (1–4)**. *Does the answer evidence actually live inside the stated 10 s segment?*
  - **4, In-Segment:** the evidence is clearly inside the stated $[X\,s, Y\,s]$ window.
  - **3, Boundary:** the evidence sits near the edge of the segment, very close to an adjacent one, but can still be attributed to the stated segment.
  - **2, Adjacent:** the evidence actually belongs to the neighbouring ± 10 s segment.
  - **1, Misaligned:** the evidence sits in a non-adjacent segment elsewhere in the video.

The phrase "caption evidence" in the answer-correctness rubric names the detail that the item asks about, not the source that the annotator consults: annotators judge every item against the clip itself, so a caption hallucination is visible to them. For misleading items (Figure 5), the tool additionally displays the misleading category (e.g., `person_identity`) and a human-readable description of the detail swap (e.g., "Changed the younger man's shirt color from maroon to blue"); the annotator flags whether the misleading premise is valid (Yes/No).

Relabel pass for flagged items. Any item receiving **Answer Correctness** ≤ 2 or a **Misleading Premise Valid = No** verdict in the initial annotation pass was treated as a candidate defect and re-examined by the authors. For each flag, the authors (i) confirmed the defect by mutual agreement against the video content, and (ii) either manually relabeled the QA item (by rewriting the incorrect distractor text, anchoring the premise swap on a caption-grounded detail), or, when the flag was judged a false positive of the initial rater, restored the original label with a corrected score. All statistics reported below and in the main text are computed *after* this relabel pass, so they reflect the final benchmark state rather than the raw output of the first annotation pass. Only a small fraction of items triggered a flag in the initial verification pass: under **10%** of the 2,000 items were flagged on either correctness or misleading-premise validity (**198/2,000**, 9.9%), and roughly two-thirds of those flags (133/198, 67.2%) turned out to be false positives where the initial rater was systematically strict rather than the item being defective. Only **63 items (3.15%)** actually required a manual QA edit. Because the correction cost scales linearly with this defect rate, the pipeline remains tractable at the 500-video benchmark size and would scale to substantially larger collections without losing quality, which we interpret as evidence of robustness: the pipeline produces items that are right on the first pass the large majority of the time, and the few defective items are easy to identify and repair with the verification tool. All authors discussed every flagged defect item jointly before relabeling or false-positive dismissal.

Post-relabel quality statistics. Aggregated over all 2,000 items, the benchmark is of consistently high quality: mean question clarity **3.92/4** (92.55% at 4, 7.30% at 3, 0.15% at 2), mean answer

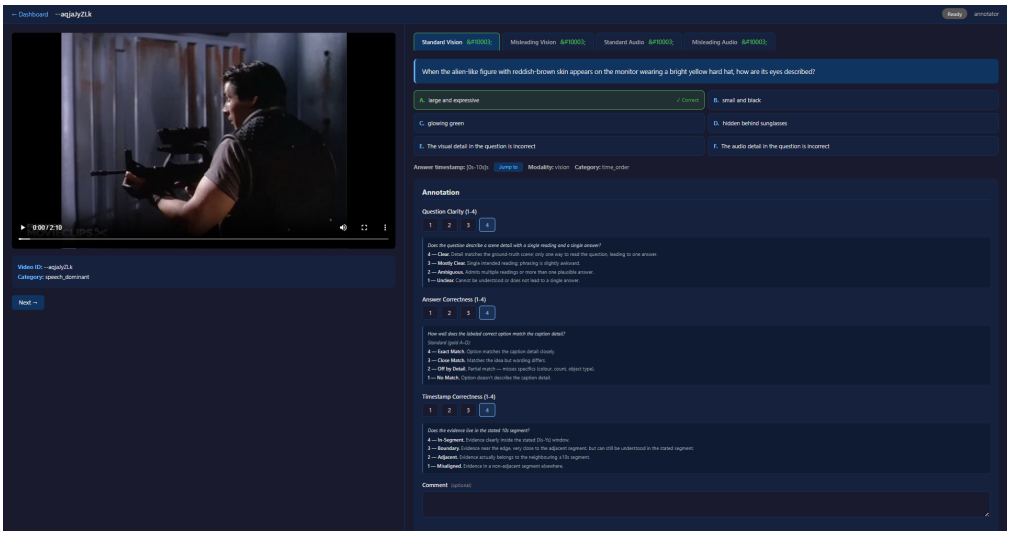


Figure 4: Annotation tool interface for **standard** questions. The annotator views the video, reads the question with six options (A–F), and rates question clarity, answer correctness, and timestamp correctness on a 1–4 Likert scale. The full rubric is shown inline beneath each radio group so raters never have to recall the scale. Metadata shows answer timestamp, modality, and question category.

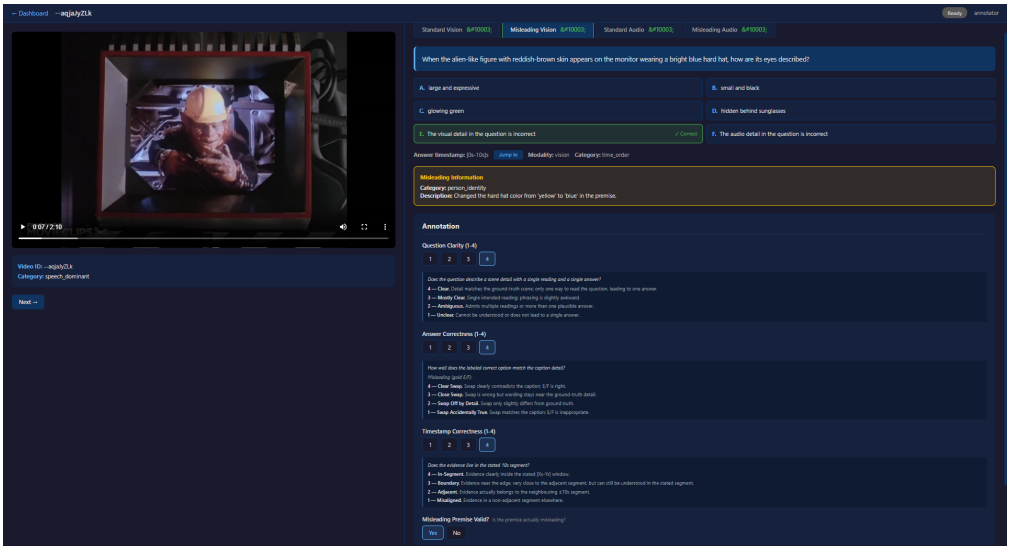


Figure 5: Annotation tool interface for **misleading** questions. In addition to the three 1–4 scales used for standard items, the tool displays the misleading category and a human-readable description of the detail swap. The annotator additionally flags whether the misleading premise is valid (Yes/No).

correctness **3.91/4** (91.45% at 4, 8.55% at 3), mean timestamp correctness **3.93/4** (92.65% at 4, 7.20% at 3, 0.15% at 2), and misleading premise validity **100%** (1,000/1,000 misleading items).

C Single-Turn Prompt Intervention Details

Purpose. The single-turn experiments (A1–A7) systematically vary prompt format to assess how different levels of explicitness affect misleading detection. Each variant is applied to all 500 videos across all four splits, yielding 2,000 evaluation instances per model per variant (inference settings in §3).

Table 6: Single-turn prompt variants (A1–A7). "Tokens" = max new tokens. A1–A5 use the full video; A6/A7 reduce temporal context to test the memory-deficit hypothesis.

ID	Name	Description	Tokens
A1	Baseline	Standard 6-choice MCQ; options shuffled (K=3) to control position bias	16
A2	Binary	TRUE/FALSE question asking whether the stated visual or audio description is accurate	64
A3	CoT	Same as A1 with chain-of-thought instruction; model reasons before selecting a letter	512
A4	Warned	A1 with an explicit warning that the visual or audio premise may be incorrect	16
A5	BinExplain	TRUE/FALSE with a 1–2 sentence explanation; used for LLM-as-Judge faithfulness analysis	256
A6	GT-Only	A1 with video trimmed to the ground-truth answer timestamp only	16
A7	GT+50%	A1 with video trimmed to the ground-truth timestamp $\pm$50% temporal padding	16

Setup. Each model receives video input (frames + audio) alongside the text prompt. The max new tokens parameter controls response length, ranging from 16 tokens (A1, forcing a single-letter answer) to 512 tokens (A3, allowing extended reasoning). All prompts follow the same structure: a system message (model-specific), followed by the multimodal input (video frames + audio waveform), followed by the text question with options. The six options are always labeled A–F, where A–D are content answers drawn from the QA generation pipeline, E is "The visual detail in the question is incorrect," and F is "The audio detail in the question is incorrect." The key differences between variants are:

- **A1 (Baseline)**: The question and six options are presented directly; the model is asked to respond with a single letter. Options are shuffled per sample (K=3 independent shuffles using a deterministic MD5 seed from `hash(video_id + correct_answer + shuffle_id)`); after the model responds, predictions are mapped back to the original letter space via the inverse permutation.
- **A2 (Binary)**: Instead of 6-choice MCQ, the model receives a TRUE/FALSE question: "Is the following visual/audio description accurate: [premise]?" This collapses the task to binary verification, removing the need to select among distractors.
- **A3 (CoT)**: Same 6-choice format as A1 but prefixed with "Think step by step before answering." The model generates up to 512 tokens of reasoning; the answer is extracted as the last valid letter (A–F) in the response.
- **A4 (Warned)**: Same as A1 but with an explicit warning prepended: "Note: the visual or audio premise in this question may be incorrect. Verify before answering."
- **A5 (BinExplain)**: Same binary TRUE/FALSE format as A2 but with "Explain your answer in 1–2 sentences" appended, generating both a verdict and a justification (used for the LLM-as-Judge analysis in Appendix F).
- **A6/A7 (Temporal trimming)**: Same prompt as A1, but the video is trimmed to the ground-truth answer timestamp only (A6) or $\pm$50% temporal padding (A7), testing whether reducing irrelevant temporal context helps the model focus on the evidence.

Interpretation. A1 is the most diagnostic setting because it reveals default trust behavior. Higher-intervention variants (A2–A5) test whether specific modifications can activate latent verification capacity, while A6/A7 test the temporal hypothesis. Table 6 summarizes all seven variants; Table 7 provides the complete per-split results.

