When Aggregate Accuracy Does Not Identify Functional Change: Sharp Bounds on AV-LLM Adaptation Effects

Anonymous authors
Paper under double-blind review

Abstract

Fine-tuning Audio-Visual Large Language Models (AV-LLMs) is commonly evaluated by in-domain and cross-domain joint accuracy, measured with both audio and video available. Yet the same reported accuracies can conceal different functional changes. Even audio-only, video-only, and joint accuracies omit whether the two isolated inputs succeed on the same questions. We define isolated-path coverage as the fraction of questions answered correctly by at least one single-modality input. Conversion is joint accuracy minus that coverage. From before-and-after aggregate accuracies, we derive a sharp identified set for the conversion change and establish exactly when its sign is identifiable. The joint-minus-video gap combines conversion with the fraction of questions answered correctly with audio alone but not video alone. Its change can therefore have the opposite sign to the conversion change. Our audit spans 56 checkpoint–dataset evaluations and 40 before-and-after fine-tuning comparisons across Qwen2.5-Omni, Qwen3-Omni, MiniCPM-o, and VITA. Aggregate reports identify none of the 40 conversion-change signs, while the conventional gap points in the opposite direction in 17 comparisons. Paired per-question evaluations recover the missing overlap. They reveal a cross-domain conversion reversal under the same adaptation and a coverage gain offset by a conversion loss behind a nearly unchanged joint score. The balance between coverage and conversion varies across adaptation datasets and model families. AV-LLM adaptation calls for paired functional profiles alongside joint accuracy to establish what changed.

1 Introduction

Modality-aware fine-tuning of Audio-Visual Large Language Models (AV-LLMs) targets modality-specific learning and cross-modal interaction (Wei et al., 2025; Zhang et al., 2026b). A gain in accuracy with both audio and video (*joint accuracy*) does not reveal which functional changes account for the improvement. Training may expand the set of questions answered correctly with either modality alone, change the joint outcome relative to those answers, or combine both effects. Distinguishing these changes matters for progress on joint-reasoning benchmarks (Chao et al., 2026) and capability retention in continual Audio-Visual Question Answering (AVQA) (Wu et al., 2025). A gain driven by isolated correct answers can be mistaken for better joint use; a flat transfer score can hide opposing changes. Such misreadings can misdirect method comparisons and subsequent training. We ask what evaluation record can establish the functional effect of AV-LLM adaptation.

Answering that question requires knowing how audio-only (A) and video-only (V) correctness overlap on the same examples. Standard reports give their separate accuracies and accuracy with joint audio-video (AV) input (Yang et al., 2022; Li et al., 2022; Chao et al., 2026), often alongside the gain over video alone, but omit that overlap. Identical reports can therefore hide oppositely signed functional effects; more questions only estimate the same marginals more precisely. The joint-minus-video contrast also changes when audio-exclusive correct coverage changes.

We formulate the adaptation effect as a partial identification problem (Manski, 2003). At each checkpoint, *isolated-path coverage* is the probability that the model answers correctly with audio alone or

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

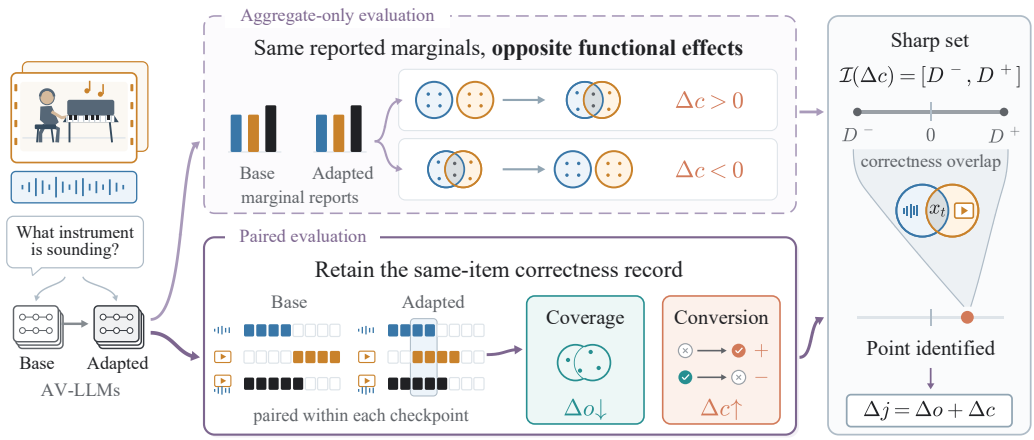


Figure 1: Aggregate versus paired evaluation of adaptation in AV-LLMs. "Base" and "Adapted" denote checkpoints before and after fine-tuning. Blue and orange circles represent the audio-only and video-only correct sets, respectively; dots are questions and x_t is their overlap for checkpoint t. Top: identical A/V/AV accuracies admit opposite adapted-minus-base conversion changes Δc. The sharp set $\mathcal{I}(\Delta c) = [D^-, D^+]$ contains all compatible effects, from the smallest (D^-) to the largest (D^+). Bottom: pairing predictions on each question recovers the correctness overlap and identifies $\Delta j = \Delta o + \Delta c$, with j joint accuracy, o isolated-path coverage, and $c = j - o$ conversion.

video alone; *conversion* is joint accuracy minus that coverage. With Δ denoting adapted minus base and AV denoting joint accuracy, every adaptation satisfies $\Delta\text{AV} = \Delta\text{coverage} + \Delta\text{conversion}$.

We derive sharp bounds on the conversion change and conditions for identifying its sign from aggregate reports. Paired item-level correctness supplies the missing A/V overlap and identifies both effects; publishing the overlap change alongside the six marginals is also sufficient. Existing per-question ablation predictions already contain this information. Aligning their correctness records reveals which questions both isolated inputs answer correctly, and how their overlap changes after adaptation. Figure 1 summarizes the progression from aggregate reports that admit multiple effects to point identification from paired correctness records.

The ambiguity is systematic in our audit. Across 40 base-to-adapted comparisons involving Qwen2.5-Omni, Qwen3-Omni, MiniCPM-o, and VITA, aggregate reports identify none of the conversion-change signs, while the conventional modality gap has the opposite sign in 17 comparisons. On independently sourced QIVD (Pourreza et al., 2026), five Qwen2.5-Omni-7B adaptations (Xu et al., 2025a) average a -0.34-point change in joint accuracy, comprising $+1.33$ points of coverage and -1.67 points of conversion. The same runs improve conversion on AVQA but reduce it on source-isolated Daily-Omni. MUSIC-AVQA adaptation and other model families yield different balances. Paired evaluation thus reveals functional changes that aggregate reports cannot resolve across domains and models.

Our contributions are as follows:

- We establish sharp identification limits for AV-LLM adaptation. Coverage and conversion distinguish two sources of accuracy change. We derive sharp bounds and necessary and sufficient conditions for identifying the conversion change or its sign from aggregate reports, and characterize modality-gap proxy reversals.
- We specify the evaluation record that resolves the ambiguity. Paired A/V correctness at each checkpoint, or equivalently the change in their correctness overlap alongside standard marginals, point-identifies the functional effects. This gives evaluation reports a precise information requirement.
- We show that this requirement changes empirical conclusions. Across datasets, training runs, and AV-LLM architectures, the paired audit uncovers proxy reversals, cross-domain changes, and opposing effects that aggregate accuracy conceals.

2 Related Work

Audio-visual evaluation and modality interaction. Joint audio-video models (Ghosh et al., 2026) motivate benchmarks for joint reasoning and temporal alignment (Chao et al., 2026; Zhao et al., 2026). Reliability studies extend evaluation to conflicting inputs (Chowdhury et al., 2025; Zhang et al., 2026a) and fusion oracles with a known attacked source (Yang et al., 2021). Per-example failures, cross-input consistency, information decomposition, and information flow then characterize how modalities interact within a fixed model (Bui et al., 2026; van Sprang et al., 2026; Fang et al., 2026; Jung et al., 2026).

