SIGHT OR SOUND: DO MULTIMODAL LARGE LANGUAGE MODELS HAVE BLIND FAITH IN VISION FOR AUDIO-VISUAL UNDERSTANDING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-visual understanding requires Multimodal Large Language Models (MLLMs) to reason over what is seen and heard, yet aligned benchmarks rarely reveal which stream controls an answer when the evidence conflicts. We construct the Audio-Visual Incongruity Corpus (AVIC), a question-conditioned probe that preserves questions while pairing answer-disagreeing streams and constructing source-attribution options. Across four MLLMs, outputs favor visual-source answers even under audio-focused instructions; conflict-aware prompting increases abstention but leaves substantial visual preference. An option-free evaluation filtered by isolated-source alignment on an independent dataset reproduces the visual preference. On a competence-controlled two-model subset, probes find locally decodable audio information under cross-modal conflict and prompt-dependent decision transfer; paired patching restores conflict-induced answer margins through late decision states. Modality masking confirms that auditory evidence can guide answers when visual access is blocked. Finally, an exploratory attention-intervention study on AVHBench tests early audio boosting and late visual suppression, revealing task-dependent effects and larger aggregate improvements from suppressing high-attention rather than low-attention visual tokens. The results connect a gap between local auditory information and decision influence to attention targets for improving audio-visual reasoning.

1 INTRODUCTION

Audio-visual understanding (AVU) Ye et al. (2025); Wei et al. (2022) requires models to jointly interpret visual events and their associated sounds Wei et al. (2022); Senocak et al. (2023); Li et al. (2022). Unlike visual-language Bai et al. (2025); Kuang et al. (2025); Ghandi et al. (2023) or audio-language Dinkel et al. (2025); Li et al. (2025); Liu et al. (2024) tasks, AVU requires not only recognizing objects and acoustic events, but also deciding how much each modality should contribute to the final answer. This modality arbitration problem is central to real-world perception because audio and vision can provide complementary, redundant, or conflicting evidence.

Recent MLLMs such as VideoLLaMA Zhang et al. (2023) and Qwen2.5-Omni Xu et al. (2025a) combine modality-specific encoders with language models to extend audio-visual reasoning to open-ended instructions. Ideally, they should weight each stream according to task relevance and signal reliability. For an audio-focused question, this means retaining the ability to follow clear auditory evidence even when a salient visual cue suggests a different answer.

However, performance on naturally synchronized clips can obscure whether such arbitration actually occurs Seth et al. (2026). When a visible object and its sound agree, a correct answer does not identify which stream the model used. In the introductory comparison in fig. 1, AV and visual-only accuracy are close, and both example responses identify the visible chicken. Such agreement leaves the contribution of audio unresolved: either stream may supply redundant evidence, or the model may favor visual cues. AudioSet Gemmeke et al. (2017) and VGGSound Chen et al. (2020) contain many congruent pairs. A complementary test must therefore preserve the question while making the two source answers disagree.

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

Conflict also separates two questions that aligned accuracy conflates: can a model recognize the auditory evidence, and does it use that evidence when competing visual cues are present? This distinction motivates examining both isolated-modality competence and conflict-time decisions, rather than inferring evidence use from accuracy alone.

We construct the Audio-Visual Incongruity Corpus (AVIC) by pairing video and audio from different samples with matching questions but different source answers. Preserving the question makes source preference measurable under controlled conflict.

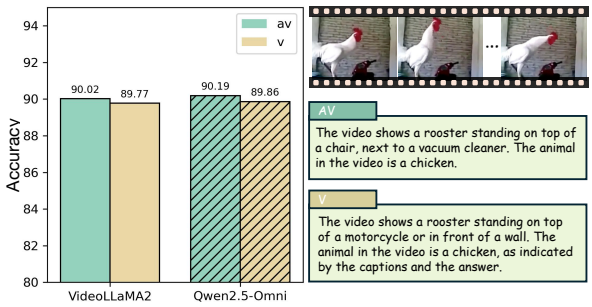

Figure 1: Audio-visual (AV) versus visual-only (V) inference: two-model accuracy (left) and a generated-response example (right).

Using this probe, we find that existing MLLMs favor visual evidence even under audio-focused instructions. An option-free replication on an independent dataset reproduces the direction. This behavioral bias raises a mechanistic question: does task-relevant audio information remain locally available, and if so, where does its influence diminish before generation? Source-disjoint probes reveal locally decodable audio information and decision transfer that varies with prompt and architecture. Paired state restoration recovers conflict-induced answer margins at late decision states. We then test whether adjusting modality attention can improve audio-visual task performance without retraining, using an exploratory AVHBench study. Our contributions are summarized as follows:

1. **Auditable conflict probe**: AVIC preserves source questions and fixes each re-matched record's options across prompt conditions; released metadata and code make the evaluation protocol inspectable.

2. **Behavioral evidence**: Four models favor visual-source answers on AVIC; an independent option-free evaluation reproduces the direction after isolated-source alignment filtering.

3. **Localization and intervention**: Competence-controlled probes reveal prompt-dependent decision transfer, and paired interventions identify late decision sensitivity in two models. Attention interventions test its practical implications on AVHBench, linking targeted visual suppression to improved aggregate accuracy and revealing task-specific trade-offs.

2 Related Work

2.1 MLLMs for Audio-Visual Understanding

MLLMs such as Qwen2.5-Omni, VideoLLaMA2, Video-SALMONN, and VITA combine audio-visual encoders with language models for joint understanding Xu et al. (2025a); Cheng et al. (2024); Sun et al. (2024); Fu et al. (2024). Synchronized-input accuracy, however, does not reveal which stream controls an answer when evidence conflicts.

Recent benchmarks examine this issue through diagnostic tasks (DAVE Radevski et al. (2025)), modality-specific labels (VGGSounder Zverev et al. (2025)), and sound modification for visible objects (AV-ConfuseBench Ye et al. (2026)). Jia et al. (2025) study conflicting cues in sound-source localization, while MMA-Bench Chen et al. (2025) and Beyond Text-Dominance Yan et al. (2026) examine tri-modal conflicts and internal modality preference. Building on this work, AVIC preserves source questions while pairing answer-disagreeing streams, enabling explicit source attribution. We connect this behavioral measurement to competence-controlled probes, paired state restoration, and inference-time attention interventions to examine whether locally available audio information influences the shared decision state.

2.2 Mitigating Modality Bias in MLLMs

Modality bias and hallucination are persistent challenges in MLLM reliability. Training-free decoding contrasts predictive distributions: CD Li et al. (2023) contrasts larger and smaller language models, DoLA Chuang et al. (2023) contrasts layers within a language model, and VCD Leng et al. (2024) contrasts original and perturbed visual inputs. CD targets text generation, while DoLA and VCD address factuality and visual hallucination, respectively. Recent work has begun to study cross-modal conflict. In vision-language models, Wu et al. (2025) identify attention imbalance and use modality mixing during instruction tuning to encourage more balanced attention. In audio-visual settings, methods such as Jung et al. (2025) combine modality masking with adaptive decoding. We connect conflict-based diagnosis to attention interventions: local representation and decision-state analyses motivate tests of audio boosting and visual suppression. These interventions adjust attention within a forward pass; their task-dependent effects are evaluated separately from source preference on AVIC.

3 Probing Modality Preference via AV Incongruities

3.1 Motivation: From Redundancy to Decision Boundaries

Redundant cues allow correct answers without revealing modality arbitration. Meaningful cross-modal conflicts can expose this preference, but require question-compatible evidence. AVE Tian et al. (2018) has limited reasoning depth, VGGSound can contain strong modality asymmetries, and Music-AVQA Li et al. (2022) focuses on a narrow domain. We build on AVQA Yang et al. (2022) for its diverse questions and temporally aligned samples.

3.2 Construction of AVIC

After verifying AVQA link availability, we retain 16,298 samples. AVIC keeps the original question and video from one record and pairs them with audio from a different record whose question matches after case normalization but whose annotated answer differs (fig. 2). For instance, an otter video can be paired with dog-barking audio under the original question "What animal is in the video?" The two source labels give distinct candidate answers for the same question.