Key patterns across variants. Several cross-variant patterns merit discussion. A recurring theme is that prompt interventions (A2–A5) can dramatically increase misleading detection, but this increase typically comes at the expense of standard accuracy, resembling the explicit verification paradigm of AVHBench [30] rather than grounding improvement. Chain-of-thought (A3) reveals extreme format sensitivity: Qwen2.5-Omni's standard accuracy collapses to 9–12% while misleading detection rises to 39–56%, suggesting that these models possess verification circuits that, once activated by extended reasoning, over-trigger and reject both genuine and fabricated premises. OmniVinci under A5 achieves 86.8% on misleading vision but only 22.4% on standard vision, indicating a FALSE-bias artifact rather than genuine improvement. The binary-with-explanation format (A5) is the most diagnostic of

Table 7: Complete single-turn accuracy (%) across all seven prompt variants. Column headers: sv/sa = standard vision/audio; mv/ma = misleading vision/audio.

	A1: Baseline				A2: Binary				A3: CoT			
Model	**sv**	**sa**	**mv**	**ma**	**sv**	**sa**	**mv**	**ma**	**sv**	**sa**	**mv**	**ma**
OLA	71.0	71.6	6.8	0.0	74.0	92.8	34.8	11.8	65.4	66.2	26.0	11.4
OmniVinci	75.4	71.4	6.6	0.0	76.6	77.6	32.0	28.4	72.0	64.6	5.8	4.0
Qwen2.5-Omni	64.4	69.0	16.0	0.6	95.4	99.0	5.8	2.0	9.4	12.0	39.4	56.2
MiniCPM-o 2.6	56.6	54.2	9.0	6.6	75.8	79.2	24.4	20.6	37.2	34.4	30.0	21.8
Uni-MoE-2.0-Omni	74.8	69.0	9.0	0.0	81.4	88.2	24.2	18.0	67.8	60.6	17.0	23.4
Baichuan-Omni-1.5	66.0	66.8	13.8	0.6	95.6	98.8	8.0	1.2	39.0	41.8	40.2	34.0
Video-SALMONN-2	69.8	66.6	16.2	0.0	95.8	98.2	12.0	4.6	44.6	43.4	66.4	5.6
Qwen3-Omni	40.6	46.6	72.8	23.6	63.4	50.2	57.8	61.0	10.4	24.8	89.8	52.0

	A4: Warned				A5: BinExplain				A6: GT-Only			
Model	**sv**	**sa**	**mv**	**ma**	**sv**	**sa**	**mv**	**ma**	**sv**	**sa**	**mv**	**ma**
OLA	68.4	69.4	10.6	0.8	60.0	97.0	46.0	8.4	73.8	72.8	7.2	0.0
OmniVinci	76.2	70.0	13.8	0.2	22.4	22.2	86.8	82.6	78.6	73.6	10.2	0.0
Qwen2.5-Omni	58.4	62.6	40.8	1.0	17.6	46.0	92.2	63.2	71.8	70.2	20.8	0.2
MiniCPM-o 2.6	55.4	52.8	13.8	7.4	55.0	42.8	53.0	57.6	77.2	72.4	5.4	0.8
Uni-MoE-2.0-Omni	72.6	67.6	11.2	0.0	37.8	40.4	69.0	60.6	77.8	69.8	13.6	0.0
Baichuan-Omni-1.5	61.4	63.0	28.0	19.0	65.4	78.2	40.8	21.2	70.8	69.0	15.8	0.4
Video-SALMONN-2	70.0	68.0	22.2	1.4	94.6	99.0	11.6	3.2	72.8	64.8	16.8	0.0
Qwen3-Omni	30.8	46.0	80.0	41.8	23.4	24.4	90.6	82.2	42.2	36.8	76.6	43.2

	A7: GT+50%			
Model	**sv**	**sa**	**mv**	**ma**
OLA	72.2	72.2	6.6	0.0
OmniVinci	77.4	74.8	7.0	0.0
Qwen2.5-Omni	66.8	67.0	20.0	2.0
MiniCPM-o 2.6	70.0	69.8	6.6	0.8
Uni-MoE-2.0-Omni	76.8	70.4	9.0	0.0
Baichuan-Omni-1.5	67.8	70.0	15.4	0.8
Video-SALMONN-2	71.2	67.6	16.0	0.0
Qwen3-Omni	45.4	48.6	72.6	28.8

true understanding: models that genuinely detect the mismatch should produce both a correct rejection *and* a faithful explanation. The LLM-as-Judge analysis (Appendix F) reveals that only Qwen3-Omni achieves tight prediction–explanation alignment under A5 (1.9% right-answer-wrong-reason), while other models frequently select E/F with factually incorrect justifications (4.9–33.2%), confirming that their elevated misleading scores reflect prompt-induced bias rather than perceptual verification. The temporal trimming experiments (A6/A7) produce limited and inconsistent gains, confirming that video duration is not the primary bottleneck.

D Position Bias Control (Option Shuffling)

Setup. All six options are randomly permuted per sample using a deterministic seed (MD5 hash of video_id + correct_answer + shuffle_id). Each model is evaluated with $K{=}3$ independent shuffles, verified to produce a uniform distribution ($16.7\% \pm 1.2$pp per position). Full results with 95% CIs are reported in Table 1.

Interpretation. Shuffling has minimal effect on misleading accuracy ($|\Delta| \leq 2$pp) for five of the eight open-source models. Per-sample consistency analysis reveals that on misleading splits, 75–95% of samples are *never* answered correctly across any shuffle, confirming the failure is systematic rather than positional. Qwen3-Omni shows the largest sensitivity ($\Delta{=}-11.2$pp), suggesting some position-dependent behavior, but even with shuffling it maintains 37.0% misleading detection, far

Table 8: Accuracy (%) averaged across 8 open-source models, stratified by (a) video duration and (b) evidence position. Standard accuracy degrades with duration; misleading detection is flat.

Split	Short	Med	Long	Δ
std_v	75.6	68.4	65.0	−10.6
std_a	68.6	68.6	61.9	−6.7
mis_v	9.8	11.6	10.3	+0.5
mis_a	0.6	1.3	0.9	+0.3

(a) By duration bin.

Split	Early	Mid	Late
std_v	67.8	69.5	67.5
std_a	64.3	64.7	70.6
mis_v	12.7	10.4	10.5
mis_a	1.2	1.3	0.9

(b) By answer position ratio.

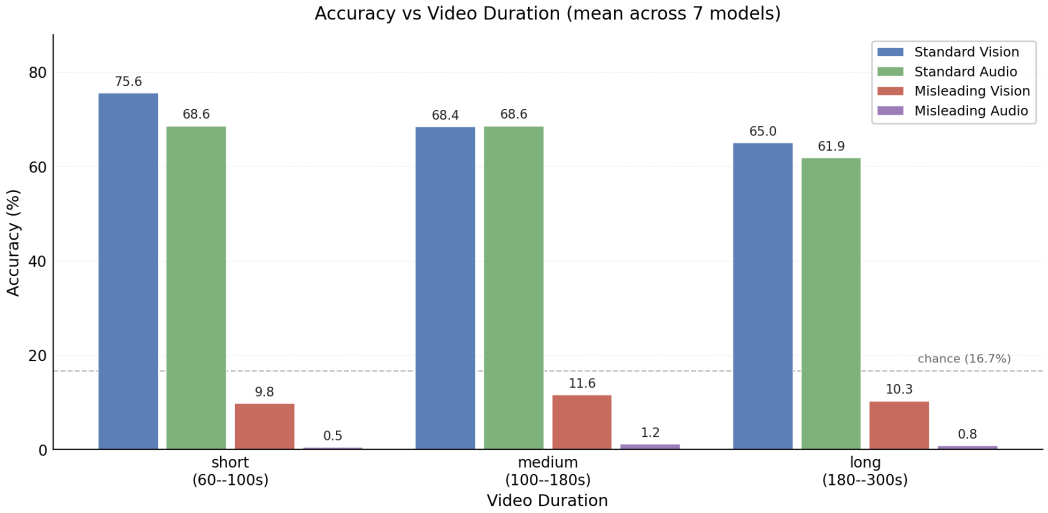


Figure 6: Accuracy by video duration bin, per split. Standard accuracy degrades with duration; misleading accuracy stays flat.

above any other model. Video-SALMONN-2 shows a notable *improvement* with shuffling (+5.4pp), suggesting its audio-misleading detection was partially suppressed by position bias in the default format.

E Temporal Decay Analysis

Purpose. If models struggle more with long videos, misleading detection should degrade with duration or with temporal distance between evidence and query. We test this hypothesis directly.

Setup. Video durations span 60–300s (mean 149s, median 140s). Videos are binned by duration (short: 60–100s, medium: 100–180s, long: 180–300s) and answer position ratio (early: 0–0.33, middle: 0.33–0.66, late: 0.66–1.0). All 16,000 samples (8 open-source models $\times$ 2,000) are included.

Interpretation. Standard comprehension degrades monotonically with duration (-10.6pp vision, -6.7pp audio), consistent with context window pressure. Misleading detection, however, is **flat** across all bins (Tables 8a–8b; Figures 6–7). A logistic regression on all 16,000 samples achieves 76.7$\pm$0.5% 5-fold CV accuracy; the `is_misleading` coefficient dominates ($\hat{\beta}=-1.34$) while temporal variables contribute negligibly ($|\hat{\beta}|<0.1$), indicating that video duration and evidence position do not explain the misleading detection failure.

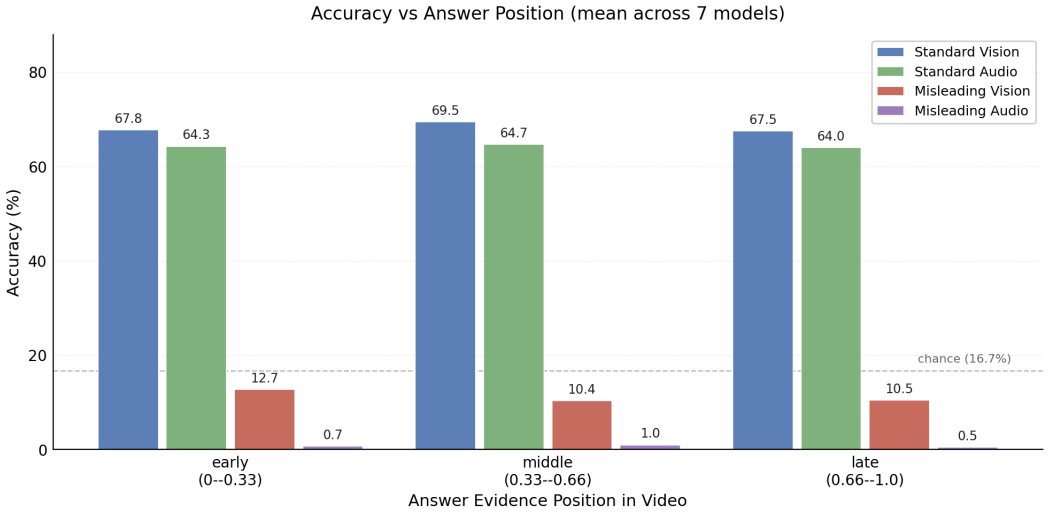


Figure 7: Accuracy by answer evidence position ratio. Misleading detection is independent of where in the video the contradicting evidence appears.