Multimodal adaptation and transfer. Modality-aware adapters balance modality-specific learning and cross-modal interaction (Wei et al., 2025; Zhang et al., 2026b), complemented by token routing and knowledge retention (Shah & Wagle, 2026; Liang et al., 2025; Xie et al., 2025). Continual tuning and AVQA benchmarks extend this goal to forgetting and transfer (Ge et al., 2025; Wu et al., 2025; Huang et al., 2025); alignment, separate reasoning paths, and missing-modality training address input use (Liu et al., 2026; Li et al., 2026; Park et al., 2024).

Identification perspective. Partial identification (Manski, 2003) and classical Fréchet–Hoeffding bounds (Fréchet, 1951) characterize what marginal reports determine. We derive a sharp set and sign conditions for the *before–after change in conversion*, show how paired evaluation identifies that change, and explain modality-gap confounding. Appendix F compares evaluation records.

3 Identifying Functional Change

This section establishes what information is needed to identify adaptation-induced changes in coverage and conversion.

3.1 Functional estimands and observables

Our estimands are the functional changes induced by adaptation; the observables are the accuracies available in an evaluation report. Let $t \in \{0, 1\}$ index the base ($t = 0$) and adapted ($t = 1$) checkpoints, both frozen during evaluation. For a question drawn from a fixed evaluation population, let $A_t, V_t, J_t \in \{0, 1\}$ indicate correctness with audio-only, video-only, and joint audio-video input, respectively. We use J_t for joint correctness to distinguish it from the product $A_t V_t$. The reported marginal accuracies are $a_t = \mathbb{E}[A_t]$, $v_t = \mathbb{E}[V_t]$, and $j_t = \mathbb{E}[J_t]$.

Define the per-question union of correct isolated paths, its average coverage, and the conversion residual by

$$O_t = A_t \vee V_t, \qquad o_t = \mathbb{E}[O_t], \qquad c_t = j_t - o_t. \tag{1}$$

o_t is the fraction of questions answered correctly by at least one isolated input. If $r_t = \Pr(J_t = 1, O_t = 0)$ is the joint-rescue rate and $f_t = \Pr(J_t = 0, O_t = 1)$ is the rate at which the combined prediction loses an available isolated answer, then $c_t = r_t - f_t$. Thus conversion is a signed, label-dependent behavioral functional.

Writing $\Delta z = z_1 - z_0$, every adaptation satisfies

$$\Delta j = \Delta o + \Delta c. \tag{2}$$

Figure 2 follows the construction behind Equation 1 and Equation 2: three controlled input views pass through the same frozen checkpoint; their paired correctness record gives o_t and c_t; comparing those quantities before and after adaptation gives the two adaptation effects. The identification question is what remains of this record after it is reduced to aggregate accuracies.

3.2 Non-identification from aggregate accuracy

Let $x_t = \Pr(A_t = 1, V_t = 1)$ denote the unreported overlap between correct isolated paths. Inclusion–exclusion gives $o_t = a_t + v_t - x_t$, and hence

$$\Delta c = \Delta j - \Delta a - \Delta v + \Delta x. \tag{3}$$

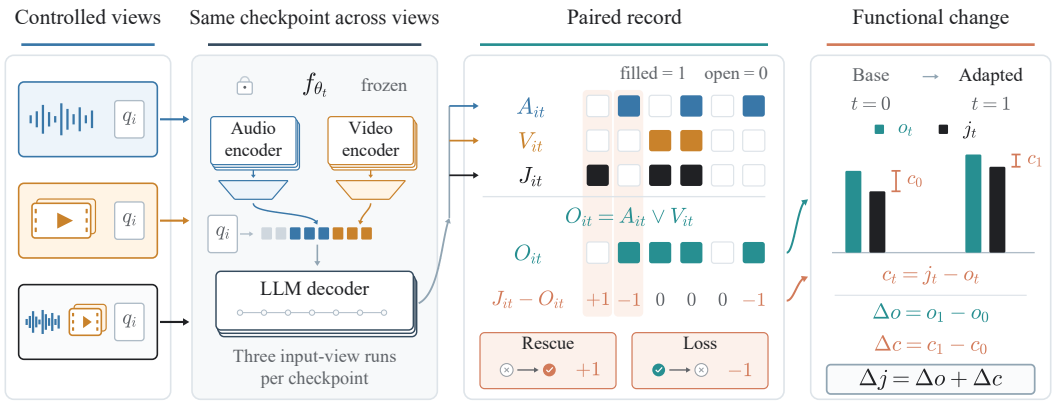


Figure 2: Paired functional evaluation. For question q_i, frozen model f_{θ_t} with parameters θ_t receives A/V/AV input; filled cells mark correct answers. Item i has correctness indicators A_{it}, V_{it}, J_{it}, coverage $O_{it} = A_{it} \vee V_{it}$, and residual $J_{it} - O_{it}$: joint rescue $(+1)$ or loss of an isolated success (-1). Row averages give o_t, c_t in Equation 1; comparing "Base" $(t = 0)$ and "Adapted" $(t = 1)$ gives Equation 2. Records and bar heights are schematic.

The six before-and-after marginal accuracies determine the first three terms but omit Δx. Consequently, even exact aggregate accuracies do not in general determine the magnitude or sign of Δc.

For example, suppose both checkpoints report $(a_t, v_t, j_t) = (0.5, 0.5, 0.6)$. A maximally disjoint A/V table has $o_t = 1$ and $c_t = -0.4$, while a nested table has $o_t = 0.5$ and $c_t = 0.1$. Using the disjoint table before adaptation and the nested table afterward gives $\Delta c = +0.5$; exchanging them gives $\Delta c = -0.5$. The two adaptations are observationally equivalent under aggregate reporting.

Formally, if $\mathcal{P}(a_t, v_t, j_t)$ denotes all joint distributions over (A_t, V_t, J_t) with the reported marginals, the identified set contains every value attained by a distribution in $\mathcal{P}(a_0, v_0, j_0) \times \mathcal{P}(a_1, v_1, j_1)$. A parameter is point identified only when this set is a singleton; its strict sign is identified only when the entire set lies above or below zero.

3.3 Sharp bounds and identification conditions

The overlap obeys the classical Fréchet–Hoeffding bounds (Fréchet, 1951)

$$\max(0, a_t + v_t - 1) \le x_t \le \min(a_t, v_t). \tag{4}$$

Substitution yields the sharp identified sets for each checkpoint:

$$L_t := \max(a_t, v_t) \le o_t \le U_t := \min(1, a_t + v_t), \qquad c_t \in [j_t - U_t, j_t - L_t]. \tag{5}$$

Maximal overlap nests the smaller correct set in the larger and attains L_t; minimal overlap makes the sets as disjoint as possible and attains U_t. Theorem 1 combines these bounds across checkpoints.

Theorem 1 (Sharp identification of adaptation effects)**.** *Given only (a_t, v_t, j_t) for $t = 0, 1$ on the same evaluation population, the sharp identified set for the conversion change is*

$$\mathcal{I}(\Delta c) = [D^-, D^+], \qquad \begin{cases} D^- = (j_1 - U_1) - (j_0 - L_0), \\ D^+ = (j_1 - L_1) - (j_0 - U_0). \end{cases} \tag{6}$$

The aggregates identify a strictly positive conversion change if and only if $D^- > 0$, and a strictly negative change if and only if $D^+ < 0$. They point identify Δc exactly when $L_0 = U_0$ and $L_1 = U_1$.

The lower endpoint pairs the smallest adapted conversion with the largest base conversion; the upper endpoint makes the opposite choices. Both are attainable because aggregate tables impose no restriction linking the two correctness overlaps. The set width is

$$D^+ - D^- = (U_0 - L_0) + (U_1 - L_1), \tag{7}$$

which depends only on overlap uncertainty, not on j_t, the model's accuracy with both modalities.

