SEEING WHAT SHOULD BE HEARD: DIAGNOSING AND REPAIRING CROSS-MODAL SHORTCUTS IN OMNI-MODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models (LLMs) are expected to answer a question using the modality it explicitly refers to. However, existing training paradigms rarely verify whether models actually follow this modality, because multimodal inputs from the same sample often provide redundant evidence for the same answer. In this work, we uncover a pervasive *cross-modal shortcut* in omni-modal LLMs: when asked an audio-related question, models rely on the image as much as on the audio, and sometimes even more. To systematically diagnose this behavior, we introduce the *Factorized Modality Diagnostic*, which independently swaps audio and images between samples to isolate each modality's causal contribution. Across two model families in different settings, we find that this shortcut persists throughout supervised fine-tuning and reinforcement learning post-training, while judge-based RL may further amplify such reliance on irrelevant visual information. Based on this finding, we propose *DMC-Repair*, which trains models on the same kind of cross-modal swapped samples while assigning supervision according to the modality specified by the question. This prevents models from exploiting the spurious correspondence between modalities within the same clip. Experiments demonstrate that DMC-Repair reduces the image-induced share of the answer effect by 59.9%, effectively suppressing the cross-modal shortcut without compromising audio-question answering performance. The reduction in shortcut reliance generalizes across two model families and zero-shot to an unseen dataset and an unseen benchmark, and persists through subsequent post-training. Code is available at `https://anonymous.4open.science/r/DMC-Repair`.

1 INTRODUCTION

When an omni-modal large language model (LLM) answers a question about the audio of a video, its answer should come from what it hears. Recent work post-trains such models with reinforcement learning (RL) (Yang et al., 2025; Zhao et al., 2025), which has improved reasoning in language models (Guo et al., 2025). These pipelines reward a correct answer and a well-formed response, and some also use a judge model to score the reasoning text. However, none of these rewards checks which modality the model actually uses to produce its answer. Because the image and the audio of a training sample usually support the same answer, a model that is asked which instrument is playing can identify the instrument from the image and still receive the full reward. We call this behavior a *cross-modal shortcut*. In the example of Figure 1, a model that sees a cello while a trumpet plays describes a cello sound and answers from the image.

As in this example, the shortcut often manifests as cross-modal hallucination, in which a model reports a sound only because its source is visible (Kim et al., 2025). Existing methods address this problem in three ways. Decoding methods contrast or reweight modality-specific branches at inference time (Jung et al., 2025; 2026; Chung et al., 2026). Preference optimization favors grounded responses over hallucinated ones (Baid et al., 2026; Chen et al., 2026a; Chaubey et al., 2026), and recent RL methods make the training objective modality-aware (Li et al., 2026a; Chen et al., 2026c; Xiao et al., 2026). However, these methods require extra decoding passes, preference pairs, or additional rewards. They are also evaluated by benchmark accuracy, and they rarely verify which modality an answer actually relies on.

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.

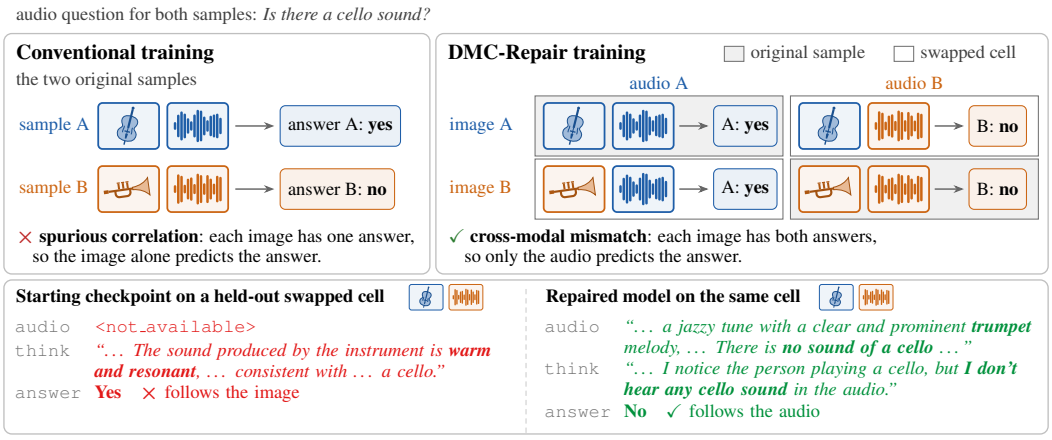


Figure 1: Conventional training and DMC-Repair training for an audio question. DMC-Repair trains on all four image–audio combinations of two samples and labels each cell by its audio. The bottom rows quote excerpts of the reasoning, with errors in red and audio evidence in green.