We restrict the source pool to the 16,213 AVQA samples annotated as requiring both modalities, excluding 61 sound-only and 24 view-only samples. Of 380 case-normalized question strings, 105 admit within-question re-matching after singleton question types are pruned. We accept a pair only when its source answers differ ($y_i \neq y_j$); this establishes a conflict between annotated source labels. Isolated-stream accuracy and a stricter competence-controlled subset are evaluated separately below.

Formally, let an original sample be $S_i = (V_i, A_i, Q_i, y_i)$, where V_i and A_i are aligned visual and audio streams, Q_i is the question, and y_i is the answer. An incongruent sample $S'_{i,j} = (V_i, A_j, Q_i)$ pairs different source clips with identical case-normalized questions and different source labels ($y_i \neq y_j$). These conditions establish question compatibility and annotated answer disagreement. The released media pipeline replaces the video's audio track and truncates the combined clip to the shorter stream. Each re-matched record has four constructed options: the two source answers, "Cannot be determined," and a distractor. These constructed options are fixed across conflict-prompt conditions. The released corpus contains 15,926 records. As shown in fig. 2 (right), AVIC covers diverse question types, including animal identification and sound-source reasoning.

3.3 Per-example Evaluation Protocol

All AVIC conflict prompts receive both re-matched streams. "Audio-focused" changes the instruction, whereas "audio-only" removes video; the analogous distinction holds for video. Each record retains its original question Q_i and fixes its four constructed options across conflict-prompt conditions. The video-focused template is `{question} Please choose the answer from the following options: {multi_choice}`. The audio-focused variant replaces only `video`/`Video` with `audio`/`Audio`, preserving each record's question content. Two questions contain neither string and remain unchanged. CAP retains the original question and options and jointly adds audio priority and conflict-triggered abstention (templates in section A).

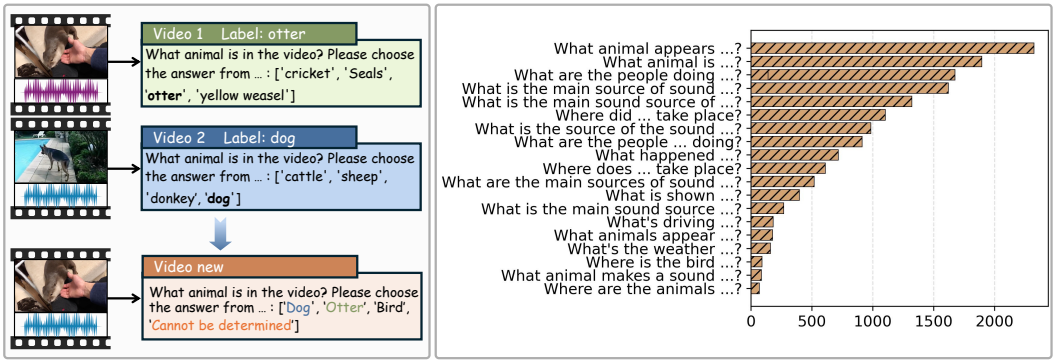

Figure 2: Overview of AVIC. (Left) An illustrative example of the Question-conditioned Re-matching pipeline, where a visual stream of an otter is paired with a contradictory dog barking track to create a cross-modal incongruity. (Right) Statistical distribution of the top 20 question types in AVIC, showing the diversity of queries.

Table 1: Quantitative results on AVIC. "In the Video" and "In the Audio" denote visual- and auditory-focused prompts. CR measures conflict abstention, VD/AD measure alignment with visual/auditory answers, and Uni-ACC. reports unimodal accuracy.

Model	**Size**	**In the Video**			**In the Audio**			**Uni-ACC.**	
		CR	VD	AD	CR	VD	AD	Vision	Audio
Qwen2.5-Omni	7B	8.94	68.67	18.27	7.29	55.90	31.81	84.16	73.06
VideoLLaMA2	7B	0.70	84.11	10.25	2.40	77.99	14.03	88.74	71.61
Video-SALMONN2^{+}	7B	11.02	83.18	2.41	15.24	77.96	2.86	81.95	34.73
Qwen3-Omni	30B	2.63	76.35	18.18	2.34	61.99	32.24	88.76	78.41

3.4 Measuring Modality Preference

AVIC has visual- and audio-source answers ($y_v \neq y_a$) for each re-matched pair. We evaluate Qwen2.5-Omni (7B), VideoLLaMA2 (7B), Video-SALMONN2-Plus (7B) Tang et al. (2025), and Qwen3-Omni (30B). We define Visual Dominance (VD) and Auditory Dominance (AD) as:

$$\text{VD} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_{v,i}), \quad \text{AD} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_{a,i}) \tag{1}$$

where $\hat{y}_i$ is the model prediction. VD and AD measure alignment with annotated sources rather than universal accuracy. The semantic evaluator maps a nonempty response to the highest-similarity option among all four using NV-Embed-v2 Lee et al. (2025); hard string matching is a separate released scoring mode. Distractor or unmatched outputs contribute to neither VD nor AD. Conflict Robustness (CR) is the rate of selecting "Cannot be determined." The three rates need not sum to 100%. Under CAP, CR reflects compliance with its abstention instruction, whereas VD and AD describe remaining source preference; unconditional abstention would also yield high CR on this all-conflict corpus.

4 Behavior under Conflict

4.1 Source Preference on AVIC

We first test whether MLLMs can follow the modality specified by the prompt when audio and vision disagree. As summarized in table 1, AVIC exposes a consistent mismatch between instruction and evidence use. Audio-focused wording shifts source preference but does not reverse the aggregate ordering: even under the "In the Audio" prompt, Qwen2.5-Omni and VideoLLaMA2 still obtain high VD scores of 55.90% and 77.99%, respectively, while AD rises to 31.81% and 14.03%. This bias is not explained by poor audio perception alone: in unimodal settings, models show clear auditory competence, such as 73.06% audio accuracy for Qwen2.5-Omni, but AD drops sharply

Table 2: Paired comparison between Base and Conflict-Aware Prompting (CAP) on the same 15,926 examples, original questions, and answer options. CAP adds only the audio-priority and conflict-reporting instruction. CR is the rate of selecting "Cannot be determined," while VD and AD measure alignment with the visual- and audio-source answers.

Model	Size	Setting	CR	VD	AD
Qwen2.5-Omni	7B	Base	8.94	68.67	18.27
		CAP	14.30 +5.36	63.24 -5.43	19.48 +1.21
VideoLLaMA2	7B	Base	0.70	84.11	10.25
		CAP	6.86 +6.16	74.44 -9.67	12.60 +2.35
Video-SALMONN2$^+$	7B	Base	11.02	83.18	2.41
		CAP	13.22 +2.20	82.00 -1.18	1.83 -0.58
Qwen3-Omni	30B	Base	2.63	76.35	18.18
		CAP	4.71 +2.08	67.19 -9.16	25.29 +7.11

Table 3: Option-free generation on Valor32k-AVQA v2.0 after relative source-alignment filtering. VD/AD (%) use nearest-answer-bank scoring; Tie denotes equal bank similarity. Intervals are pair-clustered 95% intervals over 279 source pairs.

Prompt	AD	VD	Tie	VD−AD	95% CI
Audio-focused	36.73	63.27	0.00	+26.54	[16.22, 36.70]
Video-focused	26.54	73.46	0.00	+46.92	[37.04, 56.40]

when conflicting visual evidence is present. Larger capacity also does not remove this preference, as Qwen3-Omni (30B) still reaches 61.99% VD under auditory-focused instructions. This aggregate contrast motivates the competence-controlled tests below, which examine evidence use on examples that both models answer correctly from each isolated modality.