The finite-population result is also exact. For N questions with A/V correct counts n_{A_t}, n_{V_t}, every coverage count from $\max(n_{A_t}, n_{V_t})$ to $\min(N, n_{A_t} + n_{V_t})$ is feasible. The Δc values form a contiguous grid from D^- to D^+ in steps of $1/N$; Appendix A gives both constructions.

3.4 Point identification with paired evaluation

Retaining audio-only and video-only correctness on the same questions reveals x_t and collapses Equation 5 to a point. For questions $i = 1, \ldots, N$,

$$\widehat{o}_t = \frac{1}{N}\sum_i (A_{it} \vee V_{it}), \qquad \widehat{c}_t = \frac{1}{N}\sum_i [J_{it} - (A_{it} \vee V_{it})]. \tag{8}$$

On this evaluation set, A/V pairing at both checkpoints identifies the mean $\Delta c = \widehat{c}_1 - \widehat{c}_0$. Evaluating the same questions before and after adaptation additionally identifies per-question transitions and stable-path strata, but is not required for the mean effect. Equivalently, Equation 3 shows that publishing Δx alongside the six marginals is a sufficient one-scalar augmentation for Δc.

Binary question-only correctness Q_t locates isolated-path coverage relative to success without media. Retaining it partitions coverage into two joint masses:

$$o_t = o_t^Q + o_t^{\neg Q}, \quad o_t^Q = \mathbb{E}[Q_t O_t], \quad o_t^{\neg Q} = \mathbb{E}[(1 - Q_t) O_t], \tag{9}$$

and therefore

$$\Delta j = \Delta o^Q + \Delta o^{\neg Q} + \Delta c. \tag{10}$$

The terms Δo^Q and $\Delta o^{\neg Q}$ track how coverage changes in the Q-correct and Q-wrong portions of the evaluation for each checkpoint.

3.5 Proxy confounding

Let $p_{10,t} = \Pr(A_t = 1, V_t = 0)$ and $p_{01,t} = \Pr(A_t = 0, V_t = 1)$. The binary identities $O_t = V_t + A_t(1 - V_t)$ and $O_t = A_t + V_t(1 - A_t)$ give

$$j_t - v_t = c_t + p_{10,t}, \qquad \Delta(j - v) = \Delta c + \Delta p_{10}, \tag{11}$$

$$j_t - a_t = c_t + p_{01,t}, \qquad \Delta(j - a) = \Delta c + \Delta p_{01}. \tag{12}$$

Thus the conventional AV-minus-V change equals the conversion change only under the additional restriction $\Delta p_{10} = 0$. It has the opposite sign exactly when

$$\Delta c\big(\Delta c + \Delta p_{10}\big) < 0. \tag{13}$$

An unchanged proxy can likewise hide a nonzero conversion effect. Appendix A gives attainable examples of both cases.

Coverage and conversion may also cancel in the total AV effect. Define

$$H = \frac{|\Delta o| + |\Delta c| - |\Delta j|}{2}. \tag{14}$$

Then $H = 0$ when the components share a weak sign and $H = \min(|\Delta o|, |\Delta c|)$ otherwise. Proxy confounding and cancellation are distinct: the former adds changing exclusive coverage to a modality gap; the latter hides opposing identified components in their aggregate sum.

4 Experiments and Findings

The experiments follow the information hierarchy in Section 3. We first audit what standard aggregate reports identify, then use paired predictions to measure the hidden effects across training runs, datasets, and architectures.

4.1 EXPERIMENTAL SETUP

Datasets and source isolation. AVQA supplies the main 5,000-question adaptation set and in-domain development and public-test evaluations (Yang et al., 2022). We exclude training-source overlap from those evaluations and from strict Daily-Omni (Zhou et al., 2026). QIVD is an independently sourced external test with one short-answer question per independently collected video (Pourreza et al., 2026). MUSIC-AVQA (MUSIC) supplies a second adaptation setting with source-disjoint original-media development and held-out test splits (Li et al., 2022). Appendix B gives all split sizes and roles.

Models and adaptation. Our main setting uses rank-16 low-rank adaptation (LoRA) on the query/value (Q/V) projections of Qwen2.5-Omni-7B's language model (Xu et al., 2025a; Hu et al., 2022). The main AVQA comparisons use the same 5,000 training questions, 157 optimizer steps, and global batch size 32. Model-specific media processors remain fixed, the perception encoders are frozen, and inference uses greedy answer-only decoding. MiniCPM-o 4.5 (Cui et al., 2026), Qwen3-Omni-30B-A3B (Xu et al., 2025b), and VITA-1.0/Mixtral-8x7B (Fu et al., 2025) extend the comparison across architectures; Qwen2.5-Omni-3B provides a same-family scale check. Training configurations and run counts appear in Appendix B.

Paired input conditions. Every checkpoint is evaluated on question-only (Q), audio-only (A), video-only (V), and combined audio-video (AV) input with the same question, answer space, decoding rule, and media preprocessing. Multiple-choice tasks use exact option accuracy; QIVD uses token-normalized exact match against its official concise answer. All analyses retain per-example predictions and original source IDs. Effects are in percentage points (pp). Repeated runs are summarized by mean and sample standard deviation; figures show point estimates. In results, A/V/AV denote accuracies; ΔAV, ΔO, and ΔC equal $100\Delta j$, $100\Delta o$, and $100\Delta c$, respectively.

4.2 AGGREGATE REPORTS DO NOT IDENTIFY THE ADAPTATION EFFECT

The audit comprises 16 base and 40 adapted checkpoint–dataset evaluations, forming 40 before–after comparisons. Audio-only, video-only, and joint accuracies identify conversion signs in only 7 of the 56 evaluations and none of the 40 conversion-change signs. Retaining the item-level A/V overlap point identifies the change in every comparison. The conventional change in AV-minus-V has the opposite sign in 17 comparisons and is zero while hiding a conversion change in two more. Appendix C lists all 40 comparisons and sharp sets.

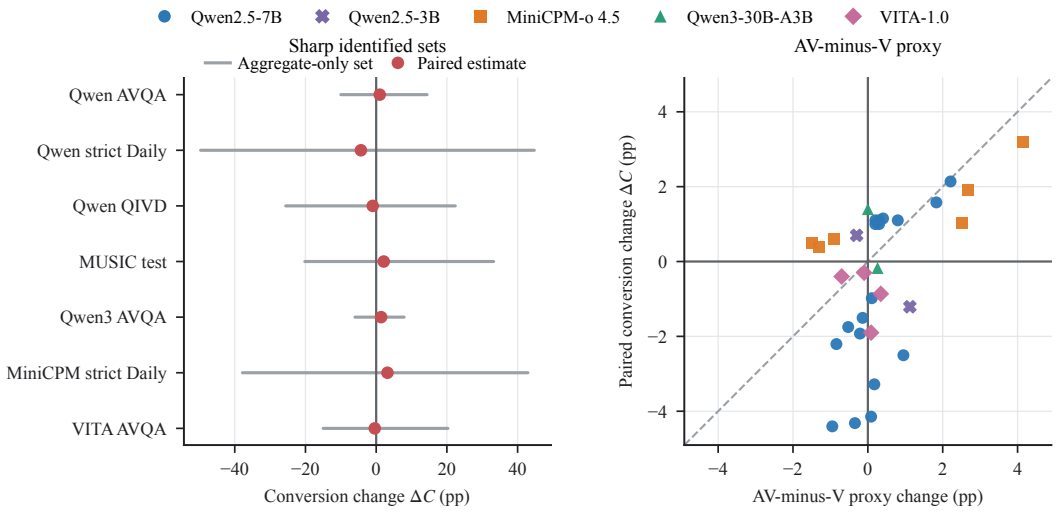


Figure 3: Aggregate accuracy leaves functional change unidentified. Left: gray sharp sets cross zero; red points are nonzero paired conversion-change estimates. Right: AV-minus-V proxy changes versus conversion changes; the dashed diagonal marks agreement, and colors identify models.