Such verification needs cases in which the image and the audio disagree, but natural data rarely contains them. We therefore create them by swapping the audio and the image between real samples and call the resulting test the Factorized Modality Diagnostic. For an audio question, the diagnostic takes two samples with different answers and forms a grid of the four combinations of their images and audio, which we call *cells*. In each cell, the *answer margin* is the gap between the log-probabilities of the two answers, and comparing the four margins separates the audio effect from the image effect. The *Shortcut Index* is the image's share of the answer effect and should be zero on audio questions. On existing omni-modal LLMs, however, the diagnostic reveals this shortcut. In two model families, the Shortcut Index is close to one half, so the image changes the answer to an audio question about as much as the audio does. The index remains at this level throughout supervised fine-tuning (SFT) and RL post-training, and judge-based RL rewards designed to encourage modality use further increase the image effect. As a result, on *conflict cells*, where the two modalities support different answers, the answer to an audio question often follows the image.

The same grid that reveals the shortcut also provides the training signal to suppress it. When a question designates the audio, we label each cell with the answer that its audio supports, so each answer appears in both original and swapped cells (Figure 1). A model therefore cannot infer the answer from the image or from whether the image and the audio originate from the same clip. We train on these *designated-modality counterfactuals* (DMC) and supervise only the answer tokens. We call this method DMC-Repair, and it requires no judge, reward model, or preference pairs.

We evaluate the repair with two metrics. The Shortcut Index shows how much the image drives the answer to an audio question, and *audio-following*, the rate at which generated answers on conflict cells follow the audio, shows whether the model relies on the audio. On held-out test questions, DMC-Repair reduces the Shortcut Index by 59.9% and improves audio-following by 13.7 points.

Contributions.

- We introduce the Factorized Modality Diagnostic, which isolates the effect of each modality on an answer, and reveal that the cross-modal shortcut persists across models and post-training stages.
- We propose DMC-Repair, which supervises only the answer tokens of counterfactual grids relabeled by the designated modality and suppresses the shortcut better than simpler constructions.
- Extensive experiments demonstrate that the repair is also effective on another model family, generalizes zero-shot to an unseen dataset and an unseen benchmark, and persists through subsequent RL post-training.

2 Related Work

Reinforcement learning for omni-modal reasoning. RL post-training has been applied across multimodal reasoning, from structured-context pipelines (Yang et al., 2025) to video and emotion

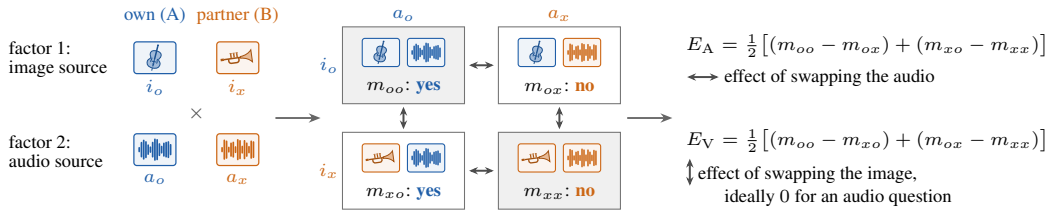


Figure 2: The 2×2 factorized design for an audio question. Crossing the image source and the audio source, each with two levels, own (A, blue) and partner (B, orange), gives four cells with answer margins m_{uv} for image u and audio v. Shaded cells contain one sample's image and audio.

reasoning (Feng et al., 2025; Zhao et al., 2025), generally using GRPO or its variants (Shao et al., 2024; Yu et al., 2025; Liu et al., 2025). These pipelines add a judged context term (Yang et al., 2025) or a frame-order term (Feng et al., 2025) to the answer and format rewards. None of these rewards verifies whether an answer follows the designated modality. Rouditchenko et al. (2025) report GRPO gains in audio QA that persist when the audio input is removed. Recent work adds modality-aware terms to the objective (Li et al., 2026a; Chen et al., 2026c; Xiao et al., 2026). We also add our own modality rewards to RL, and the diagnostic shows that none of them removes the shortcut (Section 5.5). DMC-Repair instead changes which input predicts the training label.