4.2 Prompted Conflict Awareness

The low CR scores in table 1 may reflect standard prompting behavior rather than a complete inability to detect conflict. We therefore introduce Conflict-Aware Prompting (CAP), which explicitly asks the model to answer "Cannot be determined" when audio and vision are irreconcilable. Base and CAP are paired on the same 15,926 examples, original questions, and answer options; the wrapper jointly changes audio priority and conflict handling, so their individual effects are not isolated. As shown in table 2, CAP consistently reduces VD and improves CR; for example, VideoLLaMA2 drops 9.67 percentage points in VD and rises 6.16 points in CR.

CAP's effect varies by model. CR rises more than AD in the first three models, whereas Qwen3-Omni gains 7.11 points in AD and 2.08 in CR. Thus reduced VD can become either abstention or audio-source alignment. CR reflects the requested conflict handling, but unconditional abstention would also score highly on this all-conflict corpus. We report all three rates jointly: prompting changes behavior while leaving substantial visual preference, motivating the internal analysis below.

4.3 Option-free Replication on an Independent Dataset

To test whether visual preference depends on AVIC's multiple-choice measurement, we construct question-matched, answer-disagreeing inputs from the Valor32k-AVQA v2.0 Riahi et al. (2025) test set and ask Qwen3-Omni-30B-A3B to answer without options. Selection uses only isolated-modality outputs: each must be more similar to its own source-answer bank than to the competing bank under MiniLM matching. The resulting 373 conflict compositions come from 279 source pairs. Under the audio-focused conflict prompt, 63.27% are closer to the visual bank and 36.73% to the audio bank (table 3); a pair-clustered bootstrap gives a 95% interval of [16.22, 36.70] percentage points for VD−AD. This reproduces the direction for option-free generation under relative source-alignment filtering. A scoring variant allowing ambiguous assignments also retains VD>AD (section C).

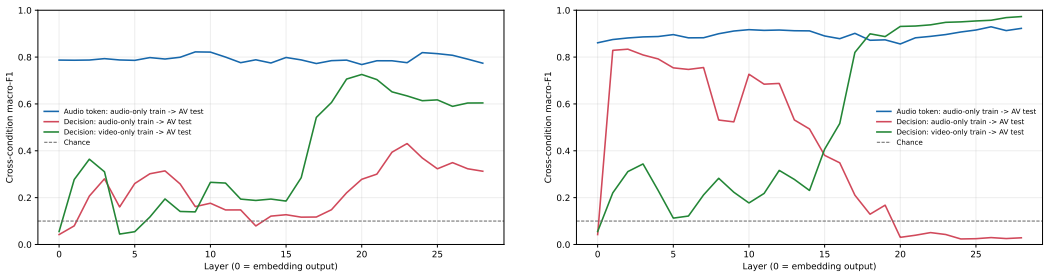

Figure 3: Video-focused unimodal-to-conflict transfer probes for Qwen2.5-Omni (left) and VideoLLaMA2 (right). Audio-focused controls are reported in section D.

5 Where Does Auditory Evidence Lose Influence?

5.1 Competence-controlled Representation and Decision Probes

To separate recognition failure from evidence use under conflict, we analyze 300 animal-question conflicts from 799 candidates. Qwen2.5-Omni and VideoLLaMA2 both answer correctly in all four isolated-modality runs; conflict outputs are not used for selection. This conditional mechanism subset controls recognition failure rather than estimating general accuracy. Source-disjoint probes test audio information at local token and decision states (section D).

Under the video-focused probe prompt, final-layer audio-token probes reach 0.809/0.924 macro-F1 for Qwen/VideoLLaMA2. Decision probes favor visual over audio labels (0.917/0.744 and 0.979/0.524; table 9). Audio-only/AV-audio CKA and RSA remain high (0.864/0.904 and 0.966/0.961), although Qwen's decision-state similarity gap is small and inconsistent. Prompt dependence is examined below.

5.2 Separating Local Decodability from Decision Transfer

To examine how locally available information relates to decision representations, we complement direct probes with unimodal-to-conflict transfer. Direct probes measure linear recoverability of source labels within conflict states. Transfer probes freeze a classifier trained on isolated-modality states and evaluate the same readout under conflict, testing whether source information remains accessible through a shared readout.

Under the video-focused prompt (fig. 3), Qwen's audio-token transfer reaches 0.774 macro-F1, versus 0.313 for audio decision transfer and 0.604 for visual decision transfer. VideoLLaMA2 yields 0.923, 0.029, and 0.973. We then change the requested modality to test how this pattern depends on the instruction.

Matched audio-focused control. Changing only the requested modality on the same 300 examples raises Qwen's audio decision transfer to 0.709 and lowers visual transfer to 0.180. VideoLLaMA2 retains a visual advantage (0.882 versus 0.130 audio transfer). Audio-token transfer remains 0.774/0.923. Thus local audio availability is stable across these prompts, while decision transfer depends on instruction and architecture (full results in section D).

Selection and controls. Of 799 candidates, 569 pass the all-four-correct filter; the final 300 have unique audio/video sources and ten classes, split into five approximately stratified folds of 60. Both prompts share ten ordered animal labels, unlike AVIC's four options. L2-regularized multinomial logistic probes select regularization within training folds and report macro-F1. Additional protocol details and opposite-source controls appear in section D.

5.3 Paired Restoration of Decision States

The probe results motivate testing which states influence the answer margin when visual evidence changes. Activation patching uses a separate audio-focused, two-option forced-choice diagnostic on the same 300 source-paired examples. Within each pair, the congruent and conflict runs share

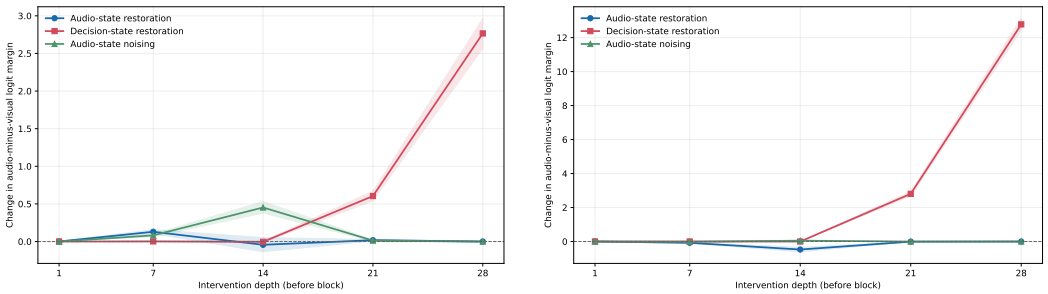

Figure 4: Paired activation patching for Qwen2.5-Omni (left) and VideoLLaMA2 (right), immediately before blocks 1, 7, 14, 21, and 28 (one-based).

audio, prompt, answer mapping, and token layout; only the video changes. We restore states from the congruent run into the conflicting run and measure the answer-logit margin $m = z_a - z_v$. For a patched state group at layer l, normalized recovery is

$$R_l = \frac{\mathbb{E}[m_{\text{patched},l} - m_{\text{conflict}}]}{\mathbb{E}[m_{\text{congruent}} - m_{\text{conflict}}]}. \tag{2}$$

The mean congruent-minus-conflict margin is 3.278 logits for Qwen2.5-Omni (95% CI [3.053, 3.501]) and 14.854 for VideoLLaMA2 ([14.441, 15.237]). Decision-state restoration reaches 18.5%/18.9% before block 21 and 84.4%/86.0% before block 28, whereas early audio-state restoration is small or absent. This depth-dependent recovery identifies late decision states as sensitive sites for restoring the conflict-induced margin change.

Self-patch and noising controls, together with interpretation of the recovery metric, appear in section D.

Patching prompt behavior. Qwen chooses audio on 299/300 congruent inputs and 266/300 conflicts; VideoLLaMA2 does so on 300/300 and 21/300. Thus Qwen mostly retains audio answers despite a reduced margin. Across models, conflict changes the answer margin and late restoration recovers it, while the frequency of audio answers differs under this two-option protocol.