Table 9: LLM-as-Judge results for A5 (16,000 entries, 0 parse errors). W+R=0 across all eight models (so E-Acc equals R+R), reflecting both the common pattern that explanations rarely contradict their own predictions and IMAVB's design, in which producing a correct explanation while predicting wrong would require simultaneously identifying and ignoring the false premise.

Model	P-Acc	E-Acc	R+R	R+W
OLA	52.8	47.9	47.9	4.9
OmniVinci	53.5	38.5	38.5	15.0
Qwen2.5-Omni	54.8	36.6	36.6	18.2
MiniCPM-o 2.6	52.1	18.9	18.9	33.2
Uni-MoE-2.0-Omni	51.9	45.9	45.9	6.0
Baichuan-Omni-1.5	51.4	37.4	37.4	14.0
Video-SALMONN-2	52.1	45.4	45.4	6.7
Qwen3-Omni	55.2	53.3	53.3	1.9

F LLM-as-Judge Prediction–Explanation Analysis

Purpose. The A5 binary-with-explanation format enables us to assess not just prediction accuracy but the faithfulness of the model's reasoning. If models rely on shallow heuristics, they should produce correct predictions paired with incorrect explanations.

Setup. Each model generates a TRUE/FALSE prediction plus a 1–2 sentence explanation for all 2,000 samples. An LLM-as-Judge (Qwen3.5-27B in non-thinking mode) evaluates whether each explanation correctly identifies the relevant evidence. A total of 16,000 entries were judged with zero parse errors (Table 9). Metrics: **P-Acc** (prediction accuracy), **E-Acc** (explanation accuracy), **R+R** (both right), **R+W** (right prediction, wrong explanation).

Interpretation. The R+W column quantifies the dissociation between prediction and explanation. MiniCPM-o 2.6 shows the most extreme mismatch: 33.2% of its predictions are correct but paired with factually wrong explanations, suggesting reliance on shallow pattern matching. OLA (4.9%) and Uni-MoE-2.0-Omni (6.0%) show tight coupling, indicating that when they succeed, the reasoning is generally faithful. Qwen3-Omni achieves the best alignment (1.9% R+W), consistent with its superior misleading detection being grounded in evidence use rather than heuristic shortcuts. The judge (Qwen3.5-27B) belongs to the same model family as the QA generator, so a same-family judge

Table 10: PGLA probe accuracy (%, eval set) at the peak layer l^* used for the confidence-gated intervention. Binary = single-layer MLP; Enhanced = multi-layer PCA + meta-classifier. Peak layers are selected by cross-validated binary probe accuracy.

Model	Layer l^*	Binary	Enhanced
Qwen3-Omni	30	72.3	71.5
OLA	14	71.9	73.3
Uni-MoE-2.0-Omni	16	71.7	71.9
MiniCPM-o 2.6	17	71.6	71.1
Qwen2.5-Omni	17	70.2	70.1
Video-SALMONN-2	15	71.3	70.0
OmniVinci	14	71.5	67.3
Baichuan-Omni-1.5	2	97.0	87.8

that scores outputs on a same-family question distribution could also reward familiarity or stylistic match. We therefore treat this result as partial evidence, not as independent evidence of grounding (Appendix N).

G PGLA Implementation and Extended Results

Purpose. Probe-Guided Logit Adjustment (PGLA) tests whether the internally encoded misleading signal is actionable by feeding it back to steer output logits at inference time, providing a diagnostic test of the Representation–Action Gap.

Hidden state extraction. For each model, we extract last-token hidden states at every transformer layer for all 2,000 samples during a single forward pass. States are stored as tensors of shape $(n_{\text{layers}}, d_{\text{hidden}})$ (e.g., $(29, 3584)$ for OLA and $(48, 2048)$ for Qwen3-Omni) alongside the model's choice logits for options A–F.

Probe training. A two-layer MLP probe (input $\rightarrow$ 256 hidden units $\rightarrow$ ReLU $\rightarrow$ 2-class output) is trained on hidden states at the peak probe layer l^* (Table 10). Training uses 25% of the data (500 samples), with `StandardScaler` normalization and Adam optimization (lr=10^{-3}, 100 epochs). An enhanced variant concatenates hidden states from a window of ± 2 layers around l^*, applies PCA (retaining 99.9% variance), and trains a deeper 3-layer MLP (512 hidden, dropout 0.2); a logistic-regression meta-classifier then stacks single-layer, multi-layer, and deep probe predictions with behavioral features (logit gap, entropy, E/F signal strength).

Confidence-gated intervention. The intervention (Eq. 5–6) has five hyperparameters tuned per model: sigmoid steepness $\gamma \in \{0.5, 1.0, 2.0\}$, probe power $p \in \{1.0, 2.0\}$, confidence threshold $\alpha \in \{0.3, 0.5, 1.0\}$, gap scale $s \in \{0.75, 1.0, 1.5\}$, and fixed boost $\delta \in \{5, 8, 12\}$. The debiasing term β is computed as the mean E$-$F logit difference on standard training samples; it corrects for systematic positional preference between the two rejection options. This yields $3 \times 2 \times 3 \times 3 \times 3 = 162$ configurations per model, swept exhaustively during each cross-validation fold.

Cross-validation protocol. We use 5-fold stratified cross-validation on the 75% evaluation set (1,500 samples, $\sim$300 per fold). For each fold: (i) the probe is trained on the fixed 25% training set, (ii) hyperparameters are tuned on four folds ($\sim$1,200 samples), selecting the configuration that maximizes balanced accuracy, and (iii) the selected configuration is evaluated on the held-out fifth fold ($\sim$300 samples). Table 4 reports the mean held-out test performance across five folds.

Pareto analysis. Table 11 shows the tradeoff between balanced accuracy gain and standard accuracy cost at different operating points. At conservative budgets ($\leq$3pp std loss), PGLA achieves +9.9pp mean balanced improvement. As the budget increases, gains grow sublinearly (from +9.9pp at 3pp cost to +15.0pp unconstrained), indicating diminishing returns from additional standard accuracy sacrifice.

Table 11: PGLA Pareto curve (5-fold CV). All rows: mean across all eight open-source models. Constrained rows ($\leq$1–10pp): per model, the configuration with highest balanced accuracy whose standard-accuracy loss from baseline stays within the budget. Unconstrained row: per model, the overall best configuration. ΔBal = balanced accuracy improvement; Uncon. = unconstrained.

Std Budget	Mean ΔBal	Mean Std	Mean Mis
$\leq$1pp	+5.4	69.0	17.9
$\leq$2pp	+8.0	67.8	24.3
$\leq$3pp	+9.9	66.9	29.0
$\leq$10pp	+15.9	60.7	47.1
Uncon.	+15.0	51.6	54.3

Table 12: Per-split accuracy (%) after PGLA intervention (best configuration per model, evaluated on the 75% split). Baselines in Table 1.

Model	std_v	std_a	mis_v	mis_a	Bal
Uni-MoE-2.0-Omni	64.8	44.7	54.4	65.8	57.4
MiniCPM-o 2.6	66.4	54.0	48.3	54.4	55.8
OLA	64.5	52.9	53.8	48.5	54.9
Qwen3-Omni	50.8	21.4	66.0	87.6	56.5
Video-SALMONN-2	59.1	38.1	56.2	69.0	55.6
OmniVinci	67.7	48.9	47.0	53.1	54.2
Qwen2.5-Omni	64.0	47.1	47.8	56.3	53.8
Baichuan-Omni-1.5	23.9	28.2	67.1	64.5	45.9

Per-split breakdown. Table 12 shows the per-split accuracy after intervention, using each model's best configuration (selected on the 75% evaluation split). The largest gains appear on mis_a, which rises from near-zero to 48–69% for the six standard models. Standard accuracy drops more on the audio side (std_a) than the vision side (std_v) across all models, suggesting the intervention interacts more strongly with the audio processing pathway.

Qwen3-Omni anomaly. As noted in §5, Qwen3-Omni is the sole model where both standard and misleading accuracy improve under PGLA. The mechanism is the debiasing term β: Qwen3-Omni over-selects E by a factor of $3-30\times$ over F, and β redistributes this mass back toward content options A–D. Both constrained and unconstrained optimization select the same configuration, confirming that there is no standard–misleading tradeoff for this model.

Tune–test consistency. Across all eight open-source models and five folds, the mean tune–test gap is 0.4pp (range: 0.1–1.7pp), indicating that the cross-validation protocol effectively controls overfitting.

H Question Type and Misleading Subcategory Breakdowns

Purpose. We categorize questions along two dimensions to identify which types of reasoning and which perceptual domains are most vulnerable to misleading failure.

Setup. Question types span 8 categories (existence, time_order, emotional, scene_description, cross_modality, plot, causal, temporal), assigned during QA generation based on the reasoning required. Misleading subcategories (9 vision, 9 audio) are determined by which perceptual element the false premise targets. Both taxonomies are assigned automatically and verified during manual quality review.

Question type analysis. Table 13 shows that existence and time_order questions are easiest for the open-source models (42–43% for the best), while causal and temporal reasoning are hardest (34–39%), reflecting their greater inferential demands. Gemini 3.1 Pro scores 57–67% across all eight question types with notably smaller variance, ranking first on every category driven by its elevated misleading detection. Speech-dominant videos yield $\sim$4–5pp higher standard accuracy than

Table 13: A1 baseline accuracy (%) by question type and video category. All 4 splits combined. **Bold** = best per row.