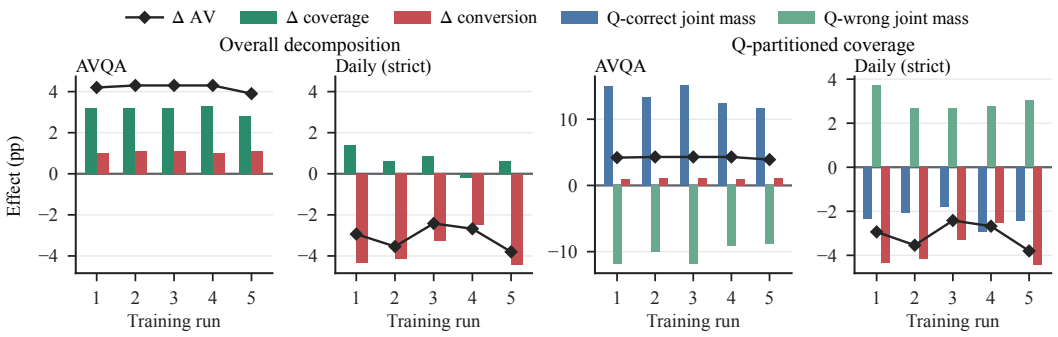


Figure 4: Conversion changes reverse across domains in every run. Left pair: AVQA gains combine coverage and conversion increases; strict, source-isolated Daily-Omni nearly preserves mean coverage but loses conversion. Right pair: partitioning coverage by Q-only correctness reveals opposing shifts in joint masses defined at each checkpoint. Bars sum to the black AV effect.

Figure 3 illustrates the ambiguity in the evaluated comparisons. Each gray sharp set spans both signs, and every point in the set is compatible with the same before-and-after marginal accuracies. More precise marginal estimates cannot recover the missing correctness overlap. The conventional gap can also mask or reverse the observed conversion change. Qwen3 on AVQA has $\Delta(\mathrm{AV} - V) = 0$ while $\Delta C = +1.40$ points. Across the three MiniCPM AVQA runs, conversion improves while the conventional gap decreases. These comparisons illustrate the confounding identity in Equation 11.

Aggregate cancellation is also common: coverage and conversion oppose one another in 17 of 40 comparisons. The hidden amount H reaches 2.25 points, and the QIVD runs repeatedly conceal roughly one to two points of opposing functional change behind a small total. Thus all three theoretical cases—an unidentified sign, a misleading proxy, and cancellation in accuracy with both modalities—occur in the audited evaluations.

4.3 A FIVE-RUN CONVERSION REVERSAL ACROSS DOMAINS

Qwen2.5-Omni-7B shows a stable cross-domain reversal. Across five independent AVQA fine-tuning runs, in-domain accuracy with both modalities improves by 4.20 ± 0.17 points, comprising 3.14 ± 0.19 points of coverage and 1.06 ± 0.05 points of conversion. On strict Daily-Omni, AV instead changes by -3.07 ± 0.58 points. Coverage is nearly preserved at $+0.66 \pm 0.56$, while conversion falls by 3.73 ± 0.82 points. All five runs have positive in-domain and negative external conversion changes (Figure 4).

The right pair of Figure 4 locates coverage changes relative to question-only correctness by partitioning coverage into the joint masses $\mathbb{E}[Q_t O_t]$ and $\mathbb{E}[(1 - Q_t)O_t]$. On AVQA, the $+3.14$-point coverage effect splits into $+13.52$ and -10.38 points in these two terms. On strict Daily-Omni, the $+0.66$ total splits into -2.31 and $+2.97$ points. External conversion therefore declines while the Q-wrong joint-mass contribution increases in every run, revealing different reorganizations of joint-input behavior and isolated-path coverage.

The 2,957-question AVQA public test provides an independent in-domain split. With both modalities, model accuracy rises from 89.65% to 93.51%, while coverage rises from 93.27% to 95.98%. The resulting $+3.86$-point effect again separates into $+2.71$ coverage and $+1.15$ conversion. The development and public-test decompositions thus agree at the component level.

4.4 EXTERNAL CANCELLATION AND STABLE-PATH PERSISTENCE

QIVD gives the clearest cancellation case. Across the same five fine-tuning runs, AV changes by -0.34 ± 0.61 points and has mixed signs across runs. The same predictions contain two stable opposing effects: coverage rises by 1.33 ± 0.28 points in all five runs, while conversion falls by 1.67 ± 0.46 points in all five. The near-zero aggregate change is therefore the sum of two repeatable changes in opposite directions.

Table 1: Primary Qwen2.5-Omni-7B adaptation effects (percentage points). Boldface marks the larger absolute component between ΔO and ΔC in each row; $\Delta\text{AV} = \Delta O + \Delta C$.

Adaptation	Evaluation	ΔAV	ΔO	ΔC
AVQA	AVQA dev	$+4.20$	$\mathbf{+3.14}$	$+1.06$
	AVQA public test	$+3.86$	$\mathbf{+2.71}$	$+1.15$
	Strict Daily-Omni	-3.07	$+0.66$	$\mathbf{-3.73}$
	QIVD external test	-0.34	$+1.33$	$\mathbf{-1.67}$
MUSIC	MUSIC dev	$+11.16$	$\mathbf{+9.58}$	$+1.58$
	MUSIC held-out test	$+10.84$	$\mathbf{+8.70}$	$+2.14$

We next remove every question whose A/V coverage membership changes across adaptation. Within this stable-coverage subset, AV changes by $+1.86$ points on AVQA and -2.87 on strict Daily-Omni, with the respective signs repeated in all five runs. A stricter subset requires the parsed audio-only and video-only answers themselves to remain unchanged; it still gives $+1.03$ and -2.57 points. The reversal therefore persists under fixed coverage membership and fixed parsed isolated answers, locating it in joint-answer changes on stable isolated paths.

Table 1 collects these effects and the MUSIC results below. Appendix B.2 illustrates representative item-level transitions.

4.5 Architectures yield distinct reorganizations

Applying the shared AVQA adaptation protocol to four architectures produces distinct reorganizations, summarized in Table 2 and Figure 5. Each model uses the same 5,000 training questions and 157 optimizer steps.

Across three MiniCPM-o runs, the $+3.53$-point in-domain AV gain is primarily a $+3.03$-point coverage expansion. On strict Daily-Omni, an almost unchanged AV total combines -2.19 points of coverage with $+2.04$ points of conversion. All three runs share these component directions. MiniCPM therefore exhibits an external coverage loss counterbalanced by increased conversion, the reverse of Qwen2.5's preserved-coverage conversion loss.

Qwen3-Omni's in-domain gain is dominated by conversion: $+1.90 = +0.50 + 1.40$ points. Its AV-minus-V gap is unchanged even though conversion improves, illustrating the zero-proxy case in Section 3.5. On strict Daily-Omni, both coverage and conversion change by less than one percentage point. The conversion-dominated in-domain gain therefore does not carry over to this external evaluation.

VITA supplies a non-Qwen, Mixtral-based language backbone. Its AVQA gain is $+6.30 = +6.65 - 0.35$ points, making the result even more coverage dominated than MiniCPM. On strict Daily-Omni, AV decreases in both runs, by 2.98 points on average, while coverage and conversion estimates are both negative.