Diagnostics for modality dependence. Media interventions are a common diagnostic, from swap, mute, and shift interventions that expose failures in audio-visual reasoning (Wen et al., 2026) to modality shuffling that shows visual dominance (Fang et al., 2026). Our diagnostic complements them by measuring each modality's effect on the answer margin with a pair of real samples. It tests whether the non-designated modality has zero effect for audio and visual questions alike.

Counterfactual media as training signal. Beyond diagnosis, Wen et al. (2026) tune models on their interventions to verify audio-visual consistency and describe mismatches, and preference methods build preference pairs from perturbed media (Baid et al., 2026; Chaubey et al., 2026; Chen et al., 2026a). Closest to our setting, Chen et al. (2026b) test whether answers follow the requested modality on clips with swapped audio and fine-tune on aligned and misaligned clips to answer the category that the question names. Our construction completes each sample pair into a 2×2 grid for one question, so mismatch status carries no information about the answer. The same grid yields both the training cells and the diagnostic. Appendix M discusses these lines of work further.

3 Cross-Modal Shortcuts in Omni-Modal LLMs

3.1 Problem Setup

To formalize the shortcut, we consider an omni-modal model that receives N modality streams $M = (M_1, \ldots, M_N)$ together with a question Q and defines a distribution $p(A \mid Q, M)$ over answers. In many audio-visual QA benchmarks, each question is annotated with the subset of streams that determines its ground-truth answer. We write this *designated* subset as $d(Q)$ and its complement as $\bar{d}(Q)$. A model uses the designated modality when its answer depends only on $M_{d(Q)}$:

$$p(A \mid Q, M_{d(Q)}, M_{\bar{d}(Q)}) \;=\; p(A \mid Q, M_{d(Q)}, M'_{\bar{d}(Q)}) \quad \text{for every } M'_{\bar{d}(Q)}. \tag{1}$$

Here, $N = 2$, and $M = (i, a)$ consists of an image and an audio clip. Accuracy on a natural corpus cannot verify Equation 1, because $M_{d(Q)}$ and $M_{\bar{d}(Q)}$ are correlated there and a cross-modal shortcut incurs no penalty. We therefore break this correlation by swapping media between samples.

3.2 The Factorized Modality Diagnostic

The diagnostic applies to any audio-visual QA dataset in which several samples share a question. We call a question *Audio*, *Visual*, or *Audio-Visual* according to whether $d(Q)$ is the audio, the image, or both. For a question q we select two real samples with different ground-truth answers. The *own* item has media (i_o, a_o) and answer g, and the *partner* item has media (i_x, a_x) and answer g'. The design has two factors, the source of the image and the source of the audio, each with two levels, own or

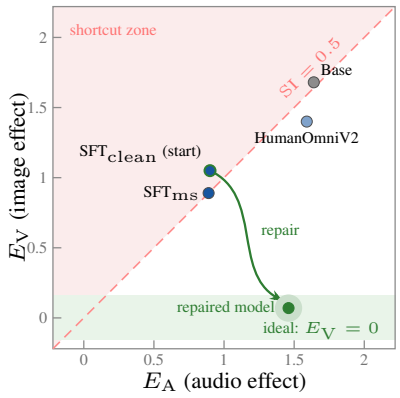


Figure 3: Audio and image effects on Audio questions under the native prompt, which retains the training system prompt and output format (65 development families).