5.4 Attention as a Descriptive Routing Diagnostic

To relate late decision sensitivity to modality allocation across layers, we examine attention during prefill and decoding. Figure 5 presents layer-wise allocation alongside global and per-modal top-K statistics for $K = 20, 50, 100$. During prefill, audio allocation drops in early layers; among audio and video, visual tokens receive more attention in later layers and in the global top-100 set. Text receives the largest share in the top-100 plot. The prefill and decode views distinguish attention while processing the input from attention during answer generation, placing the modality comparison in its computational context. Normalization and additional statistics appear in section E.

The aggregate and sparse views characterize different aspects of this allocation. Per-token mass divides each modality's received attention by its token count, summarizing average allocation at the token level. Global top-K statistics describe which modalities occupy the most attended positions across the sequence, while per-modal top-K statistics examine highly attended positions within each modality. Together, these views distinguish average allocation from concentration on selected positions. Combined with local audio decodability and late decision sensitivity, the observed patterns motivate testing early audio boosting and attention-ranked visual suppression in later layers. We first examine how removing access to an entire modality changes source preference, then evaluate graded interventions that retain both input streams on downstream tasks.

5.5 Modality-access Control

On full AVIC, we mask attention to all positions of one modality across layers, retaining the video-focused question and four options. This access control differs from the graded,

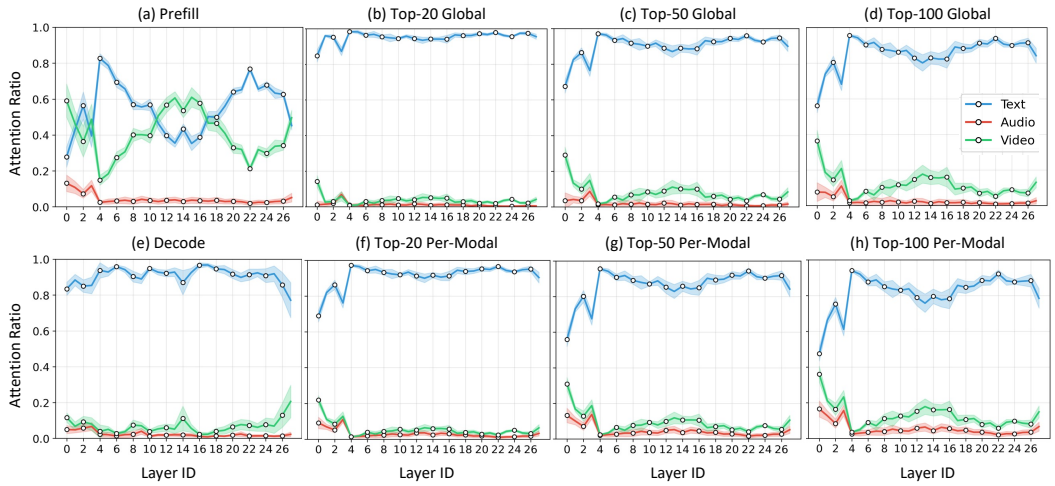

Figure 5: Full attention diagnostics: prefill/decode allocation (a,e), global top-K (b–d), and per-modal top-K (f–h), for $K = 20, 50, 100$. Text has the largest global share; among audio and video, visual allocation is generally higher.

layer-restricted AVHBench intervention below. Removing access to video raises AD from 18.27% to 72.71% for Qwen2.5-Omni and from 10.25% to 76.37% for VideoLLaMA2 (table 4). Masking audio instead leaves outputs mostly visual. The shift demonstrates output sensitivity to modality access and recoverable auditory influence; an illustrative case appears in fig. 10.

Table 4: AVIC modality-access control (%). Mask V/A blocks visual/audio attention; Base is standard inference.

Model	Setting	CR	VD	AD
Qwen2.5-Omni	Base	8.94	68.67	18.27
	Mask V	11.94	7.28	72.71
	Mask A	16.44	76.94	2.80
VideoLLaMA2	Base	0.70	84.11	10.25
	Mask V	2.41	11.81	76.37
	Mask A	1.33	90.01	4.25

6 From Diagnosis to Attention Intervention

We test these diagnosis-inspired interventions on AVHBench. State and attention analyses motivate the intervention directions; layer ranges and strengths are evaluated through sensitivity sweeps. Graded adjustments retain evidence lost through whole-modality masking, and task accuracy assesses effects beyond AVIC source attribution.

Intervention design. With weights frozen, audio boosting adds $\log \alpha$ to audio-key logits before softmax in layers 0–8, for $\alpha \in \{1.5, 2, 3\}$. Visual suppression masks $r \in \{25\%, 50\%, 75\%\}$ of visual key positions in layers 18–26 and renormalizes attention. Positions are ranked at the first active layer; selected indices are reused across active layers and decoding steps. Low-attention masking controls target selection. Both interventions operate on the current input without a congruent reference run.

Method comparison. table 5 compares baseline inference, VCD, AVCD, and combined audio boosting and visual suppression on AVHBench. The combined intervention reaches 74.93% for Qwen2.5-Omni and 72.50% for VideoLLaMA2; component effects are evaluated separately below (section G.1).

Component ablation. We separately sweep intervention strengths on 5,302 AVHBench examples with VideoLLaMA2: 2,290 V→A, 1,136 A→V, and 1,876 AV Match cases. Overall is their sample-weighted micro-average. table 6 uses the component-evaluation baseline of 70.65%. Audio boosting improves aggregate accuracy monotonically within the first-nine-layer grid; high-attention visual suppression peaks at 50% among the three displayed ratios. At that ratio, the low-attention control reaches 70.67%, compared with 71.65% for high-attention masking. AV Match improves most, while A→V declines, revealing a task-dependent trade-off. Expanded layer sweeps and the evaluation protocol appear in section G.

Table 5: Reported AVHBench method comparison (%). Result sources and component evaluation details are specified in section G.1.

Method	Qwen2.5-Omni	VideoLLaMA2
Baseline	74.33	70.65
VCD Leng et al. (2024)	73.07	65.18
AVCD Jung et al. (2025)	74.67	72.15
Combined intervention	74.93	72.50

Table 6: Component ablation on AVHBench with VideoLLaMA2 (accuracy, %). Audio boost uses layers 0–8; visual masking uses layers 18–26 (zero-based). Each component is applied separately.

Setting	Overall	V→A	A→V	AV Match
Baseline	70.65	77.60	80.11	56.45
Audio boost $\times 1.5$	70.82	77.73	80.02	56.82
Audio boost $\times 2$	71.12	77.60	79.75	58.00
Audio boost $\times 3$	71.24	77.34	79.49	58.80
Visual mask (high 25%)	71.31	77.86	80.19	57.94
Visual mask (high 50%)	71.65	78.12	79.58	58.96
Visual mask (high 75%)	71.44	77.90	79.40	58.74
Visual mask (low 50%)	70.67	77.38	79.75	56.98

7 Discussion and Limitations

Upstream causes and intervention scope. The architectures do not form matched controls for training mixture, training order, or token count. Qwen3-Omni reports 0.77T audio and 0.82T image tokens in stage-S2 pretraining Xu et al. (2025b), yet remains visually biased on AVIC; VideoLLaMA2 uses more audio than video tokens on the 300-example mechanism subset. These observations rule out a simple universal prediction from raw token totals, not training-related explanations.

The matched prompt results also suggest a concrete evaluation principle: modality preference should be measured across instructions on the same inputs. Stable local audio decodability can coexist with substantial changes in decision transfer. Reporting both measurements helps distinguish a model that adapts its representations to the requested evidence from one that retains a visual preference across prompts.

Evaluation scope. AVIC inherits AVQA's question distribution and answer space. The external replication uses automatic relative-alignment filtering; internal analyses cover two models and 300 animal-question examples. The AVHBench sweep measures intervention sensitivity, with the tuning protocol detailed in Appendix G.2; the reported intervention gains are point estimates without statistical significance tests.