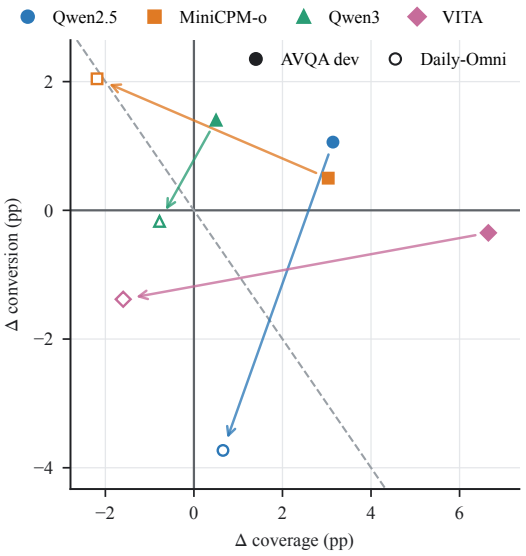


Figure 5: Functional reorganization across model families. Filled markers show AVQA dev; open markers show strict, source-isolated Daily-Omni. Arrows connect domains; the dashed line denotes zero AV change.

The shared protocol produces coverage-dominated or conversion-dominated gains across models. Their external profiles also differ in which component improves or deteriorates.

Table 2: Cross-architecture effects under AVQA adaptation (percentage points). Boldface marks the larger absolute component between ΔO and ΔC within each domain.

Model	AVQA development			Strict Daily-Omni		
	ΔAV	ΔO	ΔC	ΔAV	ΔO	ΔC
Qwen2.5-Omni-7B	$+4.20$	$\mathbf{+3.14}$	$+1.06$	-3.07	$+0.66$	$\mathbf{-3.73}$
MiniCPM-o 4.5	$+3.53$	$\mathbf{+3.03}$	$+0.50$	-0.14	$\mathbf{-2.19}$	$+2.04$
Qwen3-Omni-30B-A3B	$+1.90$	$+0.50$	$\mathbf{+1.40}$	-0.95	$\mathbf{-0.78}$	-0.17
VITA-1.0	$+6.30$	$\mathbf{+6.65}$	-0.35	-2.98	$\mathbf{-1.60}$	-1.38

Paired predictions make these profiles concrete (Appendix B.2). On the AVQA sound-source question, Qwen3's isolated answers both remain "sound of wind," while its joint answer changes from "leaf" to the correct choice. Qwen2.5 repairs the joint answer on the police-car question in all five runs, with unchanged isolated answers. On Daily-Omni, the attire question before a spoken "Whoa" shows the reverse: both isolated inputs retain the correct black dress choice, yet the joint answer switches to the incorrect silver dress choice in every run. These transitions expose conversion changes on questions with preserved isolated correct answers.

4.6 A SECOND ADAPTATION DATASET CHANGES THE BALANCE

Changing the adaptation dataset also changes the functional balance. Table 1 shows that MUSIC adaptation is strongly coverage-dominated. On the original-media development set, accuracy with both modalities rises from 68.69% to 79.85% and coverage rises from 77.27% to 86.85%. The effect is $+11.16 = +9.58 + 1.58$ points, so coverage accounts for about 86% of the net gain. Conversion improves, while joint accuracy remains below coverage after adaptation.

The Q-partitioned decomposition separates the development coverage gain into $+5.10$ points of Q-correct joint mass and $+4.49$ points of Q-wrong joint mass. AVQA instead combines a large positive Q-correct change with a negative Q-wrong change. The adaptations differ in both the amount of coverage gained and its relation to question-only success.

The held-out MUSIC test repeats the development decomposition. AV changes from 67.71% to 78.54% and coverage from 76.51% to 85.21%, yielding $+10.84 = +8.70 + 2.14$ points. The matching decompositions show that the dominant coverage gain and smaller conversion gain both carry over to the held-out split.

5 DISCUSSION AND CONCLUSION

A nearly unchanged joint score can conceal repeatable opposing changes, as on QIVD; an unchanged AV-minus-V gap can conceal improved conversion, as for Qwen3-Omni. Across domains, model families, and adaptation datasets, higher joint accuracy has different functional sources. Comparing coverage and conversion reveals whether adaptation expands isolated-path success, improves the joint outcome relative to that coverage, or trades one effect against the other.

The sharp bounds specify the information needed for that comparison. The six marginal accuracies delimit the possible conversion change; adding the A/V correctness-overlap change determines it exactly. Paired predictions recover this quantity and preserve the item-level transitions needed to locate the effect. In transfer evaluations, reporting $(\Delta o, \Delta c)$ alongside Δj distinguishes changes in isolated-path coverage from changes in conversion. Even an unchanged total can then reveal which functional component improved and which declined.

AVQA and MUSIC differ in how coverage relates to question-only success, while the Qwen2.5 reversal persists when isolated answers stay fixed. These findings give correctness-guided post-training (He et al., 2026; Nam, 2026) and modality-aware optimization (Chaubey et al., 2026) concrete evaluation targets: expand isolated-path coverage, improve conversion, and measure any trade-off across domains. AV-LLM adaptation should therefore be reported with both its total performance and the paired input record needed to establish what changed.

AI USE STATEMENT

In this work, generative AI tools were used in an assistive role for language polishing, literature discovery, code debugging, and feedback on experimental plans. The authors were responsible for the research questions, methodology, mathematical development, experiments, and conclusions. AI-assisted outputs, including text edits and code changes, were independently checked by the authors. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 3 states the assumptions; Appendix A gives complete proofs and finite-sample constructions. Section 4.1 and Appendix B describe data, training, and evaluation; Appendix C reports all 40 comparisons. Retained model revisions, configurations, split manifests, and paired predictions allow the analysis scripts to reproduce the decompositions, sharp sets, and empirical figures.

REFERENCES

Quang Bui, Shlok Jaiswal, Samuel Paik-Heintz, Kevin Zhou, Kaushik Madapati, Krittaphas Chaisutyakorn, et al. Loud or silent? A reusable framework for per-modality failure analysis in multimodal clinical AI. *arXiv preprint arXiv:2608.01462*, 2026.

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models. *arXiv preprint arXiv:2505.19616*, 2026a.

Xinyue Cai, Chaoyou Fu, Yi-Fan Zhang, Ran He, and Caifeng Shan. OmniVideo-100K: A dataset for audio-visual reasoning through structured scripts and evidence chains. *arXiv preprint arXiv:2606.14702*, 2026b.

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. JointAVBench: A benchmark for joint audio-visual reasoning evaluation. In *International Conference on Learning Representations*, 2026.

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. MoD-DPO: Towards mitigating cross-modal hallucinations in omni LLMs using modality decoupled preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. AVTrustBench: Assessing and enhancing reliability and robustness in audio-visual LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Sangyun Chung, Se Yeon Kim, Youngchae Chee, and Yong Man Ro. MAD: Modality-adaptive decoding for mitigating cross-modal hallucinations in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Yuchen Deng, Chang Sun, Hai-Tao Zheng, Feidiao Yang, and Yuxing Han. CAD: Conflict-aware decoding to mitigate cross-modal hallucinations in omnimodal large language models. *arXiv preprint arXiv:2609.04247*, 2026.

Zixuan Dong, Jiafu Tang, Zhide Lei, Zhe Cao, Zijie Zhang, Yanghai Wang, Shihao Li, Xiaodong Wang, Baoyun Peng, and Jiaheng Liu. OmniHalluc-L: Counterfactual benchmarking and modality-perturbation reliability calibration for long-form omni hallucination. *arXiv preprint arXiv:2606.03614*, 2026.

Wanlong Fang, Tianle Zhang, Wen Tao, and Alvin Chan. Towards understanding modality interaction in multimodal language models via partial information decomposition. *arXiv preprint arXiv:2606.00959*, 2026.

Maurice Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, Sciences, Section A*, 1951.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yifan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, Haoyu Cao, Di Yin, Long Ma, Xiawu Zheng, Rongrong Ji, Yunsheng Wu, Ran He, Caifeng Shan, and Xing Sun. VITA: Towards open-source interactive omni multimodal LLM. *arXiv preprint arXiv:2408.05211*, 2025.