	Category	N	OLA	OmniV.	Qwen2.5-O	MiniCPM-o	Uni-MoE-O	Baichuan-O	V-SAL-2	Qwen3-O	Gemini 3.1 Pro‡
Question Type	existence	260	42.3	43.1	42.7	33.8	41.2	41.9	43.5	49.2	**61.4**
	time_order	248	40.7	37.9	37.1	34.7	40.7	35.9	36.3	46.4	**63.1**
	emotional	264	37.5	39.4	36.4	32.6	36.0	35.2	39.0	46.6	**66.7**
	scene_description	260	37.3	37.7	39.2	30.4	40.0	37.3	40.4	45.4	**60.8**
	cross_modality	272	36.4	37.9	37.1	29.8	37.1	35.7	37.1	44.9	**57.1**
	plot	164	35.4	36.6	37.8	29.9	35.4	35.4	36.6	42.1	**62.3**
	causal	260	34.6	37.3	34.2	32.7	38.1	33.5	35.8	47.3	**59.0**
	temporal	272	34.2	36.4	35.7	28.7	36.4	39.0	36.0	44.1	**62.5**
Video	scene_dom (std)	500	69.0	71.0	64.8	53.6	70.2	64.4	65.6	43.6	48.8
	speech_dom (std)	500	73.6	**75.8**	68.6	57.2	73.6	68.4	70.8	43.6	55.2
	scene_dom (mis)	500	3.6	3.6	8.0	8.2	4.6	6.0	8.6	46.8	**66.4**
	speech_dom (mis)	500	3.2	3.0	8.6	7.4	4.4	8.4	7.6	49.6	**76.2**

Table 14: A1 baseline misleading detection (%) by subcategory. **Bold** = best per row. Red = exactly 0%.

	Subcategory	N	OLA	OmniV.	Qwen2.5-O	MiniCPM-o	Uni-MoE-O	Baichuan-O	V-SAL-2	Qwen3-O	Gemini 3.1 Pro‡
Vision	person_position	81	7.4	11.1	18.5	8.6	13.6	13.6	28.4	77.8	**93.8**
	person_action	122	7.4	7.4	17.4	10.7	9.9	17.4	16.5	79.3	**95.1**
	person_identity	90	7.8	8.9	15.6	8.9	8.9	14.4	16.7	71.1	**92.2**
	person_appearance	87	8.0	6.9	17.2	13.8	9.2	14.9	11.5	69.0	**98.9**
	object_location	38	7.9	2.6	23.7	5.3	5.3	10.5	13.2	71.1	**86.8**
	object_type	27	3.7	0.0	7.4	3.7	7.4	7.4	11.1	66.7	**92.6**
	object_attribute	26	0.0	0.0	7.7	7.7	0.0	3.8	11.5	61.5	**96.2**
	location_setting	15	0.0	0.0	6.7	0.0	0.0	13.3	6.7	66.7	**93.3**
	location_detail	14	7.1	0.0	7.1	0.0	14.3	14.3	7.1	64.3	**85.7**
	Vision avg	500	6.8	6.6	16.0	9.0	9.0	13.8	16.2	72.8	**94.0**
Audio	speech_tone	36	0.0	0.0	2.8	11.1	0.0	0.0	0.0	13.9	**58.3**
	sound_type	77	0.0	0.0	0.0	10.4	0.0	2.6	0.0	27.3	**59.7**
	ambient_sound	29	0.0	0.0	0.0	10.3	0.0	0.0	0.0	17.2	**31.0**
	speech_content	95	0.0	0.0	0.0	8.5	0.0	0.0	0.0	18.1	**49.5**
	sound_source	69	0.0	0.0	0.0	5.8	0.0	1.4	0.0	30.4	**47.8**
	sound_intensity	19	0.0	0.0	0.0	5.3	0.0	0.0	0.0	36.8	**42.1**
	speech_context	95	0.0	0.0	1.1	3.2	0.0	0.0	0.0	23.2	**52.6**
	background_music	29	0.0	0.0	0.0	3.4	0.0	0.0	0.0	**34.5**	27.6
	speech_speaker	51	0.0	0.0	2.0	2.0	0.0	0.0	0.0	19.6	**41.2**
	Audio avg	500	0.0	0.0	0.6	6.6	0.0	0.6	0.0	23.6	**48.6**
Vision–Audio Gap			6.8	6.6	15.4	**2.4**	9.0	13.2	16.2	49.2	45.4

scene-dominant ones for open-source models, and on misleading splits video category makes little difference across those models, indicating that the failure to detect false premises is independent of the dominant modality in the video; Gemini 3.1 Pro shows a larger scene–speech gap on misleading splits (66.4% vs. 76.2%) consistent with its higher overall detection rate.

Misleading subcategory analysis. Table 14 shows that within vision, person-related swaps cluster around 7–17% detection for most open-source models, while object and location attributes are harder (0–14%). The audio table is dominated by red cells: most open-source models score exactly 0% across nearly all subcategories. MiniCPM-o 2.6 is the sole 7B model with non-trivial audio detection (3–11%), yielding the smallest vision–audio gap (2.4pp). Qwen3-Omni achieves 61–79% on vision subcategories but only 14–37% on audio, with non-speech elements (sound_intensity 36.8%, background_music 34.5%, sound_source 30.4%) detected more reliably than speech-related claims (speech_tone 13.9%, speech_content 18.1%), suggesting that salient environmental sounds are easier to verify than linguistic content. Gemini 3.1 Pro extends both trends to a much higher absolute level: 86–99% on vision subcategories and 28–60% on audio, with sound_type (59.7%) and speech_tone (58.3%) leading; the vision–audio gap nevertheless remains large (45.4pp), indicating that the asymmetry persists even at proprietary scale.

I Evaluation Prompt Templates

I.1 Single-Turn Prompt Templates

A1: Baseline (max_new_tokens=16)

```
Question:
{question}
```

```
Options:
A. {option_a}  B. {option_b}  C. {option_c}  D. {option_d}
E. The visual detail in the question is incorrect
F. The audio detail in the question is incorrect

Answer with the option's letter from the given choices directly.
```

A2: Binary (max_new_tokens=64)
Vision variant:

```
The following question describes a visual scene from a video:
"{question}"
Based on what you actually SEE in this video, is the visual
description in the question accurate? Does the scene described
actually appear?
Answer with only: TRUE or FALSE
```

Audio variant:

```
The following question describes audio/sounds from a video:
"{question}"
Based on what you actually HEAR in this video, is the audio
description in the question accurate? Do the sounds described
actually occur?
Answer with only: TRUE or FALSE
```

A3: CoT (max_new_tokens=512)

```
Think step by step. First, does the video match the audio/visual
premise? Carefully compare what you actually see and hear against
the description. If visual description is wrong, select E.
If audio is wrong, select F.
[Question + Options as A1]
Reasoning and answer:
```

A4: Warned (max_new_tokens=16)

```
Important: Check closely -- the visual or audio premise may be
incorrect. Consider whether what you see/hear matches the
description before answering.
If visual description does not match, select E.
If audio description does not match, select F.
[Question + Options as A1]
Answer with the option's letter from the given choices directly.
```

A5: Binary Explain (max_new_tokens=256)

Same TRUE/FALSE format as A2 with the addition: "Explain your reasoning in 1–2 sentences."

A6/A7: Temporal Trimming (max_new_tokens=16)

Same prompt as A1; video is trimmed to the ground-truth answer timestamp (A6) or the ground-truth timestamp $\pm 50\%$ padding (A7). This tests whether reducing temporal context helps models focus on the relevant evidence segment.

I.2 Annotation Pipeline Methodology

Figure 2 summarizes the annotation pipeline; here we provide the per-pass details and design rationale.

Pass 1: Raw Caption Generation. Each video is segmented into non-overlapping 10 s clips with three caption streams per segment. **Vision**: GPT-4o [48] receives frames at 1 fps (10 frames) and provides a vision-only caption for each segment. **Audio**: Qwen3-Omni-30B-A3B-Captioner [3] receives the mono audio track and provides an audio-only caption for each segment. This dedicated captioning model runs with no text prompt (only the audio segment is in the chat template), so its output register is not under our control. Because a sound must often be described through an inference

about its source, where a visible object can be named directly, audio captions tend toward a more abstract register than vision captions, and this register carries into the audio premises (Appendix N). **Omni**: Qwen3-Omni-30B-A3B-Thinking [3] receives both 10 frames and audio, with the previous segment's caption as context for continuity and provides an omnimodal caption for each segment.

Pass 2: Detail Enhancement. Qwen3.5-27B [49] integrates all three streams into a unified caption per segment. Because streams occasionally disagree, a strict trust hierarchy governs conflict resolution: *Omni* $>$ *Vision* $>$ *Audio*. Omni captions capture both modalities accurately, and vision captions provide reliable visual detail. Audio captions offer detailed and thorough descriptions of audible content, but without visual context they sometimes attempt to infer visual events and can be incorrect. This hierarchy ensures such guesses do not propagate into the fused caption.

Pass 3: Sequential Narrative Unification. Qwen3.5-27B performs a single-shot global merge of all enhanced segment captions into one unified, deduplicated, timestamped narrative per video, resolving remaining inter-segment inconsistencies.

QA Generation. From each unified narrative, Qwen3.5-27B generates four question variants per video in the 2×2 design (modality $\times$ premise condition). The generator is instructed to target a single modality per question and to swap exactly one premise detail to produce misleading variants.

I.3 Annotation Pipeline Prompts

Level-1 Vision Caption
User prompt:

```
Describe this 10-second video clip.

Focus on visible content only: the people, actions, objects,
setting, and any notable visual changes. Write a short,
concrete caption.
```

Level-1 Audio Caption No explicit text prompt is used. The model receives only the audio segment with the chat template scaffold.