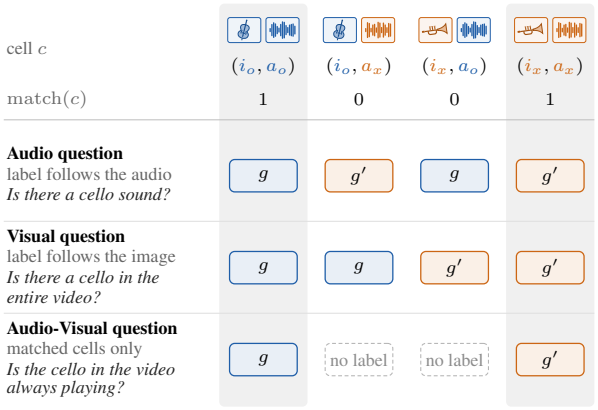


Figure 4: Labels of the four cells of a pair under Equation 4, with an example question per type in italics. Blue g is the own sample's answer and orange g' the partner's. Shaded columns are the matched cells ($\mathrm{match}(c) = 1$), which contain one sample's image and audio.

partner (Figure 2). We evaluate all four combinations, which we call *cells*, so the two factors are fully crossed as in a classical 2×2 factorial experiment (Fisher, 1935). The cells (i_o, a_o) and (i_x, a_x) contain a sample's own image and audio and are *matched*, and the other two are *mismatched*. For an Audio question the correct answer for every cell is the one that its audio supports, so swapping the image should leave the preferred answer unchanged and swapping the audio should reverse it. In cell (i_u, a_v), the teacher-forced answer margin $m_{uv} = \log p(g \mid q, i_u, a_v) - \log p(g' \mid q, i_u, a_v)$ compares the two answers under direct answer scoring, using the prompt followed by `<answer>`. Its sign indicates which answer the model prefers. The audio and image main effects are

$$E_{\mathrm{A}} = \tfrac{1}{2}\big[(m_{oo} - m_{ox}) + (m_{xo} - m_{xx})\big], \quad E_{\mathrm{V}} = \tfrac{1}{2}\big[(m_{oo} - m_{xo}) + (m_{ox} - m_{xx})\big], \qquad (2)$$

with interaction $\psi = \frac{1}{2}(m_{oo} - m_{ox} - m_{xo} + m_{xx})$ (Appendix D). Each main effect averages the change from swapping one modality over both levels of the other, and ψ measures how much that change depends on the other modality. Swapping or removing one modality at a time would measure its effect at a single level of the other only, and could not separate a main effect from an interaction. Equation 1 applied to an Audio question requires $E_{\mathrm{V}} = 0$, so we summarize each configuration by the *Shortcut Index* $\mathrm{SI} = |E_{\mathrm{V}}|/(|E_{\mathrm{A}}| + |E_{\mathrm{V}}|)$, which is zero for a model that ignores the image and one half when the two modalities contribute equally. On Visual questions the two modalities swap roles, and the ideal index is one. Unlike accuracy, the index exposes a violation of Equation 1.

A *question family* groups the samples that share one question. We draw two sample pairs from each family and orient each pair so that the own item has the answer `yes`. We compute SI from the main effects averaged over a family's two pairs and report its mean over families.

3.3 Pilot Study

We first run the diagnostic as a pilot study on a development set of MUSIC-AVQA (Li et al., 2022), with checkpoints from two training routes. HumanOmniV2 (Yang et al., 2025) adds SFT and GRPO post-training to Qwen2.5-Omni (Xu et al., 2025), and we diagnose its released checkpoint. Our repair route starts from $\mathrm{SFT}_{\mathrm{clean}}$, an SFT checkpoint trained on HumanOmniV2's cold-start data with separate audio and visual descriptions, and we call it the *starting checkpoint*. We apply DMC-Repair to $\mathrm{SFT}_{\mathrm{clean}}$ to obtain the *repaired model* and continue with GRPO to obtain the *RL endpoint* (Section 5.5). We also diagnose $\mathrm{SFT}_{\mathrm{ms}}$, which uses an earlier version of these descriptions.

The pilot study finds the same cross-modal shortcut in every checkpoint of both model families, including MiniCPM-o-2.6 (Yao et al., 2025; OpenBMB, 2025). Regardless of the prompt and the post-training stage, the image changes the answer about as much as the audio does (Figure 3 and Table 4 in Appendix B). In free generation, the starting checkpoint follows the audio on fewer than half of the conflict cells. Validity checks confirm that the diagnostic measures modality use rather

than noise (Appendix B). The tested SFT and RL post-training stages do not teach a model which modality a question refers to. On natural samples, an accuracy reward cannot resolve this, because an image-derived answer receives the same reward as an audio-derived one.