8 Conclusion

Question-matched conflicts reveal visual-source preference across four MLLMs, with an option-free replication under relative source-alignment filtering. Two-model probes reveal locally decodable audio information and prompt-dependent decision transfer; late state restoration recovers conflict-induced answer margins. Guided by these findings, attention adjustments yield task-dependent improvements in the AVHBench sensitivity study, with high-attention visual suppression outperforming the corresponding low-attention control. These findings highlight a central distinction for audio-visual reasoning: recognizing information within a modality and allowing that information to shape the final answer are separate capabilities. Evaluating both capabilities under matched conflicts can help distinguish perceptual failures from evidence-use failures and guide more targeted improvements to multimodal reasoning. Future work can couple modality weighting with input reliability, evaluated through paired conflicts and naturally aligned tasks.

AI USE STATEMENT

Generative AI tools were used for language polishing and improving the presentation of this manuscript. The authors take responsibility for the final content.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Tianle Chen, Chaitanya Chakka, Arjun Reddy Akula, Xavier Thomas, and Deepti Ghadiyaram. Some modalities are more equal than others: Decoding and architecting multimodal integration in mllms. *arXiv preprint arXiv:2511.22826*, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

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. *arXiv preprint arXiv:2309.03883*, 2023.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. MiDashengLM: Efficient audio understanding with general audio captions, 2025. URL https://arxiv.org/abs/2508.03983.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. *ACM Comput. Surv.*, 2023.

Yanhao Jia, Ji Xie, S. Jivaganesh, Hao Li, Xu Wu, and Mengmi Zhang. Seeing sound, hearing sight: Uncovering modality bias and conflict of ai models in sound localization. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding. *arXiv preprint arXiv:2505.20862*, 2025.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with MLLM in visual question answering: A survey. *ACM Comput. Surv.*, 2025.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.17428.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering, 2025. URL https://arxiv.org/abs/2503.11197.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2023.

Jizhong Liu, Gang Li, Junbo Zhang, Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Yujun Wang, and Bin Wang. Enhancing automated audio captioning via large language models with optimized audio encoding. In *Interspeech 2024*, 2024.

Gorjan Radevski, Teodora Popordanoska, Matthew B Blaschko, and Tinne Tuytelaars. Dave: Diagnostic benchmark for audio visual evaluation. *arXiv preprint arXiv:2503.09321*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://arxiv.org/abs/1908.10084.

Ines Riahi, Abduljalil Radman, Zixin Guo, Rachid Hedjam, and Jorma Laaksonen. Valor32k-AVQA v2.0: Open-ended audio-visual question answering dataset and benchmark. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13097–13103, 2025. doi: 10.1145/3746027.3758261. URL https://doi.org/10.1145/3746027.3758261.

Arda Senocak, Junsik Kim, Tae-Hyun Oh, Dingzeyu Li, and In So Kweon. Event-specific audio-visual fusion layers: A simple and new perspective on video understanding. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023.

Ashish Seth, Xinhao Mei, Changsheng Zhao, Varun Nagaraja, Ernie Chang, Gregory P. Meyer, Gael Le Lan, Yunyang Xiong, Vikas Chandra, Yangyang Shi, Dinesh Manocha, and Zhipeng Cai. Egoavu: Egocentric audio-visual understanding, 2026. URL https://arxiv.org/abs/2602.06139.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022.

Chen Henry Wu, Neil Kale, and Aditi Raghunathan. Mitigating modal imbalance in multimodal reasoning. *arXiv preprint arXiv:2510.02608*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, et al. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Xinru Yan, Boxi Cao, Yaojie Lu, Hongyu Lin, Weixiang Zhou, Le Sun, and Xianpei Han. Beyond text-dominance: Understanding modality preference of omni-modal large language models. *arXiv preprint arXiv:2604.16902*, 2026.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, 2022.

Qilang Ye, Zitong Yu, Rui Shao, Yawen Cui, Xiangui Kang, Xin Liu, Philip Torr, and Xiaochun Cao. Cat+: Investigating and enhancing audio-visual understanding in large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Qilang Ye, Wei Zeng, Meng Liu, Jie Zhang, Yupeng Hu, Zitong Yu, and Yu Zhou. When eyes and ears disagree: Can mllms discern audio-visual confusion? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11955–11963, 2026. doi: 10.1609/aaai.v40i14.38183. Issue 14.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, 2023.

Daniil Zverev, Thaddäus Wiedemer, Ameya Prabhu, Matthias Bethge, Wieland Brendel, and A. Sophia Koepke. Vggsounder: Audio-visual evaluations for foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Supplementary Material

A Details of Prompts

To ensure reproducibility, we provide the exact per-example prompt construction used in each evaluation setting. Here, `{question}` and `{multi_choice}` are read from the current AVIC record.

A.1 Prompt Focus and Input Modality

Prompt focus and input modality are separate settings. Both video-focused and audio-focused conflict runs contain video and audio. Video-only/audio-only runs instead supply the corresponding isolated stream. The video-focused prompt retains the record's original AVQA question. For the audio-focused condition, only `video`/`Video` in that question is replaced by `audio`/`Audio`; all other words and the per-example options remain fixed. Two of 15,926 records contain no `video` in the question, so their audio-focused question strings remain unchanged.

Video-focused Prompt

`{question}` Please choose the answer from the following options: `{multi_choice}`

Audio-focused Prompt

`{audio-focused question}` Please choose the answer from the following options: `{multi_choice}`

For example, an animal-identification record uses "What animal is in the video?" in the video-focused condition and "What animal is in the audio?" in the audio-focused condition. These sentences illustrate one of the 105 retained question forms.

A.2 CONFLICT-AWARE PROMPT (CAP)

CAP and Base both receive the same conflicting streams, original question, and constructed options. CAP changes the wrapper to request both audio priority and abstention on irreconcilable evidence; this comparison measures their joint effect, not separate effects of the two instructions.

CAP Prompt Template

`{question}` Options: `{multi_choice}`. Please choose the correct answer from these options. **Focus primarily on the audio/sound information.** If the visual information conflicts with the audio information, choose **"Cannot be determined"**. For example, if you see a dog but hear a cat meowing, the visual and audio conflict, so choose "Cannot be determined".

Implementation Logic:

- **Audio-priority instruction:** The wrapper tests whether an explicit request can increase the influence of auditory evidence when the streams disagree.
- **Conflict example:** The dog–cat example defines the intended use of "Cannot be determined" without changing the current record's question or options.

B THE AVIC DATASET

To evaluate the model's ability to detect cross-modal semantic conflicts, we construct the **Audio-Visual Incongruity Corpus (AVIC)** based on the original AVQA dataset.

B.1 DATA SELECTION AND FILTERING

The original AVQA dataset provides *Question Relation* annotations, categorizing questions based on the modality required to answer them. Among the 16,298 available samples in our source pool, the distribution is as follows:

- **Both (Audio-Visual):** 16,213 samples (99.48%)
- **Sound-only:** 61 samples (0.37%)
- **View-only:** 24 samples (0.15%)

To reduce missing-modality artifacts, we exclusively select the **Both** category (16,213 samples) as our foundation. We separately evaluate each isolated stream below to check whether the paired source answers are recoverable by the evaluated models.

B.2 QUESTION TAXONOMY AND REFINEMENT

The initial pool contained 380 distinct question types. However, a long-tail distribution was observed, where many types contained only a single sample, lacking the requisite diversity for robust evaluation and cross-matching. We performed the following refinement:

1. **Pruning:** Question types with only one sample were removed because they do not admit within-question re-matching. After pruning, we retained **105** question forms.
2. **Re-matching:** Within each of the 105 question forms, we pair a visual source with an audio source from a different AVQA record that has the same question under case-insensitive text matching but a different annotated answer. Each released record stores four re-matched options containing both source answers, "Cannot be determined," and one distractor. The original source option lists are stored separately for inspection. Evaluation reads the stored re-matched list rather than regenerating or reshuffling it.