Pass 1 (Omni Caption): First Segment
User prompt:

```
Describe this video clip in detail.

Include:
1. What you see: people, actions, setting, objects
2. What you hear: speech (transcribe it), sounds, music

Write a natural description combining visual and audio.
```

Pass 1 (Omni Caption): Subsequent Segments
User prompt:

```
For context, here is what happened in the last 10s of the video:
{previous_caption}

Use this context to understand continuity (same people, ongoing
actions, conversation flow). Now describe THIS current video clip.

Include:
1. What you see: people, actions, setting, objects
2. What you hear: speech (transcribe it), sounds, music

Write a natural description combining visual and audio.
```

Pass 2: Detail Enhancement
System prompt:

```
You are an expert video caption enhancer. Add specific details
while maintaining the primary source's accuracy.
```

User prompt:

```
You are enhancing a video segment caption by combining details
from multiple sources. Your output must be a clean, natural
description of what happens in the video - NEVER mention the
sources or any conflicts between them.

## Primary Caption (MOST RELIABLE):
{primary_caption}

## Vision Caption (RELIABLE for visual details):
{vision_caption}

## Audio Caption (may have errors - lacks visual context):
{audio_caption}

## Your Task:
Combine these sources into ONE enhanced caption:
1. Primary Caption = TRUTH - Start with this as your base
2. Add visual details from Vision Caption
3. Add audio details ONLY if they match the scene
4. Resolve conflicts internally - Trust Primary > Vision > Audio

## CRITICAL RULES:
- Never reference models, captions, sources, or conflicts
- Write ONLY what a viewer would see and hear
- If sources conflict, choose the most reliable source silently

Enhanced caption:
```

Pass 3: Sequential Narrative Unification

System prompt:

```
You are a novelist adapting a video into prose. Your cardinal rule: describe everything ONCE, then never
      again. The input segments contain massive repetition because they were captioned independently -
     your job is to deduplicate them. Setting details go in the opening [0s-10s] only. Each character's
     appearance is described once on first appearance, then only pronouns or names. Ambient sounds noted
      once, repeated only if they genuinely change. Dialogue lines appear ONCE (they often repeat across
      adjacent input segments - include each line only in its first occurrence). You never reference
     cameras, microphones, audio equipment, recording techniques, sound engineering terms (reverb,
     fidelity, stereo field, sub-bass, frequency, HVAC, soundscape, etc.), captions, models, or any
     technical process.
```

User prompt:

```
You are writing a chapter of a novel based on a video. The video has been captioned in {num_segments}
     overlapping 10-second segments. **IMPORTANT**: These segments were captioned independently by a
     vision+audio model, which means:

1. **Each segment describes the FULL scene** as if the model had never seen prior segments
2. **Settings are re-described** in nearly every segment (walls, furniture, lighting, room layout)
3. **Character appearances are re-described** every time they appear (clothing, hair, features)
4. **Ambient sounds are re-introduced** repeatedly (music, hum, traffic, chatter)
5. **Dialogue lines may appear in 2-3 adjacent segments** because the model doesn't know what was
     already captured

**Your job is to deduplicate these redundant descriptions into a flowing narrative.**

## SEGMENT CAPTIONS (WITH EXTENSIVE REPETITION):
{all_segments}

## YOUR TASK:
Rewrite these as a single flowing narrative with [0s-10s] through [{last_start}s-{last_end}s] timestamp
     markers. Remove ALL repeated descriptions while preserving the timeline.

**CRITICAL STRUCTURE -- THE OPENING vs CONTINUATION RULE:**
- **[0s-10s] (OPENING)**: Establish everything once and for all:
  - **Setting**: Full room/location description (walls, floor, furniture, lighting, windows, decor).
     This is the ONLY segment where you describe the space.
  - **Characters**: Every person's complete appearance (clothing from head to toe, hair color/style,
     distinguishing features, accessories). This is the ONLY segment where you describe what they're
     wearing.
```

```
  - **Ambient sounds**: Background audio that persists (music genre, traffic hum, machine noise, crowd
      chatter). These stay constant unless explicitly changed later.

- **[10s-20s] through [{last_start}s-{last_end}s] (CONTINUATION)**: Write ONLY what is NEW:
  - **Actions and movements** (walking, gesturing, picking up objects)
  - **New dialogue** (each spoken line appears ONCE, in the segment where it first occurs)
  - **Changes to the scene** (new character enters -> describe them once; light turns off -> mention it;
      music stops -> mention it)
  - **Use pronouns/names for characters**: "he", "she", "they", character names, or role labels ("the
      host", "the driver"). NEVER use clothing or hair as identifiers.

**DEDUPLICATION EXAMPLES:**

**BAD (repeating setting):**
"[0s-10s] The couple sits in a dimly lit living room with cream-colored walls and a burgundy sofa. [10s
     -20s] They continue talking in the dimly lit living room with cream-colored walls. [20s-30s] The
     conversation unfolds in the same dimly lit space with burgundy furniture."

**GOOD (setting described once):**
"[0s-10s] The couple sits in a dimly lit living room with cream-colored walls and a burgundy sofa. [10s
     -20s] They continue talking, leaning closer. [20s-30s] She gestures toward the window as he nods."

---

**BAD (repeating character appearance):**
"[0s-10s] A woman in a red jacket and blonde hair speaks. [10s-20s] The woman in the red jacket nods.
     [20s-30s] The blonde woman in the red jacket smiles."

**GOOD (appearance described once, pronouns after):**
"[0s-10s] A woman in a red jacket, her blonde hair pulled back, speaks into the microphone. [10s-20s]
     She nods, leaning forward. [20s-30s] She smiles and glances at the audience."

---

**BAD (repeating ambient sound):**
"[0s-10s] Piano music plays softly in the background. [10s-20s] The piano continues its gentle melody.
     [20s-30s] Soft piano notes fill the air. [30s-40s] The piano music persists."

**GOOD (ambient sound mentioned once):**
"[0s-10s] Piano music plays softly in the background. [10s-20s] She opens the envelope and pulls out a
     letter. [20s-30s] She reads silently, her expression shifting. [30s-40s] She sets the letter down
     and exhales."

---

**BAD (repeating dialogue):**
"[20s-30s] He says, 'Where did you go?' [30s-40s] 'Where did you go?' he asks again."

**GOOD (dialogue appears once):**
"[20s-30s] He says, 'Where did you go?' [30s-40s] She hesitates, then looks away without answering."

**OTHER RULES:**
1. ALL {num_segments} timestamp markers must appear: [0s-10s], [10s-20s], ..., [{last_start}s-{last_end}
     s]. Never skip or combine.
2. Each segment includes both visual AND audio details.
3. Each segment continues from where the previous ended -- flowing narrative, not independent paragraphs.

4. NEVER mention technical terms: "the audio model", "the vision model", "the caption", "recording
     equipment", audio engineering vocabulary (reverb, fidelity, stereo field, sub-bass, frequency, HVAC,
      soundscape, etc.)

Begin writing with [0s-10s]:
```

QA Generation

System prompt:

```
You create benchmark questions to test whether video understanding models truly watch the video or just
     guess from text patterns.

QUESTION FORMAT -- each question is one sentence with two parts:
  PREMISE (describes a specific scene moment) + QUESTION (asks one detail)

Example: "When the man in the red jacket sits down at the table, what does he pick up first?"
         ^^^^^^^^^^^^^^^^ premise ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^  ^^^^ question ^^^^^^^^^^^^

THE 4 VARIANTS you must create:
  Q_std_v -- correct premise + vision question (answer = something you SEE)
  Q_mis_v -- premise with ONE wrong visual detail + SAME vision question
  Q_std_a -- correct premise + audio question (answer = something you HEAR)
```

```
  Q_mis_a -- premise with ONE wrong audio detail + SAME audio question

RULES (violating ANY of these makes the output invalid):

1) MODALITY -- the question part determines the modality.
   Vision = answer is visible: gesture, position, object, expression, movement, color, clothing.
   Audio  = answer is audible: spoken words, sounds, music, tone of voice, volume.
   The premise CAN mention both audio+visual to set context.
   WRONG vision Q: "what does he shout?" -> shouting is audio.
   WRONG audio Q: "what gesture does he make?" -> gesture is visual.

2) MISLEADING = SIMPLE SWAP.
   Write Q_std first, then COPY it to Q_mis and change exactly ONE detail in the premise.
   The question part (everything after the premise) must be WORD-FOR-WORD IDENTICAL.
   Do NOT rewrite the sentence. Do NOT add or remove words in the question part.

   Good:
     Q_std_v: "When the man in the RED jacket sits at the table, what does he pick up?"
     Q_mis_v: "When the man in the BLUE jacket sits at the table, what does he pick up?"
     (Only RED->BLUE changed. Rest is identical.)

   Bad -- question part changed:
     Q_std_a: "After the glass shatters, what sound cuts through the air?"
     Q_mis_a: "After the glass shatters, what soft piano melody cuts through the air?"
     (WRONG: "sound" was replaced with "soft piano melody" in the question part.)

3) PREMISE SPECIFICITY -- must pinpoint ONE moment.
   Include: timestamp [Xs-Ys], character-identifying details, specific action.
   Bad: "When the man walks" -- too vague.
   Good: "At [20s-30s], when the tall man in the gray suit pauses at the doorway"

4) TIMESTAMPS -- use [Xs-Ys] from the caption's 10-second segments.
   answer_timestamp = the exact segment where the ANSWER appears in the caption.
   Double-check against the caption text.

5) CHOICES -- all 4 choices (A/B/C/D) must be UNIQUE and DIFFERENT from each other.
   No two choices can have the same text. The correct answer must directly match
   something stated in the caption -- do not ask for details the caption does not describe.

6) ANSWER GROUNDING -- the correct answer must be a fact DIRECTLY stated in the caption.
   Do NOT ask about details the caption does not mention (e.g., do not ask "what color"
   if the caption only says "large" without naming a color).

7) NO ANSWER IN PREMISE -- the standard question's premise must NOT contain the answer.
   The premise sets the scene; the question asks for a DIFFERENT detail from that moment.
   Bad: premise "man in red jacket" + question "what color is his jacket?" (answer in premise)
   Good: premise "man in red jacket sits at table" + question "what does he pick up?"
```

User prompt:

```
=== VIDEO CAPTION ===
{unified_caption}
=== END CAPTION ===

Create 4 question variants. Correct answer at position {correct_position}.
Question focus (target this reasoning type): {question_focus}
Vision misleading category (pick one): {vision_categories}
Audio misleading category (pick one): {audio_categories}

CRITICAL RULES -- your output will be rejected if any are violated:
- Q_mis_v must be a COPY of Q_std_v with exactly ONE detail swapped in the premise. The question part
     must be word-for-word identical.
- Q_mis_a must be a COPY of Q_std_a with exactly ONE detail swapped in the premise. The question part
     must be word-for-word identical.
- Vision questions: the answer must be something VISIBLE. Never ask about sounds/speech.
- Audio questions: the answer must be something AUDIBLE. Never ask about appearance/position.
- All 4 choices (A/B/C/D) must have UNIQUE, DIFFERENT text. No duplicates.
- The correct answer must be a fact DIRECTLY stated in the caption.

Output ONLY valid JSON:
{
  "shared_intro": "Brief video description",
  "visual_element": {
    "correct_detail": "the real visual detail from caption",
    "wrong_detail": "the swapped-in wrong visual detail",
    "timestamp_range": "[Xs-Ys]"
  },
  "audio_element": {
    "correct_detail": "the real audio detail from caption",
    "wrong_detail": "the swapped-in wrong audio detail",
```

```
    "timestamp_range": "[Xs-Ys]"
  },
  "Q_std_v": "premise with correct visual detail + vision question",
  "Q_mis_v": "SAME sentence with ONE visual detail swapped",
  "Q_std_a": "premise with correct audio detail + audio question",
  "Q_mis_a": "SAME sentence with ONE audio detail swapped",
  "vision_choices": {"A": "...", "B": "...", "C": "...", "D": "..."},
  "audio_choices": {"A": "...", "B": "...", "C": "...", "D": "..."},
  "correct_answer": "{correct_position}",
  "vision_answer_timestamp": "[Xs-Ys]",
  "audio_answer_timestamp": "[Xs-Ys]",
  "vision_misleading": {"category": "from list", "description": "what was swapped"},
  "audio_misleading": {"category": "from list", "description": "what was swapped"}
}
```