4 DMC-REPAIR

The pilot study shows that the model needs training examples in which the image and the audio disagree and the label follows the designated modality. Our key idea is to build these examples from the same factorized grid that the diagnostic uses. DMC-Repair consists of three components. Counterfactual grid construction creates the conflict between the modalities (Section 4.1), designated-modality relabeling makes the designated modality the only predictor of the label (Section 4.2), and answer-token supervision optimizes the answer margin that the diagnostic measures (Section 4.3). The recipe requires only the question types and answers of a dataset (Section 4.4).

4.1 COUNTERFACTUAL GRID CONSTRUCTION

Let $\mathcal{D}$ be a training corpus of real samples (q_n, i_n, a_n, y_n), where q_n is the question, i_n the image, a_n the audio clip, and y_n the ground-truth answer. Training first requires inputs in which the two modalities support different answers, which are rare in natural samples. Masking or dropping one modality does not suffice, because a missing input does not contradict the other. We instead pair samples of $\mathcal{D}$ that share a question but have different answers, and collect these pairs in $\mathcal{P}$. Each pair $\{o, x\} \in \mathcal{P}$ yields a grid of four cells, one for each image–audio combination,

$$\mathcal{G}(o, x) = \big\{(q, i_u, a_v) \;:\; u, v \in \{o, x\}\big\}. \tag{3}$$

The two cells with $u = v$ are the original samples, and the two *counterfactual* cells with $u \neq v$ combine the image of one sample with the audio of the other (Figure 1, right). The grid contains only real media and both swap directions, so every image and audio clip of a pair appears in a matched and a mismatched cell. Because the two samples have different answers, the image and the audio of the two counterfactual cells disagree, for example a cello on screen while a trumpet plays.

4.2 DESIGNATED-MODALITY RELABELING

Each cell of the grid then needs a label. A mismatched cell could be labeled as a mismatch, which trains a model to detect inconsistent media (Wen et al., 2026). Such a label depends on whether the media match and not on the designated modality. We instead assign each cell the answer of the sample that supplied its designated modality, so for a cell $c = (q, i_u, a_v)$ the label is

$$y^\star(c) = \begin{cases} y_v & \text{if } d(q) = \{\text{audio}\}, \\ y_u & \text{if } d(q) = \{\text{image}\}, \\ y_u & \text{if } d(q) = \{\text{image}, \text{audio}\} \text{ and } u = v. \end{cases} \tag{4}$$

For a question designating both modalities, a counterfactual cell has no known label, so we retain only the two original cells. Figure 4 shows the labels for a pair with answers $g = y_o$ and $g' = y_x$.

With these labels, the designated input is the only input that predicts the label. In a complete Audio or Visual grid, each answer labels two cells. These two cells contain different non-designated inputs, and one is matched while the other is mismatched. Formally, let $\mathrm{match}(c) = \mathbb{I}[u = v]$ indicate whether a cell is matched, and let $c_{\bar{d}}$ denote its non-designated input, which is the image for an Audio question and the audio for a Visual question. For a cell c drawn uniformly from the grid,

$$\Pr\big[y^\star(c) = g \mid c_{\bar{d}}\big] \;=\; \Pr\big[y^\star(c) = g \mid \mathrm{match}(c)\big] \;=\; \tfrac{1}{2}. \tag{5}$$

Neither the non-designated input nor the match indicator alone predicts the label better than chance, whereas the designated input determines it through Equation 4. In the original samples alone, the two inputs always occur together, so both predict the label equally well, which is the origin of the shortcut. In joint training on Audio and Visual grids, the same image determines the label under a Visual question and is uninformative under an Audio question. The model must therefore infer from the question which modality to use, rather than acquire a fixed preference.