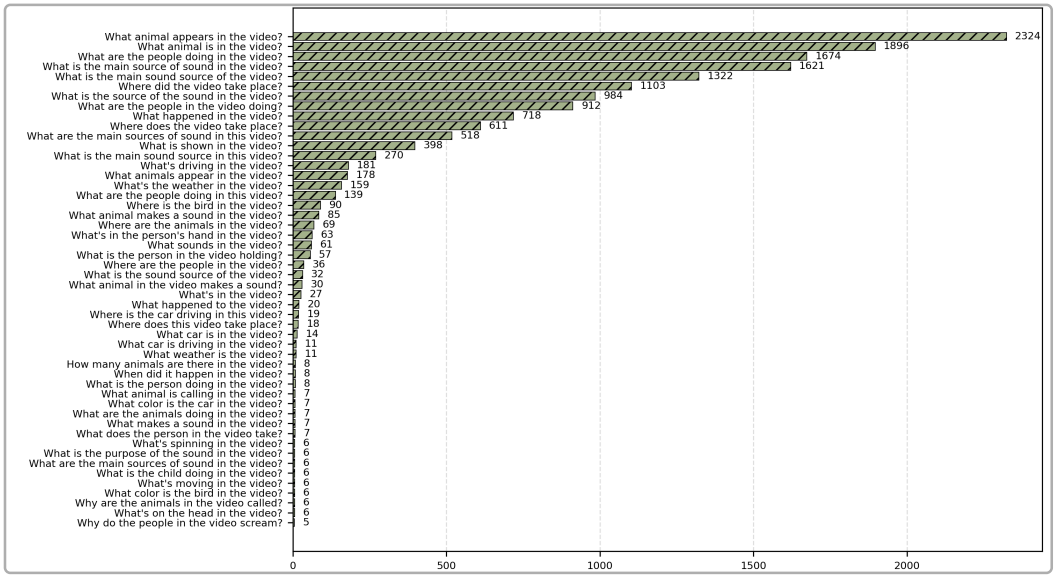

Figure 6: Distribution of the top 50 question types in AVIC after refinement.

Table 7: Unimodal source-answer accuracy and AVIC source alignment (%). V−A and VD−AD are percentage-point differences.

Model	Video-only	Audio-only	V−A	AVIC VD	AVIC AD	VD−AD
Qwen2.5-Omni	84.16	73.06	11.10	68.67	18.27	50.40
VideoLLaMA2	88.74	71.61	17.13	84.11	10.25	73.86

B.3 Dataset Statistics

After filtering and re-matching, AVIC contains **15,926** samples across 105 question forms after case normalization. fig. 6 shows the 50 most frequent questions, spanning animal identification, counting, location, events, sound sources, and other AVQA categories.

B.4 Unimodal Validity after Re-matching

Each stream is inherited from an AVQA source record annotated with the corresponding source answer. We additionally evaluate the re-matched streams in isolation. As shown in table 7, both tested models recover source answers from either isolated modality, while the visual–audio gap under conflict is substantially larger than the corresponding unimodal capability gap.

C Option-Free External Replication

We construct question-matched, answer-disagreeing conflicts from the independent Valor32k-AVQA v2.0 test set. Qwen3-Omni-30B-A3B generates answers without candidate options. For each source, its answer bank contains the reference and its supplied rephrasings. We embed the generated answer and bank entries with the Sentence Transformers model `paraphrase-MiniLM-L3-v2` Reimers & Gurevych (2019) and use the maximum cosine similarity within each bank, denoted s_a and s_v.

Selection and primary scoring. The 373 retained compositions from 279 source pairs satisfy $s_a > s_v$ for the audio-only output and $s_v > s_a$ for the video-only output. Selection does not use conflict outputs. This filter measures relative source alignment rather than establishing absolute answer correctness. For conflict outputs, the primary rule assigns audio when $s_a > s_v$, visual when $s_v > s_a$, and a tie for equal scores. Thus a vague or irrelevant response can still receive a source

Table 8: Scoring sensitivity on the same 373 option-free outputs (%). Amb. is the ambiguous or exact-tie category, depending on the rule. The ambiguity threshold is 0.06.

Prompt	Scoring rule	AD	VD	Amb.
Audio-focused	Nearest bank	36.73	63.27	0.00
	Gap below 0.06 ambiguous	32.44	54.42	13.14
Video-focused	Nearest bank	26.54	73.46	0.00
	Gap below 0.06 ambiguous	22.25	65.15	12.60

assignment; zero ties does not mean every response is unambiguous. The generator sees no options, while the evaluator uses the source references.

Scoring sensitivity. We also evaluate an ambiguity-aware scoring rule, which marks $|s_a - s_v| < 0.06$ as ambiguous. Re-aggregating those labels on the same 373 outputs yields table 8; the primary counts and confidence intervals remain those of nearest-bank scoring. Allowing ambiguity reduces both source rates while preserving their ordering for both prompts. This checks the sensitivity of automatic source attribution to the scoring rule.

Each source pair contributes one or two compositions. The intervals in table 3 use 100,000 percentile bootstrap replicates over the 279 pair clusters (seed 42), retain all compositions within each sampled cluster, and compute sample-weighted VD$-$AD. The retained set contains 200 action, 131 temporal, 36 description, and 6 count cases.

D Representation and Decision-State Controls

We analyze 300 conflicts for which Qwen2.5-Omni and VideoLLaMA2 both recover the corresponding source answer from each isolated modality. Source videos do not overlap across five approximately stratified folds; regularization is selected within each training fold; models, prompts, and option order are fixed. At every layer, the *decision state* is the final input-token state before answer generation. Direct probes are trained and tested on AV-conflict states, whereas transfer probes are trained on the corresponding unimodal states and evaluated on AV-conflict states without refitting. All probe values are macro-F1. Direct probes measure decodability rather than the model's use of that information in its answer. Reduced transfer can reflect changes in representational coordinates or feature mixing, even when source information remains present. Opposite-stream controls are near the 0.1 chance level.

Probe prompt and readout. The video-focused representation results use "What animal appears in the video? Choose exactly one answer from: bird, cat, cattle, chicken, dog, dove, gibbon, horse, lion, pheasant." This fixed ten-class list is shared across examples and differs from the full-corpus four-option evaluation. Frozen hidden states are used to train external linear probes; the reported macro-F1 is probe classification performance, not the MLLM's generated-answer accuracy.

Matched prompt control. The audio-focused prompt changes only "in the video?" to "in the audio?" in the ten-class question above. The 300 source pairs, labels, option order, and folds are identical across the two prompt conditions. At the final layer, Qwen's direct decision probes predict audio/visual labels at 0.744/0.917 under video focus and 0.786/0.831 under audio focus; its audio/visual decision-transfer scores change from 0.313/0.604 to 0.709/0.180. VideoLLaMA2's corresponding direct scores are 0.524/0.979 and 0.739/0.961, while transfer scores are 0.029/0.973 and 0.130/0.882. AV audio-token direct scores remain 0.809/0.924 for Qwen/VideoLLaMA2, and audio-token transfer remains 0.774/0.923. Prompting therefore changes decision representations more than local audio decodability in this comparison. For Qwen, audio-focused transfer favors audio even though the direct decision probe still predicts visual labels somewhat better; the diagnostics measure different readouts.

The recognition filter does not control confidence or interference susceptibility. Under the video-focused prompt, Qwen's opposite-source controls are 0.097 (visual-only decision states to paired audio labels) and 0.125 (audio-only states to paired visual labels), near ten-class chance.

Table 9: Representative source-disjoint probe and representation-similarity results. A/V denote audio-/visual-source labels. CKA/RSA entries compare audio-only audio states with AV audio states.