I.4 LLM-as-Judge Prompt

The prompt below is used by the LLM-as-Judge analysis (Appendix F) to evaluate prediction-explanation alignment for the A5 binary-with-explanation outputs. The `GROUND TRUTH` paragraph has two branches selected by the gold label of the item; only one branch is shown to the judge per item.

```
You are a judge evaluating whether a model's explanation for a TRUE/FALSE question is correct.

TASK: A model was shown a video and asked whether a description in a question matches what's in the
      video. The model answered {pred} and gave an explanation. You need to judge if the explanation is
      correct.

QUESTION GIVEN TO MODEL:
"{question}"

[Branch A: gold == "FALSE"]
GROUND TRUTH: The description in the question is INACCURATE (FALSE).
The question contains a misleading {modality} premise of type: {misleading_category}.
This means the {modality} event described in the question does NOT actually occur in the video.

[Branch B: gold == "TRUE"]
GROUND TRUTH: The description in the question is ACCURATE (TRUE).
The {modality} event described in the question actually occurs in the video.

MODEL'S RESPONSE:
"{explanation}"

EVALUATION CRITERIA:
1. extraction_correct: Does the model's response clearly state "{pred}" as its answer? (true/false)
2. explanation_correct: Is the model's reasoning valid and consistent with the ground truth? Consider:
   - If ground truth is FALSE (misleading): Does the model correctly identify that the described event
     is inaccurate or doesn't occur? The model doesn't need to identify the exact misleading category,
     but its reasoning should align with the ground truth.
   - If ground truth is TRUE (accurate): Does the model correctly validate that the described event
     occurs?
   - A model that gives the right answer (TRUE/FALSE) but with wrong reasoning should be marked as
     explanation_correct=false.
   - A model that gives the wrong answer but explains something reasonable about the content should
     still be marked as explanation_correct=false (since the conclusion is wrong).
   - If the explanation is empty or nonsensical, mark explanation_correct=false.

Respond ONLY with a JSON object. Keep judge_reasoning under 15 words:
{"extraction_correct": true/false, "explanation_correct": true/false, "judge_reasoning": "short reason"}
```

J Evaluation Implementation Details

Evaluation Framework. As described in §3, all experiments use the lmms-eval framework [42]. We publicly release the benchmark dataset and all code (including hidden-state extraction, probe training, and PGLA). The dataset is available at `https://huggingface.co/datasets/ngqtrung/IMAVB`, and the code is available at `https://github.com/ngquangtrung57/imavb`.

Dataset Preprocessing. The 500 IMAVB clips are drawn from three sources (§2.1) under the following selection criteria: (i) minimum 720p resolution, (ii) clear audio with meaningful sound design, (iii) 60–300 s duration, and (iv) no duplicate content across clips. Preprocessing: the final 20 s of each clip are removed to exclude outros; durations are further trimmed to multiples of 10 s for

segment-aligned captioning; video is extracted at $\geq$25 fps; audio is resampled to 16 kHz mono. Each video is additionally labeled *speech-dominant* or *scene-dominant* based on which modality carries the majority of perceivable information, enabling stratified analysis. Source video and audio content are preserved without further editing throughout.

Multimodal Input Processing. All models receive identical input: 50 uniformly sampled video frames and the full-length audio track resampled to 16 kHz mono. The eight open-source models process these through their official processor and tokenizer; Gemini 3.1 Pro receives equivalent input via the Gemini API. For modality ablation experiments (open-source models only), audio or video tokens are selectively omitted from the input to isolate cross-modal interference effects.

Hidden State Extraction. For representational analysis (linear probing, logit lens), we extract the last-token hidden state at every transformer layer using forward hooks registered on each decoder layer via `register_forward_hook()`. The layer access pattern varies by architecture (e.g., `model.thinker.model.layers` for the Qwen family). Each extraction produces a tensor of shape `(num_layers, hidden_dim)` per sample, saved as `.pt` files. A total of 16,000 hidden-state extractions (2,000 samples $\times$ 8 open-source models) were performed; Gemini 3.1 Pro is excluded from representational analysis as hidden-state access is not available via the API.

Logit Lens Normalization. Consistent with the logit-lens convention [43], we apply the model's final pre-LM-head normalization to every intermediate hidden state before projecting through the unembedding matrix (Eq. 3). All eight open-source models' LLM backbones use RMSNorm as the final pre-LM-head normalization; we load the trained norm weights from each model's `model.norm` module (or the architecturally equivalent module) and apply them uniformly across intermediate layers rather than using per-layer norm weights. This makes layer-wise $\mathbf{z}_l$ values scale-comparable; omitting the normalization produces spurious monotonic trajectories driven by residual-stream magnitude rather than semantic signal. The cross-model and cross-layer $P_l(\text{correct})$ values reported in Table 3 are therefore computed under a common, architecture-consistent procedure.

Reproducibility. All open-source experiments use temperature=0 with top-p=top-k=1; Gemini 3.1 Pro is evaluated via the Google API with temperature=0. Shuffled evaluations use MD5-seeded permutations for reproducibility. For method comparisons, we perform paired bootstrap resampling over the per-sample held-out prediction differences (each evaluation sample appears in exactly one held-out fold under 4-fold stratified group cross-validation; B=10,000); reported p-values reflect the proportion of bootstrap resamples with non-positive mean difference, and values at the resolution limit of B^{-1} are reported as $p < 1/B$. Open-source models are evaluated on 8$\times$ NVIDIA H100 80 GB GPUs.

K Related Work

Omnimodal LLMs. The omnimodal landscape has expanded rapidly [50]. Connector-based approaches combine pre-trained modality encoders with frozen LLMs via learned adapters [51, 52], exemplified by OLA [35], OmniVinci [36], Video-SALMONN-2 [41], and MiniCPM-o 2.6 [38]. Mixture-of-experts designs such as Uni-MoE-2.0-Omni [39] route different modalities through specialized sub-networks, while native omni-models like Qwen2.5-Omni [37] and Qwen3-Omni [3] train end-to-end on interleaved multimodal token sequences. All share the challenge of aligning visual, acoustic, and textual representations in a shared latent space, and as we show, all fail when the textual premise conflicts with sensory evidence, even when their hidden states contain linearly decodable signal for the mismatch.

Modality Laziness and Cross-Modal Hallucination. A growing body of work identifies systematic imbalances in multimodal processing. Modality laziness [53] describes disproportionate reliance on the dominant modality, a pattern that interacts with prompt-driven hallucination [54–56] where misleading textual premises override conflicting sensory evidence. Mechanistic analyses reveal that cross-modal information flow can be disrupted under modality conflict [57, 58], and modality bias pervades MLLM behavior at multiple stages [59, 60]. Recent work on sycophancy in vision-language models [19] and gaslighting negation attacks [61] further demonstrates that MLLMs systematically defer to textual cues even when they contradict visual evidence. We extend this line by showing that

the failure goes beyond behavior: hidden states contain linearly decodable signal for the premise–perception mismatch, yet this signal does not propagate to the output distribution, a finding consistent with concurrent work showing that LLMs' hidden states encode information not reflected in their outputs [9] and that chain-of-thought explanations can systematically misrepresent the true basis of a model's prediction [62]. Three contemporaneous papers report related encoding–grounding dissociations in vision-language models [10–12]; we extend this line to audio and omnimodal grounding and quantify the resulting modality asymmetry.

Inference-Time Intervention and Contrastive Decoding. A parallel line of work addresses hallucination through inference-time logit modification rather than retraining. Visual Contrastive Decoding (VCD) [46] contrasts output distributions from original and noise-perturbed visual inputs; Instruction Contrastive Decoding (ICD) [63] perturbs the instruction instead. AVCD [47] extends this paradigm to audio-visual LLMs using attention-guided modality perturbation, and Fork-Merge Decoding [64] processes modalities separately through early layers before merging. OPERA [65] penalizes over-trusted summary tokens in attention, while Self-Introspective Decoding [66] identifies unimportant vision tokens internally to construct a hallucination signal without a second forward pass. At the LLM level, DoLa [67] contrasts early and late layer distributions to improve factuality, Inference-Time Intervention (ITI) [8] shifts activations along "truthful directions" identified by probes, and Representation Engineering [7] applies population-level reading vectors to steer model behavior. Our PGLA shares the probe-then-intervene philosophy of ITI and RepE but applies it to the audio-visual multimodal setting, using a linear probe to adjust output logits in a single forward pass, avoiding the double-inference cost of contrastive decoding methods.

Multimodal Evaluation Benchmarks. Existing benchmarks vary along two axes: modalities evaluated and whether premises are cooperative or adversarial. Video-only benchmarks such as MVBench [22], Video-Bench [23], and TemporalBench [68] assess visual comprehension without audio. Video-MME [26] includes audio but evaluates only under cooperative premises, while long-video understanding remains challenging even for state-of-the-art models [69, 70]. *Omnimodal benchmarks* incorporate audio but assume cooperative premises: OmniBench [27] tests tri-modal binding, Daily-Omni [28] targets audio-visual temporal alignment, Video-Holmes [71] tests complex video reasoning over suspense short films, and WorldSense [29] covers eight real-world domains. The closest adversarial work, AVHBench [30], targets cross-modal hallucinations and identifies failure modes such as audio-driven visual hallucination. IMAVB differs in three respects: (i) AVHBench uses short clips ($\sim$10 s) with temporally concentrated evidence, while IMAVB uses 1–5 minute videos demanding sustained cross-modal tracking; (ii) AVHBench employs *explicit* verification prompts, while IMAVB embeds false premises *implicitly* as presuppositions testing default trust behavior; and (iii) AVHBench evaluates only behavioral accuracy, while IMAVB adds representational analysis to diagnose *why* models fail.