4.3 Answer-Token Supervision

Relabeling gives each counterfactual cell a target answer but no reference response, so standard supervised fine-tuning on complete responses is not applicable. The pilot study also shows that RL rewards on the generated text do not remove the shortcut. DMC-Repair therefore supervises the answer directly and trains the model to prefer the label of each cell over the other answer of its pair. For a cell c and a candidate answer $y = (y_1, \ldots, y_{|y|})$, we teacher-force the prompt followed by `<answer>` and score the answer by its length-normalized log-likelihood,

$$s_\theta(y \mid c) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log p_\theta\big(y_t \mid c, \texttt{<answer>}, y_{<t}\big). \tag{6}$$

Let $\bar{y}(c)$ be the other answer of the pair. The loss of a cell is a hinge on the score gap with threshold τ, and DMC-Repair minimizes its average over the set $\mathcal{C}$ of all labeled cells,

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \max\Big(0,\ \tau - \big[s_\theta(y^\star(c) \mid c) - s_\theta(\bar{y}(c) \mid c)\big]\Big). \tag{7}$$

For one-token answers, the score gap is the answer margin of Section 3.2 oriented toward the label, so the objective optimizes the quantity that the diagnostic measures. The hinge is zero once this margin reaches τ, so only cells whose margin remains below τ contribute gradients. Since the loss covers only the answer tokens, it sets no target for the generated description or reasoning, and training requires no rollouts, reward model, or preference pairs.

4.4 The Recipe in General

The construction extends beyond images and audio to any dataset that provides a designation $d(Q)$ for every question and real samples that share a question but differ in their answers. With N modality streams, we take each stream k from a source sample s_k to form a cell $c_{\mathbf{s}} = (Q, M_1^{(s_1)}, \ldots, M_N^{(s_N)})$. Its label is the answer of the sample that supplied all designated streams,

$$y^\star(c_{\mathbf{s}}) = y_{s_k} \ \text{ for } k \in d(Q), \qquad \text{defined only if } s_k = s_{k'} \text{ for all } k, k' \in d(Q). \tag{8}$$

The label depends only on the designated streams, as Equation 1 requires. We next measure how closely models trained on these labels meet this requirement.

5 Experiments

We organize the experiments around four questions. **Q1:** Does DMC-Repair suppress the cross-modal shortcut on held-out questions and in another model family (Section 5.2)? **Q2:** Which parts of the grid construction produce the gain, and do the loss and parameterization matter (Section 5.3)? **Q3:** Does the suppression generalize zero-shot to unseen data, and how does it compare with published methods (Section 5.4)? **Q4:** Can a modality reward in RL achieve the same effect, and does the suppression persist through subsequent RL without degrading other abilities (Section 5.5)?

5.1 Datasets and Metrics

Datasets. Before any measurement, we split the 387 MUSIC-AVQA diagnostic families (Appendix A) into a development set of 130 families (65 Audio, 27 Visual, and 38 Audio-Visual), which serves as the validation set to evaluate all design choices, and a confirmation set of 257 families, which serves as the test set and remains unseen until all checkpoints are frozen. After excluding every diagnostic family and its media, we build 6,176 training cells from 2,036 MUSIC-AVQA pairs in 643 question families. AVQA (Yang et al., 2022) with 104 families and AVHBench (Kim et al., 2025) with 5,302 questions are used to evaluate the zero-shot generalization ability, and IntentBench (Yang et al., 2025) with 2,689 questions is used to evaluate the general omni-modal reasoning ability.

Metrics. Under direct scoring, we report the *Shortcut Index* $\mathrm{SI} = |E_\mathrm{V}|/(|E_\mathrm{A}|+|E_\mathrm{V}|)$ of Section 3.2, the image's share of the answer effect, which is ideally zero on Audio questions and one on Visual questions. We also report the non-designated effect, which is E_V on Audio questions and E_A on

Table 1: Repair results on held-out development families. The confirmation-set row uses the 128 Audio families of the sealed confirmation set. Full and LoRA (Hu et al., 2022) both start from $\mathrm{SFT}_{\mathrm{clean}}$ and use the same cells (Appendix A). Full (bold) updates all LLM parameters, and LoRA trains a rank-16 adapter. The released HumanOmniV2 checkpoint (Yang et al., 2025) is evaluated by direct scoring only. Δ is the change of Full from $\mathrm{SFT}_{\mathrm{clean}}$ with its paired 95% confidence interval (CI). In the own-reasoning block, SI is rescored after the model's own reasoning, and the other two rows use free generation with unparseable outputs counted as failures (Appendix L).