Model	Metric	L7	L14	L21	L28
Qwen2.5-Omni	AV audio state → A	0.798	0.822	0.817	0.809
	AV decision state → A	0.607	0.541	0.688	0.744
	AV decision state → V	0.631	0.666	0.948	0.917
	Audio-state transfer A→AV	0.792	0.775	0.784	0.774
	Decision-state transfer A→AV	0.314	0.121	0.300	0.313
	Decision-state transfer V→AV	0.194	0.194	0.704	0.604
	A/AV audio CKA/RSA	.998/.998	.982/.982	.962/.966	.864/.904
VideoLLaMA2	AV audio state → A	0.882	0.910	0.911	0.924
	AV decision state → A	0.885	0.798	0.685	0.524
	AV decision state → V	0.798	0.900	0.974	0.979
	Audio-state transfer A→AV	0.883	0.912	0.882	0.923
	Decision-state transfer A→AV	0.755	0.493	0.039	0.029
	Decision-state transfer V→AV	0.212	0.231	0.933	0.973
	A/AV audio CKA/RSA	1.000/.999	.998/.998	.994/.989	.966/.961

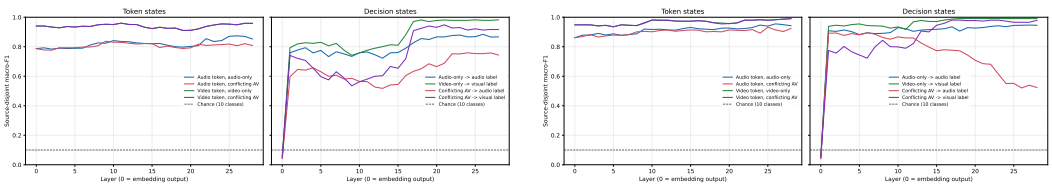

Figure 7: All-layer direct-probe results for Qwen2.5-Omni (left) and VideoLLaMA2 (right) on the competence-controlled subset.

Patching prompt. The separate patching diagnostic uses "What animal appears in the audio? Choose from A={answer A}; B={answer B}. Respond with exactly one letter." The two candidates are the paired source answers; their letter mapping is fixed within each congruent/conflict pair. Both runs contain audio and video. The reported margin compares the audio-target and visual-target letter logits before answer generation.

Indexing conventions. Probe plots index representation depth from 0 (embedding output) to 28. Patching labels L1/L7/L14/L21/L28 refer to pre-block hooks with one-based block numbers, implemented at block indices 0/6/13/20/27. In contrast, the AVHBench attention intervention specifies zero-based block ranges 0–8 and 18–26. A patch at L28 is therefore not a post-block-28 probe readout or masking layer 28.

Under the video-focused prompt, audio-source labels remain decodable at AV audio-token positions, whereas late decision probes favor visual-source labels, most strongly for VideoLLaMA2. Qwen2.5-Omni shows a weaker and non-monotonic pattern, and its CKA/RSA decision gap is small and inconsistent; its late preference is therefore supported by direct and transfer probes rather than representation similarity alone. These diagnostics characterize linear decodability and sample geometry.

For paired counterfactual activation patching, matched and conflict runs keep audio, prompt, answer mapping, token IDs, and layout fixed and change only the paired video. Normalized margin recovery is the patch-induced change in the audio-minus-visual answer-logit margin divided by the matched-minus-conflict margin within each model. As shown in table 10, final-layer decision-state patches recover most of this margin, while early audio-state recovery is small or absent. This localizes sensitivity of the answer margin to late decision states. Since the restored state can carry the reference answer preference, the result does not identify the upstream pathway that produced the conflict effect. Recovery is a signed ratio of mean margin changes, rather than an answer-flip rate or accuracy gain.

Self-patch and noising controls. Self-patches replace a state with its own cached value. Across the 300 examples and both audio/decision self-patches, Qwen's L1 mean absolute mar-

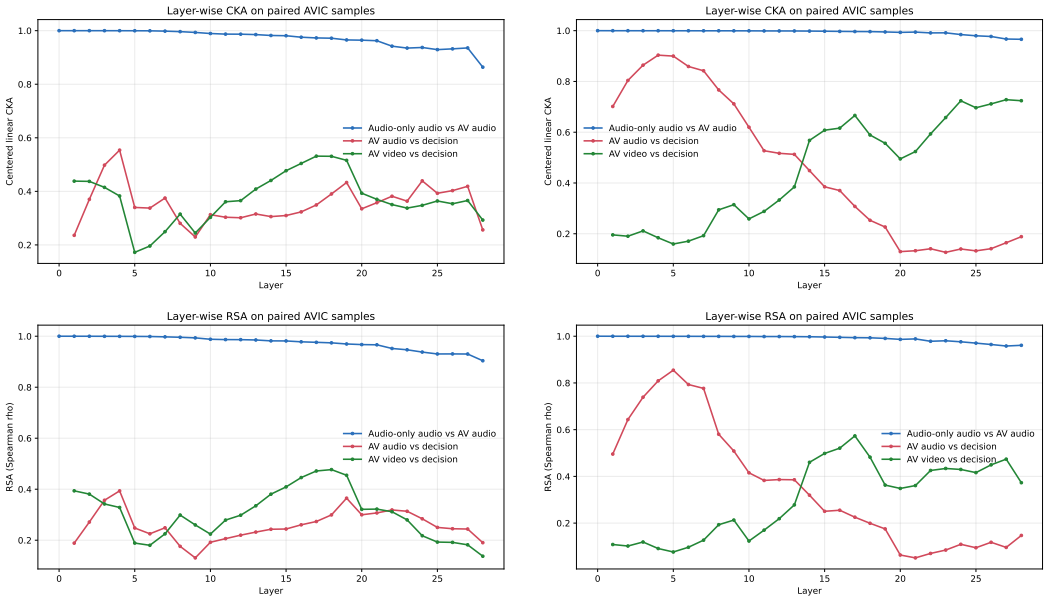

Figure 8: All-layer CKA (top) and RSA (bottom) results for Qwen2.5-Omni (left) and VideoLLaMA2 (right).

Table 10: Normalized margin recovery from paired counterfactual activation patching.

Model	L21 decision	L28 decision	Early audio state
Qwen2.5-Omni	18.5%	**84.4%**	L7: 4.0%; L14: no positive recovery
VideoLLaMA2	18.9%	**86.0%**	no positive recovery

gin change is 0.000417 logits (maximum 0.125); all other tested depths and all VideoLLaMA2 self-patches have zero recorded change. Audio noising inserts conflict audio states into the congruent run; its signed effect is $m_{\text{congruent}} - m_{\text{noised}}$. Mean effects at L1/L7/L14/L21/L28 are 0.0004/0.0819/0.4527/0.0127/-0.0010 for Qwen and 0.0000/-0.0140/0.0499/-0.0027/-0.0001 for VideoLLaMA2. These values are averaged over paired examples. Noising and restoration need not have equal effects because their receiving runs differ.

E Additional Representation and Attention Diagnostics

For modality m with token set S_m, the layer-l mean hidden-state magnitude is $E_m^{(l)} = |S_m|^{-1} \sum_{i \in S_m} \|\mathbf{h}_i^{(l)}\|_2$. The audio and video trajectories do not show a simple magnitude collapse (fig. 9); hidden-state magnitude provides a complementary diagnostic of representation scale.

For attention weights $\mathbf{A} \in \mathbb{R}^{H \times L_Q \times L_K}$, we normalize each modality's received mass by its number of tokens:

$$\bar{\mathcal{M}}_m^{(l)} = \frac{1}{H L_Q |S_m|} \sum_{h=1}^{H} \sum_{i=1}^{L_Q} \sum_{j \in S_m} \mathbf{A}_{h,i,j}^{(l)}. \tag{3}$$

During prefill, audio allocation drops sharply in early layers; during decoding, the remaining cross-modal attention mainly targets video. Text receives the largest share of global top-K attention; among the two non-text modalities, visual allocation generally exceeds audio allocation for $K \in \{20, 50, 100\}$, and per-modal top-K audio peaks remain smaller (fig. 5). These statistics describe modality allocation across layers and decoding stages; attention weights alone do not establish causal information flow.