False-Premise and Adversarial Evaluation. The implicit false-premise design of IMAVB connects to a broader literature on testing models' resistance to flawed assumptions. In the text-only domain, ECHOMIST [72] is the first benchmark for implicit misinformation embedded in queries, and MultiHoax [73] extends this to multi-hop false-premise questions requiring chained reasoning. The Multimodal Uncertainty Benchmark [74] addresses misleading premises in vision-language settings but lacks audio and temporal dynamics. IMAVB generalizes the implicit-premise paradigm to the audio-visual multimodal setting, combining it with multi-minute video and representational analysis (Table 15).

L Residualized Probe Analysis

This appendix details the residualized probing procedure used to isolate multimodal signal from text-predictable variance in hidden-state representations (§4.3).

Method. For each model, we extract last-token hidden states at the peak probe layer l^* and compute 384-dimensional sentence-BERT embeddings (all-MiniLM-L6-v2) for the corresponding question texts. To remove text-predictive features without data leakage, we employ a nested 4-fold stratified group cross-validation protocol grouped by video, so all four query variants of a given video stay in the same fold: within each group fold, a Ridge regression (α=1.0, no intercept) is fit on the

Table 15: Benchmark comparison. A=audio, V=video, I=image. MCk=k-choice MCQ. Long=videos>60 s. AV=both audio and vision required. IMAVB is unique in pairing each cooperative item with a stimulus-matched implicit misleading variant, enabling representational analysis of the standard/misleading distinction on identical inputs.

Benchmark	Modality	#Items	#QA	Format	Long	Premise Type	AV
OmniBench	A+I	1,142	1,142	MC4	✗	Cooperative	✓
AVHBench	A+V	2,136	5,302	Y/N	✗	Explicit verification	✓
WorldSense	A+V	1,662	3,172	MC4	✓	Cooperative	✓
Daily-Omni	A+V	684	1,197	MC4	✗	Cooperative	✓
Video-Holmes	V+A	270	1,837	MC6	✓	Cooperative	✗
IMAVB (Ours)	A+V	500	2,000	MC6	✓	Cooperative + implicit misleading	✓

training split to map hidden states to text embeddings, yielding a weight matrix $W \in \mathbb{R}^{d_{\text{text}} \times d_{\text{hidden}}}$. We compute the SVD $W = USV^\top$ and project both training and test hidden states onto the null space of W via $\mathbf{h}_{\text{res}} = \mathbf{h} - V_k V_k^\top \mathbf{h}$, where V_k retains components with singular values above $10^{-5} \cdot s_{\max}$. A logistic regression probe (C=1.0, L-BFGS, 1000 iterations) is then trained on the residualized training states and evaluated on the residualized test states. Crucially, the Ridge projection is fit only on training data within each fold, preventing any information from test samples from influencing the residualization.

Full results. Table 16 reports the complete per-model breakdown. Across all eight open-source models, residualized probe accuracy remains well above chance (50%) and above the sentence-BERT text-only ceiling (66.8%/57.3%), with the held-out per-sample correctness gap positive for every model and modality (paired bootstrap on per-sample held-out predictions, $p < 1/B$ where B=10,000). Against the stronger TF-IDF baseline (73.4%/71.4%), the residualized margin narrows to approximately 3–7pp for vision, indicating that lexical features capture a portion of the signal. Even so, the consistent residualized advantage over both baselines on vision supports multimodal encoding there. For audio, residualized accuracy falls below the TF-IDF ceiling for several 7B models, so the audio-side evidence is weaker. Baichuan-Omni-1.5 retains near-perfect accuracy (99.8%/100.0%) even after text removal, consistent with its layer-2 peak reflecting raw modality encoder outputs that are trivially separable and remain decodable after residualization. Qwen3-Omni's residualized vision accuracy (79.8%) is comparable to the best 7B models (76.7–80.7%) despite operating at a deeper layer (l^*=30), so its signal also remains decodable after residualization at a deeper layer.

Probe signal decay. The binary probe (standard vs. misleading) peaks at mid-layers and then decays toward the final layer, quantifying the gap between encoding and output discussed in §6. Table 17 reports peak and final-layer accuracy for each model. Across the six standard 7B models the decay ranges from 5.6 to 9.0pp (≈6–9pp); Baichuan-Omni-1.5 decays more sharply (12.6pp) because its near-ceiling signal at layer 2 has many more layers to traverse before the output. One plausible mechanism for this attenuation is exposure bias in training data, under which features that are rarely needed for the output, such as low-level background sound, are filtered out by later layers. Because our layer-wise analysis is not audio-specific, however, it cannot separate low-level audio features in this way, so this direction remains open.

Table 16: Residualized probe analysis (4-fold stratified group CV, grouped by video). l^* = peak probe layer. **Original**: within-modality probe (std_v vs. mis_v for vision; std_a vs. mis_a for audio), which is more conservative than Table 3; Table 3 uses cross-modality negatives (all standard samples as negatives), making it an easier task; the within-modality setup here is the appropriate control for residualization. **Residualized**: probe after projecting out text-predictive features. **SBERT**: sentence-BERT text-only baseline. Δ = residualized $-$ SBERT; all values positive with paired bootstrap $p < 1/B$ on per-sample held-out differences ($B = 10{,}000$). Note that TF-IDF (73.4%/71.4%) is a stronger text-only baseline than SBERT; margins over TF-IDF are approximately 3–7pp for vision. Text features explain at most 1–4pp of probe accuracy for the six 7B models under this residualization, while on audio residualized accuracy falls below TF-IDF for several 7B models.

		Original		**Residualized**		**SBERT Baseline**		**Δ (Resid. $-$ SBERT)**	
Model	l^*	Vis	Aud	Vis	Aud	Vis	Aud	Vis	Aud
OLA	14	82.8	69.6	79.0	67.7	66.8	57.3	+12.2	+10.4
OmniVinci	14	80.4	70.1	78.9	66.1	66.8	57.3	+12.1	+8.8
Qwen2.5-Omni	17	79.4	70.4	77.7	67.0	66.8	57.3	+10.9	+9.7
MiniCPM-o 2.6	17	81.2	70.3	78.7	69.5	66.8	57.3	+11.9	+12.2
Uni-MoE-2.0-Omni	16	78.8	66.9	76.7	65.6	66.8	57.3	+9.9	+8.3
Baichuan-Omni-1.5	2	99.9	100.0	99.8	100.0	66.8	57.3	+33.0	+42.7
Video-SALMONN-2	15	82.5	66.7	80.7	65.8	66.8	57.3	+13.9	+8.5
Qwen3-Omni	30	83.4	66.9	79.8	65.0	66.8	57.3	+13.0	+7.7

Table 17: Binary probe accuracy (%) at peak layer l^* vs. final layer. Decay = final $-$ peak. The misleading signal partially dissipates during late-layer processing across all models, consistent with an output-translation bottleneck rather than an absence of representational signal.

Model	l^*	**Peak**	**Final**	**Decay**
OLA	14	71.9	62.9	-9.0
Uni-MoE-2.0-Omni	16	71.7	63.0	-8.7
MiniCPM-o 2.6	17	71.6	63.9	-7.7
Qwen2.5-Omni	17	70.2	63.8	-6.4
OmniVinci	14	71.5	65.3	-6.2
Video-SALMONN-2	15	71.3	65.7	-5.6
Baichuan-Omni-1.5	2	97.0	84.4	-12.6

M Runner-up Analysis

Purpose. A model that is aligned to give its best guess may still rank the reject option high. We therefore test whether the second-best option points to the correct reject option.

Setup. We reuse the per-item hidden states that we collected for the probing and logit-lens analyses. For every item, this collection stores the hidden state at the answer position of the forward pass at every layer, together with each model's final normalization and unembedding weights. We project the final-layer state through the normalization and the unembedding, which recovers the model's first-answer-token logits over the six option letters exactly, and we apply this to all 2,000 items. We rank the six letter logits and record the rank of the correct reject option (E on misleading-vision items, F on misleading-audio items) under the option order of the stored forward pass (fixed order). If the ranking carried no information, a given option would be the runner-up one time in six (16.7%), and we use this uniform-ranking reference as the chance level.

Columns. **Ranked 1st** is the percentage of the 1,000 misleading items on which the correct reject option has the highest logit. Because greedy decoding outputs the rank-1 option, this column equals the fixed-order misleading accuracy in Table 1, pooled over the two modalities. **Ranked 2nd** is the percentage on which it has the second-highest logit. **1st or 2nd** is the sum of these two columns. **Runner-up on standard twins** is the rank-2 percentage for the same option on the 1,000 standard

twins, where the premise is correct, so a high value there would indicate a generic attraction to the reject option rather than detection.

Table 18: Rank (%) of the correct reject option among the six option logits under fixed option order, sorted by the 1st-or-2nd column.

Model	Ranked 1st	Ranked 2nd	1st or 2nd	Runner-up on std. twins
Qwen3-Omni	48.2	25.7	73.9	17.8
Qwen2.5-Omni	8.3	26.4	34.7	15.6
Video-SALMONN-2	8.1	25.2	33.3	13.9
MiniCPM-o 2.6	7.8	21.4	29.2	14.6
OLA	3.4	24.9	28.3	14.1
Uni-MoE-2.0-Omni	4.5	23.6	28.1	16.8
OmniVinci	3.3	22.7	26.0	15.3
Baichuan-Omni-1.5	7.2	16.6	23.8	16.1

Interpretation. The correct reject option is the runner-up above chance for six of the seven under-rejecting models. Pooled over these seven models, the rate is 23.0% against the 16.7% reference, while the pooled rate on the standard twins is 15.2%, below chance. The effect is stronger on vision than on audio (25.8% against 20.2%). Baichuan-Omni-1.5 is the exception, at chance level. This agrees with the layer-wise picture in Table 17, where its signal peaks at layer 2 and decays the most on the way to the output, so the least residue survives into the final ranking. Qwen3-Omni, the strongest behavioral detector, is also the only model that places the reject option in the top two on more than half of the misleading items. We read the result as confirming a weak version of the runner-up hypothesis: a residue of the mismatch signal reaches the output logits, but the second-best option does not systematically point to the reject option.

N Limitations

Domain and stimulus length. IMAVB uses movie clips, which may not generalize to all video domains (e.g., surveillance, lectures, or user-generated content with weaker audio-visual correlation), and our misleading premises swap a single detail per question; longer or multi-step adversarial compositions may exhibit different signatures. Our clips run 60–300 s (median 140 s), which is long relative to the $\sim$10 s clips of AVHBench but not long in absolute terms. Within this range, standard accuracy declines with duration while misleading detection stays flat (Appendix E). We have no data beyond 30 minutes, where we expect standard accuracy to degrade further while detection stays flat, which would widen the gap for reasons of retrieval rather than grounding.