		HumanOmniV2	$\mathrm{SFT}_{\mathrm{clean}}$	LoRA	Full	Δ (Full)	[95% CI]
Audio q., direct	SI (ideal = 0)	0.498	0.549	0.194	**0.187**	-0.362	$[-0.43, -0.29]$
	E_{V} (non-designated)	$+1.40$	$+1.05$	-0.04	$+\mathbf{0.07}$	-0.98	$[-1.39, -0.59]$
	SI, confirmation set	—	0.559	0.208	**0.224**	-0.335	$[-0.39, -0.28]$
Visual q., direct	SI (ideal = 1)	0.716	0.713	0.872	**0.872**	$+0.159$	$[+0.09, +0.22]$
	E_{A} (non-designated)	$+1.86$	$+1.28$	-0.10	$+\mathbf{0.02}$	-1.26	$[-1.65, -0.87]$
Audio q., own reasoning	SI (ideal = 0)	—	0.612	0.256	**0.271**	-0.341	$[-0.42, -0.27]$
	Audio-Following, %	—	43.1	60.8	**61.2**	$+18.1$	$[+10.0, +26.2]$
	Matched-Cell Accuracy, %	—	61.5	60.0	**67.3**	$+5.8$	$[-0.4, +11.9]$

Visual questions. Under free generation, we report *Audio-Following*, the rate at which answers on conflict cells follow the audio, and *Matched-Cell Accuracy*. Unless stated otherwise, intervals in brackets are 95% family-cluster bootstrap confidence intervals (CIs) over 10,000 resamples, paired between checkpoints. Appendices A and L give the training and statistical details and all prompts.

5.2 Main Results

Results on MUSIC-AVQA. DMC-Repair shifts answers to each question's designated modality (Table 1). On Audio questions, it reduces the Shortcut Index by two thirds, whereas the released HumanOmniV2 remains near one half. The repair lowers the image effect for either audio clip and brings it close to zero (Appendix D). On Visual questions, the index moves toward its ideal value of one, and the audio effect also falls close to zero. The model thus learns which modality each question requires rather than suppressing one. The index reduction persists when answers are rescored after the model's own reasoning. In free generation, 84 of the 260 conflict-cell answers switch toward the audio and 37 away from it. Matched-cell accuracy on Audio questions rises as well (Table 1 and Appendix E). The gain extends to generated answers without reducing Audio-question accuracy.

Results on the confirmation set. On the sealed confirmation set, the repair reduces the Shortcut Index by 59.9% (Table 1) and improves audio-following by 13.7 points. All seven pre-registered tests meet their criteria (Appendix F), so the gain is not an artifact of tuning on the development set.

Cross-model generalization. The repair is also effective on MiniCPM-o-2.6 (Yao et al., 2025; OpenBMB, 2025), whose encoders and output format differ from Qwen2.5-Omni's. LoRA training on the same cells reduces its Shortcut Index by about half on both sets under its own answer scoring and improves its zero-shot AVHBench accuracy by 1.4 points $[+0.4, +2.3]$ (Appendices K and H). In both model families, DMC-Repair suppresses the cross-modal shortcut on held-out questions.

5.3 Ablation Study

Construction controls. We next ask which elements of the construction produce the gain. Three pre-registered controls each remove one element (Figure 5) and show that the gain requires both conflicting inputs and designated-modality labels. Modality dropout retains the labels but removes the conflicts, and mismatch labels retain the conflicts but teach the model to report a mismatch. Both leave the index at about twice that of the complete grid. Single-direction cells cover one swap direction per pair and retain about 80% of the index reduction. In free generation, the complete grid leads every control by 12 to 31 points under three treatments of unparseable outputs (Appendix F). On the development set, training only on the original cells yields about a third of the index reduction without improving audio-following (Appendix J). Every element of the construction matters.