Chendi Ge, Xin Wang, Zeyang Zhang, Hong Chen, Jiapei Fan, Longtao Huang, Hui Xue, and Wenwu Zhu. Dynamic mixture of curriculum LoRA experts for continual multimodal instruction tuning. In *International Conference on Machine Learning*, 2025.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Siddharth Gururani, Hanrong Ye, Pritam Biswas, Yuanhang Su, Ehsan Hosseini-Asl, et al. Audio-Visual Flamingo: Open audio-visual intelligence for long and complex videos. *arXiv preprint arXiv:2607.16107*, 2026.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, et al. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. In *International Conference on Learning Representations*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chengyue Huang, Brisa Maneechotesuwan, Shivang Chopra, and Zsolt Kira. FRAMES-VQA: Benchmarking fine-tuning robustness across multi-modal shifts in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jihoo Jung, Chaeyoung Jung, Ji-Hoon Kim, and Joon Son Chung. Probing cross-modal information hubs in audio-visual LLMs. In *International Conference on Machine Learning*, 2026.

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.

Xuanchen Li, Yuheng Lu, Chenrui Cui, Tianrui Wang, Zikang Huang, Yu Jiang, Long Zhou, Longbiao Wang, and Jianwu Dang. Separate first, fuse later: Mitigating cross-modal interference in audio-visual LLMs reasoning with modality-specific chain-of-thought. *arXiv preprint arXiv:2605.09906*, 2026.

Jian Liang, Wenke Huang, Guancheng Wan, Qu Yang, and Mang Ye. LoRASculpt: Sculpting LoRA for harmonizing general and specialized knowledge in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Hongcheng Liu, Yuhao Wang, Zhe Chen, Pingjie Wang, Zhiyuan Zhu, Yixuan Hou, Yanfeng Wang, and Yu Wang. Cross-modal coreference alignment: Enabling reliable information transfer in omni-LLMs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2026.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Hyeonuk Nam. Learning from audio-dependency errors: Data curation strategies based on model confusion patterns in audio question answering. *arXiv preprint arXiv:2606.22276*, 2026.

Kyu Ri Park, Hong Joo Lee, and Jung Uk Kim. Learning trimodal relation for audio-visual question answering with missing modality. In *European Conference on Computer Vision*, 2024.

Reza Pourreza, Rishit Dagli, Apratim Bhattacharyya, Sunny Panchal, Guillaume Berger, and Roland Memisevic. Can vision-language models answer face to face questions in the real-world? In *International Conference on Learning Representations*, 2026.

Shrey Shah and Justin Wagle. MoLoRA: Composable specialization via per-token adapter routing. In *International Conference on Machine Learning*, 2026.

Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M. Asano. Same content, different answers: Cross-modal inconsistency in MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yake Wei, Yu Miao, Dongzhan Zhou, and Di Hu. MokA: Multimodal low-rank adaptation for MLLMs. In *Advances in Neural Information Processing Systems*, 2025.

Kaixuan Wu, Xinde Li, Xinling Li, Chuanfei Hu, and Guoliang Wu. AVQACL: A novel benchmark for audio-visual question answering continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jingyi Xie, Jintao Yang, Zhunchen Luo, Yunbo Cao, Qiang Gao, Mengyuan Zhang, and Wenpeng Hu. AdaDARE-gamma: Balancing stability and plasticity in multi-modal LLMs through efficient adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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 ACM International Conference on Multimedia*, 2022.

Chenshuang Zhang, Kyeong Seon Kim, Chengxin Liu, and Tae-Hyun Oh. SVHalluc: Benchmarking speech-vision hallucination in audio-visual large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Mingfang Zhang, Yunhong Wang, Lu Wang, and Jiaxin Chen. Parameter-efficient adaptation for MLLMs via implicit modality decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Chen Zhao, Jiajun Ma, Qilong Huang, Tiehan Fan, Hongyu Li, Zhuoliang Kang, Xiaoming Wei, Jian Yang, and Ying Tai. Temporal and cross-modal alignment for enhanced audiovisual video captioning. In *European Conference on Computer Vision*, 2026.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-Omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2026.

A Proofs and Finite-Sample Constructions

A.1 Sharp bounds for a fixed checkpoint

Omitting the checkpoint subscript t, write $x = \Pr(A = 1, V = 1)$. The A/V table with marginals (a, v) has cells $(x, a-x, v-x, 1-a-v+x)$ for $(A, V) = (1,1), (1,0), (0,1), (0,0)$, respectively. Cell nonnegativity is equivalent to

$$\max(0, a + v - 1) \le x \le \min(a, v). \tag{15}$$

Every x in this interval defines a valid table, so the bounds are sharp. Since $o = a + v - x$, the maximum overlap $x = \min(a, v)$ yields $o = L = \max(a, v)$, whereas the minimum overlap

$x = \max(0, a + v - 1)$ yields $o = U = \min(1, a + v)$. For any such table, let J be Bernoulli with mean j independently of (A, V). This completes a valid joint distribution with the reported j, and $c = j - o$ realizes every point of $[j - U, j - L]$. No relationship between J and (A, V) is encoded in the three reported marginal accuracies.

For N questions, let n_A, n_V, n_J be the A/V/AV correct counts and n_{11} the A/V overlap count. Choose any integer n_{11} satisfying

$$\max(0, n_A + n_V - N) \le n_{11} \le \min(n_A, n_V). \tag{16}$$

Assign $n_{10} = n_A - n_{11}$, $n_{01} = n_V - n_{11}$, and $n_{00} = N - n_A - n_V + n_{11}$ questions to the other A/V cells. These counts are nonnegative by Equation 16. Mark any n_J of the N questions correct under J, regardless of their A/V cell. The resulting coverage counts are every integer from $\max(n_A, n_V)$ to $\min(N, n_A + n_V)$. Dividing by N gives the exact finite-sample coverage set: a contiguous grid with spacing $1/N$ and both bounds attained.

A.2 Proof of Theorem 1

Equation 5 yields $c_t \in [j_t - U_t, j_t - L_t]$. Minimizing $c_1 - c_0$ selects $o_1 = U_1$ and $o_0 = L_0$, giving D^-. Maximizing selects $o_1 = L_1$ and $o_0 = U_0$, giving D^+. Both pairs are attainable on the same evaluation population: construct the base and adapted A/V tables with their chosen overlaps and couple the two correctness tables in any way (the product coupling suffices). Conditional on each table, assign J_t with the specified mean j_t. Their marginals match all six reported accuracies. Varying the overlaps realizes every intermediate population effect, proving that the interval cannot be narrowed from these reports.

The finite-sample argument uses the N question indices in place of a population coupling. Construct the A/V cells independently for each checkpoint using Equation 16, then assign the binary outcomes at each checkpoint to the same indices in any order. Assign n_{J_0} and n_{J_1} correct joint answers independently of the A/V arrangements. If $n_{O_t} \in \{NL_t, \dots, NU_t\}$ denotes the feasible integer coverage counts, then the feasible effect counts are

$$N\Delta c = n_{J_1} - n_{J_0} - n_{O_1} + n_{O_0}. \tag{17}$$

Differences of the two contiguous integer grids form the entire grid $\{ND^-, ND^- + 1, \dots, ND^+\}$. The construction uses the original N indices and permits arbitrary coupling between checkpoints.

Sharpness gives the sign criterion immediately: every feasible effect is positive exactly when $D^- > 0$, and every feasible effect is negative exactly when $D^+ < 0$. The width of the set is $(U_0 - L_0) + (U_1 - L_1)$; both widths are nonnegative, so the effect is point-identified exactly when both checkpoint-specific sets have zero width.