Text confounds and the strength of the representational claim. The automated QA generation pipeline introduces text-distributional cues that text-only classifiers detect (TF-IDF: 73.4%, SBERT: 66.8% on vision). The residualized vision probe exceeds both baselines for all eight models, but the margin over TF-IDF is only 3–7pp, and for audio the residualized probe falls below TF-IDF for several 7B models. The audio-side representational evidence is therefore weaker than the vision side. A contextual text-only classifier (e.g., a fine-tuned text encoder over the question text) is the control that would further harden this result. It is informative only if it is tuned as carefully as the probe, on the same group-by-video folds and without leakage across the four variants of a video, and we leave it to future work. Human-written misleading premises could further strengthen this control.

What the probes show. Our probes show that mismatch-related information is linearly decodable from hidden states after residualization, reliably for vision and more weakly for audio. However, they do not show that the model has localized the exact contradictory detail, and the residual could partly reflect a more generic plausibility signal. The sharper test is a counterfactual in which a non-targeted detail is altered and the probe should not move. It requires new items that pass the full generation and verification pipeline (Appendices I and B), and we leave it to future work. None of our claims depends on its outcome.

Vision–audio comparability. Audio captions come from a dedicated captioner that runs without a text prompt and describes sounds in a more abstract register than vision captions describe visible

content (Appendix I.2). Audio premises inherit this register (e.g., a music style rather than a color). We control what a question asks about through eight question types and eighteen misleading subcategories, but not how abstractly its premise is phrased. Part of the vision–audio difference may therefore reflect that an abstract audio premise is harder to verify against the signal, rather than weaker audio grounding alone. The Representation–Action Gap is measured within each modality and does not depend on this comparison.

Generation–evaluation family overlap. The captioning and QA-generation pipeline uses Qwen-family models (Qwen3-Omni-30B for audio and omnimodal captions, Qwen3.5-27B for fusion and QA generation), and our evaluation includes Qwen2.5-Omni and Qwen3-Omni. Qwen3-Omni's outlier behavior on misleading detection could in principle partly reflect family-specific familiarity with the generated question distribution rather than superior grounding. The LLM-as-Judge analysis (Appendix F) also uses a Qwen-family judge, so it is partial rather than independent evidence that these rejections are grounded. A misleading subset regenerated by a non-Qwen generator and scored by a non-Qwen judge is the direct control for this overlap.

Logit-lens scope. The vanilla logit lens is known to be unreliable for architectures that are not trained to be early-exit-friendly, and we apply a single final normalization across all layers (Appendix J). The two regimes therefore describe what this lens shows, not stable model properties, and a tuned lens [43] could show more. The representational claim rests on the probes, which do not depend on unembedding alignment.

Model coverage and PGLA. Our representational analyses (probing, logit lens, PGLA) require hidden-state access and are therefore limited to the eight open-source models. We include one proprietary model (Gemini 3.1 Pro) for behavioral evaluation, but API cost constraints and the lack of hidden-state access preclude extending the full diagnostic pipeline to additional closed-source models. PGLA is trained and tuned on splits of the same benchmark distribution; no external transfer test is reported, and out-of-distribution generalization remains an open question. The intervention trades standard accuracy for misleading detection on seven of the eight open-source models (+9.9pp at a 3pp standard-accuracy budget, +15.0pp unconstrained) and should be interpreted as diagnostic evidence for the Representation–Action Gap rather than a deployment-ready solution.

O Broader Impact

IMAVB is designed to diagnose a safety-relevant failure mode in omnimodal LLMs: the systematic inability to detect when textual premises contradict sensory evidence. Identifying this vulnerability enables developers to build more trustworthy multimodal systems, particularly for applications where adversarial or erroneous text inputs could mislead model behavior (e.g., content moderation, accessibility tools, autonomous agents). A potential negative consequence is that the detailed characterization of the Representation–Action Gap could inform adversarial prompt crafting strategies that exploit models' tendency to trust text over perception. We believe the diagnostic value outweighs this risk, as the failure mode is already exploitable without our analysis. The benchmark uses movie clips from publicly available sources; these may contain violence or mature content typical of commercial cinema, and users should be aware of this when working with the dataset. The video content is sourced from publicly available movie scenes and is distributed solely for non-commercial research purposes under fair use. We open-source all annotations, benchmark code, and evaluation tools for non-commercial research use at `https://huggingface.co/datasets/ngqtrung/IMAVB`. The video content remains the intellectual property of the original copyright holders, and we will remove any clip upon request from rights holders.

P Qualitative Examples

To complement the in-text Figure 3, we present two more annotated samples from IMAVB illustrating further types of misleading premises the benchmark employs. Red boxes in the filmstrips highlight the visual region targeted by the misleading premise. Red text in the questions marks the fabricated or swapped details.



Audio cue: synthesised electronic music reminiscent of late-1980s video games with "bip bip" bursts; then a processed voice announces "And now, Danny boy! Let's talk about safety in the workplace," followed by "oops."

	Standard premise	**Misleading premise**
Vision	Q^{v}_{std} *(A):* When the alien-like figure with reddish-brown skin appears on the monitor wearing a bright yellow hard hat, how are its eyes described? **A. large and expressive** B. small and black C. glowing green D. hidden behind sunglasses E. The visual detail in the question is incorrect. F. The audio detail in the question is incorrect.	Q^{v}_{mis} *(E):* When the alien-like figure with reddish-brown skin appears on the monitor wearing a bright blue hard hat, how are its eyes described? A. large and expressive B. small and black C. glowing green D. hidden behind sunglasses **E. The visual detail in the question is incorrect.** F. The audio detail in the question is incorrect.
Audio	Q^{a}_{std} *(A):* While the synthesised electronic music loop reminiscent of late-1980s video games plays in the background, what specific phrase does the processed voice announce? **A. And now, Danny boy! Let's talk about safety in the workplace.** B. Welcome to the factory floor, everyone stay safe. C. Danger ahead, please evacuate immediately. D. The system is ready for inspection. E. The visual detail in the question is incorrect. F. The audio detail in the question is incorrect.	Q^{a}_{mis} *(F):* While the synthesised electronic music loop reminiscent of 1950s jazz club plays in the background, what specific phrase does the processed voice announce? A. And now, Danny boy! Let's talk about safety in the workplace. B. Welcome to the factory floor, everyone stay safe. C. Danger ahead, please evacuate immediately. D. The system is ready for inspection. E. The visual detail in the question is incorrect. **F. The audio detail in the question is incorrect.**

Figure 8: **Example 1.** Frames 1–5: a man in a dark industrial setting approaches machinery. Frames 6–10: a creature on a monitor wears a yellow hard hat (red boxes). Visual swap: hard-hat colour (yellow → blue). Audio swap: music style (late-1980s video games → 1950s jazz club).



Audio cue: a woman tells the man softly to stop eating, then the man murmurs "I know, I know" in a soft, resigned tone.

	Standard premise	**Misleading premise**
Vision	Q^{v}_{std} *(A):* When Billy responds to Heidi's remark while wearing a brown suit jacket, what color is his shirt underneath? **A. white** B. black C. gray D. blue E. The visual detail in the question is incorrect. F. The audio detail in the question is incorrect.	Q^{v}_{mis} *(E):* When Billy responds to Heidi's remark while wearing a blue suit jacket, what color is his shirt underneath? A. white B. black C. gray D. blue **E. The visual detail in the question is incorrect.** F. The audio detail in the question is incorrect.
Audio	Q^{a}_{std} *(A):* After Heidi tells Billy to stop eating like that, how does Billy murmur 'I know, I know'? **A. soft and resigned** B. loud and angry C. whispering fiercely D. shouting in panic E. The visual detail in the question is incorrect. F. The audio detail in the question is incorrect.	Q^{a}_{mis} *(F):* After Heidi tells Billy to stop eating like that in a loud and angry tone, how does Billy murmur 'I know, I know'? A. soft and resigned B. loud and angry C. whispering fiercely D. shouting in panic E. The visual detail in the question is incorrect. **F. The audio detail in the question is incorrect.**

Figure 9: **Example 2.** Frames 1–5: a car drives at night. Frames 6–10: interior shot of a man and a woman; red boxes on frames 8–9 mark the brown suit jacket, a small-region detail that makes even the standard variant challenging. Visual swap: jacket colour (brown → blue). Audio insertion: a fabricated "loud and angry tone" is added to the woman's speech, though she actually speaks softly.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 state three contributions—documenting the *Representation–Action Gap*, proposing the probe-guided logit adjustment (PGLA), and releasing the IMAVB harness—each supported by results in §4–§5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix N.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is empirical; no theorems or proofs are presented.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §3 and Appendices J and I document evaluation protocol, per-model input configuration, and prompt templates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The IMAVB dataset (videos and annotations) is publicly released at `https://huggingface.co/datasets/ngqtrung/IMAVB`. All code (including hidden-state extraction, probe training, and PGLA) is publicly released at `https://github.com/ngquangtrung57/imavb`, and IMAVB is evaluated with the publicly available lmms-eval framework. Prompt templates and implementation details are fully described in Appendices I–J.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Models, evaluation protocol, and metrics are in §3; PGLA training and hyperparameters are in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: 95% bootstrap CIs are reported for shuffled baselines (Table 1); probing and PGLA use cross-validation with paired bootstrap tests where applicable (§4.3, §5, Appendix L, Appendix G).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments run on 8× NVIDIA H100 80 GB GPUs (Appendix J).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. The authors carried out all manual verification themselves, no other people took part in the study, and no personal data were collected (Appendix B).

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: See Appendix O.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: IMAVB is a diagnostic evaluation benchmark, not a generative model or pre-trained system. The benchmark does not enable generation of harmful content. Movie clips are from publicly available YouTube channels and a research dataset.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All evaluated models, video sources (BingeSociety, BoxofficeMovieScenes, Condensed Movies), and the lmms-eval framework are cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: IMAVB construction, quality verification, and statistics are documented in §2 and Appendix B, and the dataset is released on HuggingFace (`https://huggingface.co/datasets/ngqtrung/IMAVB`).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and no research with human subjects. The authors carried out all manual verification themselves (Appendix B).

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper involves no study participants other than the authors.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are used for IMAVB construction (caption generation and QA generation; §2.2, Appendix I) and as the LLM-as-Judge evaluator (Appendix F).

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.