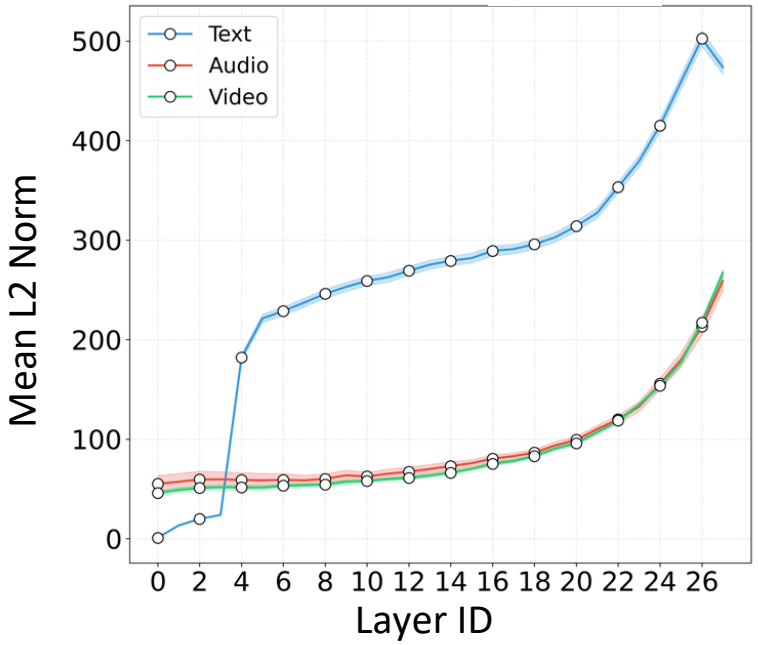

Figure 9: Layer-wise hidden-state norms for text, audio, and video tokens.

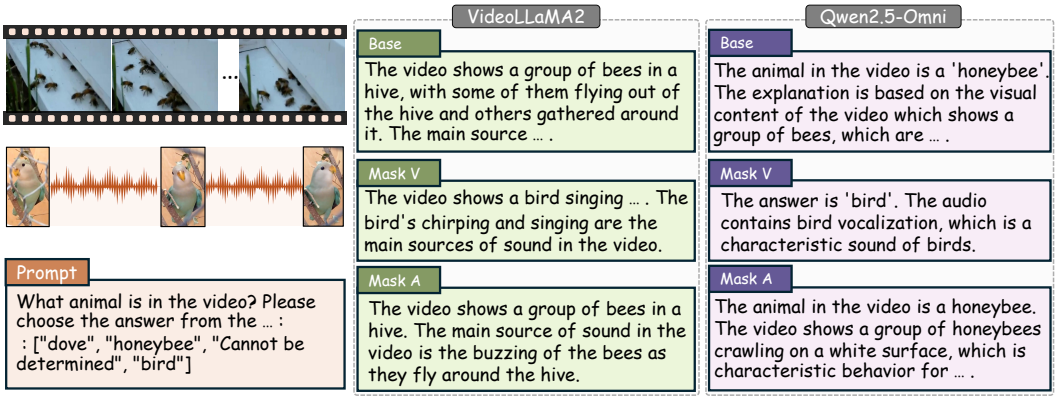

Figure 10: An animal-identification record with the original question "What animal is in the video?" and conflicting bee video and bird audio. Under standard inference the output follows the visual source; masking visual access shifts it toward the auditory source.

F Illustrative Modality-masking Case

G Additional Ablation Results

G.1 Method Comparison and Component Evaluation

table 5 reports AVHBench method-level results for two backbones. VideoLLaMA2's VCD and AVCD scores (65.18/72.15) are taken from prior work Jung et al. (2025). The combined audio-boosting and visual-suppression intervention scores are 74.93 for Qwen2.5-Omni and 72.50 for VideoLLaMA2.

table 6 evaluates audio boosting and visual masking separately on VideoLLaMA2. Relative to the 70.65 baseline, applying the 50% high-attention visual mask alone yields an accuracy of 71.65.

G.2 Expanded Component Sweeps

This section reports the expanded VideoLLaMA2 sensitivity sweep referenced in table 6. We vary the number of audio-boosted layers, the boost multiplier, and whether high- or low-attention visual keys are masked. High-attention masking produces larger directional changes than low-attention masking, while the response to audio boosting depends on both layer range and task. We report a sensitivity sweep on all 5,302 AVHBench test examples, selecting intervention settings on the same evaluation set.

Implementation and scope. Audio boosting adds $\log \alpha$ to audio-key logits before softmax. For attention-ranked visual masking, the implementation computes mean received attention over heads and query positions at the first active layer, caches the selected visual indices for the example, and reuses them across the active layers and decoding steps. It zeros the selected attention columns and renormalizes; it does not zero the key vectors themselves. The late range is layers 18–26 with zero-based indexing, excluding layer 27.

Ablation (VideoLLaMA2 on AVHBench). We sweep the two intervention families in isolation (table 6): early audio boosting over the first 9 layers with multipliers $\{1.5, 2, 3\}$, and late high-attention visual masking over the layers 18–26 (zero-based) with mask ratios $\{25\%, 50\%, 75\%\}$. All entries are measured point estimates.

Increasing the audio boost strength raises Overall monotonically within this grid ($+0.17 \rightarrow +0.47 \rightarrow +0.59$ relative to the component-evaluation baseline) and yields an upward trend on AV Match ($+0.37 \rightarrow +1.55 \rightarrow +2.35$). The other hallucination probes are not monotonic: V$\rightarrow$A drifts from 77.73 at $\times 1.5$ to 77.34 at $\times 3$, while A$\rightarrow$V falls from 80.02 to 79.49. The sensitivity is therefore task-dependent rather than a uniform performance gain.

For late high-attention visual masking, a mask ratio of 50% attains the best Overall and AV Match in the $\{25, 50, 75\%\}$ sweep. At 25%, Overall lies between the baseline and the 50% result, while A$\rightarrow$V reaches its highest value within this grid (80.19). Increasing the ratio to 75% reduces Overall and A$\rightarrow$V relative to 50%, while V$\rightarrow$A remains above baseline. Thus aggregate accuracy peaks at the middle ratio, whereas audio boosting improves monotonically over the tested multipliers. The preferred suppression strength varies across tasks.

Additional controls in table 11 show that masking *low*-attention visual keys produces smaller aggregate changes than masking high-attention visual keys. Layer choice also matters: boosting the first 6 layers at $\times 3$ reaches 71.48%, exceeding the corresponding first-9-layer result of 71.24%; first-13-layer boosts range from 70.60% to 71.01%. These patterns show sensitivity to both the layer range of audio boosting and the attention rank of suppressed visual positions.

Table 11: Full ablation results on AVHBench for VideoLLaMA2. We report the individual effects of audio attention boosting and visual masking. Overall, V→A Hallu., A→V Hallu., and AV Match are accuracies (%). Best results are in bold.

Experiment / Method	Boost	Visual Mask	Overall	V→A Hallu.	A→V Hallu.	AV Match
Baseline (No Intervention)	–	–	70.65	77.60	80.11	56.45
Audio Attention Boosting Only						
first6_boost1.5	×1.5	–	70.75	77.69	79.75	56.82
first6_boost2	×2	–	71.03	77.86	80.02	57.25
first6_boost3	×3	–	71.48	78.08	80.11	58.21
first9_boost1.5	×1.5	–	70.82	77.73	80.02	56.82
first9_boost2	×2	–	71.12	77.60	79.75	58.00
first9_boost3	×3	–	71.24	77.34	79.49	58.80
first13_boost1.5	×1.5	–	70.61	77.42	79.75	56.77
first13_boost2	×2	–	70.60	77.47	79.75	56.66
first13_boost3	×3	–	71.01	77.42	79.93	57.78
Visual Masking Only on Layers 18–26						
last9_attnhigh_12	–	High 12.5%	71.31	**78.12**	**80.19**	57.62
last9_attnhigh_25	–	High 25%	71.31	77.86	**80.19**	57.94
last9_attnhigh_50	–	High 50%	**71.65**	**78.12**	79.58	**58.96**
last9_attnhigh_75	–	High 75%	71.44	77.90	79.40	58.74
last9_attnlow_12	–	Low 12.5%	70.52	77.51	79.93	56.29
last9_attnlow_25	–	Low 25%	70.56	77.47	80.11	56.34
last9_attnlow_50	–	Low 50%	70.67	77.38	79.75	56.98
last9_attnlow_75	–	Low 75%	70.97	77.60	80.02	57.41