For the example with $a_t = v_t = 0.5$ and $j_t = 0.6$ at both checkpoints, a maximally disjoint A/V table has $o_t = 1$ and $c_t = -0.4$, while a nested table has $o_t = 0.5$ and $c_t = 0.1$. Use the disjoint table before adaptation and the nested table afterward to obtain $\Delta c = 0.5$; exchanging them gives $\Delta c = -0.5$. Both constructions preserve the same six marginal accuracies.

A.3 Exact modality-gap identity

The binary identity $O_t = V_t + A_t(1 - V_t)$ gives $o_t - v_t = \Pr(A_t = 1, V_t = 0) = p_{10,t}$. Therefore $j_t - v_t = (j_t - o_t) + (o_t - v_t) = c_t + p_{10,t}$. Taking the difference between adapted and base checkpoints proves the first row of Equation 11. Interchanging A and V proves Equation 12.

A.4 Attainable proxy counterexamples

Here we again omit t; $p_{11}, p_{10}, p_{01}, p_{00}$ are the probabilities of $(A, V) = (1, 1), (1, 0), (0, 1), (0, 0)$, respectively. First consider

$$\begin{aligned} \text{Base: } & (0.3, 0.2, 0.2, 0.3), \quad j = 0.6, \\ \text{Adapted: } & (0.5, 0.1, 0.2, 0.2), \quad j = 0.8. \end{aligned}$$

Here $j - v = 0.1$ at both checkpoints, while c increases from -0.1 to 0. For a reversal, consider

$$\begin{aligned} \text{Base: } & (0.4, 0.1, 0.2, 0.3), \quad j = 0.6, \\ \text{Adapted: } & (0.3, 0.3, 0.1, 0.3), \quad j = 0.5. \end{aligned}$$

Now $j - v$ increases from 0 to 0.1, while c falls from -0.1 to -0.2. In both examples J can be assigned with the stated marginal accuracy using the construction above. Both examples are therefore jointly realizable.

A.5 Cancellation identity

By Equation 2, the triangle inequality gives $H \geq 0$. If Δo and Δc have the same weak sign, their absolute values add and $H = 0$. If their signs differ, then $|\Delta j| = \big||\Delta o| - |\Delta c|\big|$, so substitution gives $H = \min(|\Delta o|, |\Delta c|)$.

B Evaluation and Reporting Details

B.1 Data and adaptation configurations

Table 3 records the role of each split. Both-5k denotes a 5,000-question adaptation set trained with both audio and video inputs.

Table 3: Data splits and their role in the evidence chain.

Split	Questions	Sources	Role
AVQA Both-5k	5,000	4,934	Main adaptation
AVQA dev	1,000	975	In-domain development
AVQA public test	2,957	2,938	Held-out in-domain test
Strict Daily-Omni	1,158	662	External development
QIVD short-answer test	2,856	2,856	Independent external test
MUSIC Both-5k	5,000	3,814	Second adaptation
MUSIC original-media dev	5,574	1,080	Second development
MUSIC original-media test	5,611	1,098	Held-out second test

Table 4 reports the training configurations. All rows use rank-16 Q/V-LoRA for 157 optimizer steps with global batch size 32. The supplementary experiment manifest records exact run identifiers, checkpoint paths, and model revisions.

Table 4: Controlled adaptation settings and the number of repeated training runs. The final column is the number of parameters updated by Q/V-LoRA.

Model	Adaptation data	Runs	Parameters
Qwen2.5-Omni-7B	AVQA Both-5k	5	5,046,272
Qwen2.5-Omni-7B	MUSIC Both-5k	1	5,046,272
Qwen2.5-Omni-3B	AVQA Both-5k	1	3,686,400
MiniCPM-o 4.5	AVQA Both-5k	3	7,667,712
Qwen3-Omni-30B-A3B	AVQA Both-5k	1	6,684,672
VITA-1.0/Mixtral-8x7B	AVQA Both-5k	2	6,815,744

B.2 Representative paired predictions

Table 5 makes the aggregate effects concrete. In each row, an isolated answer remains available while the combined answer alone is repaired or lost. Both Qwen2.5 transitions recur in all five training runs.

C Complete Identification Audit

Tables 6 and 7 list all 40 comparisons. The audit covers 56 checkpoint–dataset evaluations: 16 base evaluations are reused across 40 comparisons with adapted checkpoints. Run numbers distinguish independent adapters; the supplementary manifest gives exact identifiers. Every paired

Table 5: Representative paired predictions. Entries list A/V/AV answer choices; bold denotes the correct prediction.

Setting	Question and correct answer	Base	Adapted
Qwen2.5, AVQA	What car makes a sound in the video? A: police car	$\mathbf{A}/B/C$	$\mathbf{A}/B/\mathbf{A}$
Qwen2.5, Daily	What was the contestant's attire right before male voice first exclaimed 'Whoa' during her performance? B: Short strapless black dress	$\mathbf{B}/\mathbf{B}/\mathbf{B}$	$\mathbf{B}/\mathbf{B}/D$
Qwen3, AVQA	What is the source of the sound in the video? C: sound of wind	$\mathbf{C}/\mathbf{C}/D$	$\mathbf{C}/\mathbf{C}/\mathbf{C}$

$\Delta C = 100\Delta c$ lies inside an aggregate-only sharp set crossing zero. The proxy is $\Delta(\mathrm{AV} - V) = 100\Delta(j - v)$; H from Equation 14 is also in percentage points.

Table 6: Complete aggregate-only identification audit, part I. Effects and sharp-set endpoints are percentage points.

Comparison	ΔC	Aggregate-only set	$\Delta(\mathrm{AV} - V)$	H
Qwen2.5 AVQA dev, run 1	$+1.00$	$[-10.00, 14.40]$	$+0.20$	0.00
Qwen2.5 AVQA dev, run 2	$+1.10$	$[-9.90, 14.40]$	$+0.20$	0.00
Qwen2.5 AVQA dev, run 3	$+1.10$	$[-9.90, 15.00]$	$+0.80$	0.00
Qwen2.5 AVQA dev, run 4	$+1.00$	$[-9.90, 14.50]$	$+0.30$	0.00
Qwen2.5 AVQA dev, run 5	$+1.10$	$[-10.30, 14.50]$	$+0.30$	0.00
Qwen2.5 strict Daily-Omni, run 1	-4.32	$[-49.65, 44.73]$	-0.35	1.38
Qwen2.5 strict Daily-Omni, run 2	-4.15	$[-50.26, 44.99]$	$+0.09$	0.60
Qwen2.5 strict Daily-Omni, run 3	-3.28	$[-49.14, 46.20]$	$+0.17$	0.86
Qwen2.5 strict Daily-Omni, run 4	-2.50	$[-49.40, 45.85]$	$+0.95$	0.00
Qwen2.5 strict Daily-Omni, run 5	-4.40	$[-50.52, 44.91]$	-0.95	0.60
Qwen2.5 MUSIC dev	$+1.58$	$[-19.61, 32.60]$	$+1.83$	0.00
Qwen2.5 MUSIC test	$+2.14$	$[-20.17, 33.22]$	$+2.21$	0.00
MiniCPM AVQA dev, run 1	$+0.50$	$[-11.60, 14.00]$	-1.50	0.00
MiniCPM strict Daily-Omni, run 1	$+1.04$	$[-39.64, 41.28]$	$+2.50$	1.04
MiniCPM AVQA dev, run 2	$+0.60$	$[-12.10, 14.60]$	-0.90	0.00
MiniCPM AVQA dev, run 3	$+0.40$	$[-12.20, 14.20]$	-1.30	0.00
MiniCPM strict Daily-Omni, run 2	$+3.20$	$[-37.82, 42.92]$	$+4.15$	2.25
MiniCPM strict Daily-Omni, run 3	$+1.90$	$[-39.29, 41.45]$	$+2.68$	1.90
Qwen3 AVQA dev	$+1.40$	$[-6.00, 7.90]$	0.00	0.00
Qwen3 strict Daily-Omni	-0.17	$[-40.76, 40.41]$	$+0.26$	0.00

D Stable-Path Analysis

Figure 6 evaluates the AV effect on two increasingly strict subsets: examples with unchanged A/V coverage and examples with unchanged parsed audio-only and video-only answers. Both retain the cross-domain reversal.

E Additional Diagnostic Experiments

E.1 Counterfactual modality replacement

We retain 942 questions with the same isolated path correct and the other wrong for the base and all five adapted Qwen checkpoints. A matched distractor comes from a different source, has a different semantic answer, and matches question type and relation. Input replacement has also been used to probe hallucination and modality interference (Dong et al., 2026; Cai et al., 2026a). In Table 8, self, distractor, and neutral denote joint accuracy with the original unsupported input, a matched

Table 7: Complete aggregate-only identification audit, part II. Effects and sharp-set endpoints are percentage points.

Comparison	ΔC	Aggregate-only set	$\Delta(\mathrm{AV}-V)$	H
Qwen2.5 QIVD, run 1	-0.98	$[-25.60, 22.37]$	$+0.11$	0.98
Qwen2.5 QIVD, run 2	-1.93	$[-26.44, 22.06]$	-0.21	1.65
Qwen2.5 QIVD, run 3	-1.51	$[-26.37, 22.13]$	-0.14	1.44
Qwen2.5 QIVD, run 4	-2.21	$[-26.86, 21.43]$	-0.84	0.88
Qwen2.5 QIVD, run 5	-1.75	$[-26.23, 21.74]$	-0.53	1.40
Qwen2.5 OmniVideo, run 1	-1.98	$[-33.66, 37.82]$	$+4.36$	0.00
Qwen2.5 OmniVideo, run 2	-2.38	$[-34.85, 37.82]$	$+3.17$	0.00
Qwen2.5 OmniVideo, run 3	-3.56	$[-35.25, 36.24]$	$+2.77$	0.00
Qwen2.5 OmniVideo, run 4	-0.40	$[-33.86, 40.20]$	$+2.77$	0.00
Qwen2.5 OmniVideo, run 5	-4.36	$[-37.03, 37.43]$	$+1.19$	1.19
MiniCPM OmniVideo, run 1	$+0.40$	$[-44.95, 48.71]$	0.00	0.40
MiniCPM OmniVideo, run 2	-1.78	$[-46.93, 47.72]$	-0.99	0.00
MiniCPM OmniVideo, run 3	$+1.39$	$[-42.77, 50.30]$	$+2.57$	1.39
Qwen2.5 AVQA public test	$+1.15$	$[-8.29, 12.55]$	$+0.41$	0.00
Qwen2.5-3B AVQA dev	$+0.70$	$[-12.30, 14.00]$	-0.30	0.00
Qwen2.5-3B strict Daily-Omni	-1.21	$[-49.83, 45.68]$	$+1.12$	0.00
VITA AVQA dev, run 1	-0.40	$[-15.00, 20.30]$	-0.70	0.40
VITA strict Daily-Omni, run 1	-0.86	$[-32.21, 32.90]$	$+0.35$	0.00
VITA AVQA dev, run 2	-0.30	$[-14.40, 20.90]$	-0.10	0.30
VITA strict Daily-Omni, run 2	-1.90	$[-33.16, 32.64]$	$+0.09$	0.00

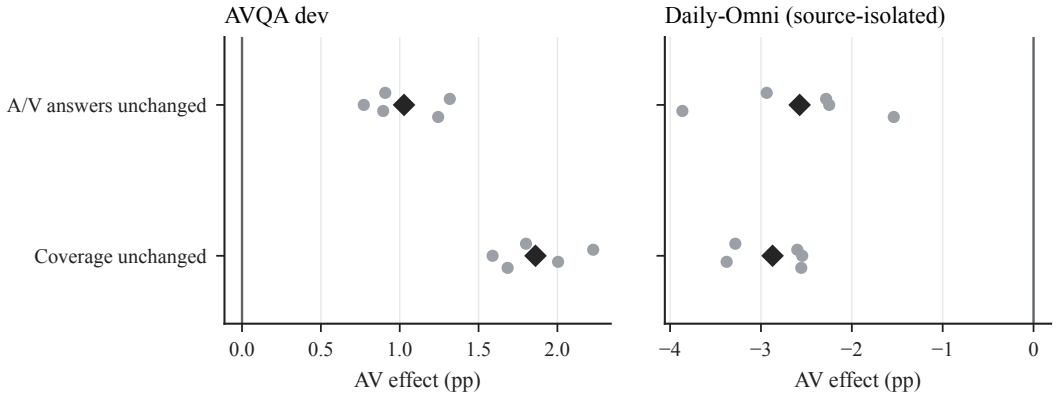


Figure 6: Joint-answer changes persist when isolated paths are unchanged. Circles show five independent Qwen fine-tuning runs; diamonds show their means. Both the in-domain improvement and external degradation remain.

replacement, and silent audio or black video in its place, respectively. Δ is the adapted-minus-base change in each accuracy difference. The distractor-versus-self effect changes sign across datasets, while the content-versus-neutral effect is negative on QIVD and positive on Daily. The response to unsupported modality content is therefore dataset-specific.

E.2 Prospective external test

OmniVideo-Test (Cai et al., 2026b) contains 505 manually reviewed questions from 264 sources with zero overlap with AVQA Both-5k. The prospective protocol specified rows, checkpoints, component predictions, and the decision rule before inference. The predicted Qwen conversion and MiniCPM coverage directions occur in every run; Qwen coverage instead averages -0.59 points.

Table 8: Mean counterfactual interference effects across five Qwen runs (percentage points).

Evaluation	Δ(distractor–self)	Δ(distractor–neutral)
AVQA	-2.08	$+4.42$
Strict Daily-Omni	$+7.04$	$+2.13$
QIVD	-2.16	-5.79
Pooled	-0.04	-3.14

Table 9: Prospective OmniVideo-Test effects in percentage points. Values are mean $\pm$ sample standard deviation across independent fine-tuning runs.

Model	ΔAV	ΔO	ΔC
Qwen2.5 (5 runs)	-3.13 ± 0.92	-0.59 ± 1.06	-2.53 ± 1.52
MiniCPM-o (3 runs)	-1.85 ± 0.89	-1.85 ± 0.75	$+0.00 \pm 1.62$

E.3 Controls for alternative explanations

Table 10 summarizes representative controls over answer representation, media budget, decoding, and source-external replay. MAD denotes modality-adaptive decoding (Chung et al., 2026); CAD denotes conflict-aware decoding (Deng et al., 2026).

Table 10: Representative controls. All changes are point estimates in percentage points.

Probe	Result
Option permutation	Strict Daily-Omni ΔC versus the base model: -4.23.
Semantic targets	Strict Daily-Omni ΔC versus the base model: -2.68.
Matched media budget	Strict Daily-Omni ΔC versus the base model: -3.80.
MAD/CAD decoding	On strict Daily-Omni, the base model's AV changes by -1.90 (MAD) and -2.16 (CAD) versus standard decoding.
Source-external replay	Candidate versus ordinary-replay AV: $+0.26$ on strict Daily-Omni and -0.40 on AVQA.

F Evaluation Paradigms

Table 11 compares JointAVBench (Chao et al., 2026), the source-robust oracle of Yang et al. (2021), and per-modality failure analysis (Bui et al., 2026) with our adaptation-effect analysis.

Table 11: Evaluation paradigms. Views: A/V/AV evaluations; paired items: within-example comparisons; pre/post: before–after adaptation comparisons. ✓/– indicate records included/absent from the analysis.

Paradigm	Primary target	Views	Paired items	Pre/post
JointAVBench	Joint reasoning	✓	–	–
Robust-fusion oracle	Source robustness	✓	–	–
Failure analysis	Modality interaction	✓	✓	–
Ours	**Adaptation-effect identification**	✓	✓	✓