Loss and parameterization. The loss and parameterization matter much less. Cross-entropy achieves 89% of the reduction obtained with the hinge loss, and LoRA matches full fine-tuning

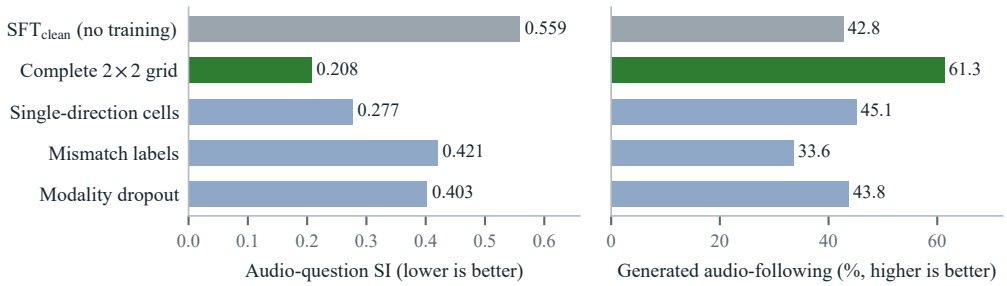


Figure 5: Construction controls on the confirmation set (128 Audio families, 510 conflict cells). The complete grid is the LoRA reference, and paired differences with 95% CIs are in Appendix F.

Table 2: Comparison with published methods on AVHBench, all built on Qwen2.5-Omni. Base is the audio-hallucination accuracy that each paper reports for its base model, and each change is measured against that base model on the same task. Other numbers are copied from the papers (—, not reported), and Table 14 gives all tasks for our models. †Evaluated by Chung et al. (2026).

Method	Training data	Base	Δ Audio hall.	Δ AV matching
AVCD† (Jung et al., 2025)	none (decoding)	73.0	+2.8	—
MAD (Chung et al., 2026)	none (decoding)	73.0	+5.7	—
ACPO (Baid et al., 2026)	audio-swap preferences	66.7	+2.6	—
OmniDPO (Chen et al., 2026a)	audio-visual preferences	67.6	+9.9	—
MoD-DPO++ (Chaubey et al., 2026)	modality-perturbed preferences	77.4	+6.0	+15.0
Chen et al. (2026b)	(mis)aligned AudioSet clips	71.7	+8.2	+0.0
DMC-Repair (full)	MUSIC-AVQA grid cells	61.9	+5.8	+7.2
+ subsequent RL	+ GRPO	61.9	**+11.3**	+7.3

with 148 times fewer trainable parameters. The index falls by 0.30 to 0.36 for all four seeds (Appendix J). The gain originates from the training cells rather than the training settings.

5.4 Zero-Shot Generalization Performance

Since all training cells are drawn from MUSIC-AVQA, we next apply the repaired model directly to an unseen dataset and an unseen benchmark. In this zero-shot setting, the checkpoints are frozen before evaluation, and neither target is used for training, model selection, or prompt tuning.

Generalization to an unseen dataset. DMC-Repair generalizes zero-shot to AVQA (Yang et al., 2022), an everyday audio-visual QA dataset built on VGGSound (Chen et al., 2020a). The selected questions, taken from the AVQA items of OmniInstruct (Li et al., 2025), ask whether an object is the main sound source. On these questions, all three pre-registered criteria are met (Appendix G). The repair reduces the image effect by 62% and improves audio-following by 18 points (as on MUSIC-AVQA), and matched-cell accuracy rises by 14 points. This suggests that the model learned a rule for selecting the modality rather than knowledge about particular instruments or scenes.

Generalization to a hallucination benchmark. Suppressing the shortcut also improves zero-shot accuracy on the audio-hallucination and audio-visual matching tasks of AVHBench (Kim et al., 2025), the two tasks whose answers depend on the audio. Video-hallucination accuracy is nearly unchanged (Appendix H). Without any hallucination-specific training data, the repair achieves an audio-hallucination gain over its base that is comparable to that of the strongest published decoding method. Subsequent RL nearly doubles this gain, which is then the largest in Table 2. Most of the audio gain is due to audible objects that the starting checkpoint missed. Accuracy on audible objects rises from 49.1% to 61.7% after repair and to 75.4% after RL. The suppression generalizes beyond the training data and improves the abilities that require listening.

5.5 Analysis

Modality rewards in RL. Adding a modality reward to RL in place of the repair does not remove the shortcut (Appendix C). Four of our six RL variants use a judge to score the model's own description

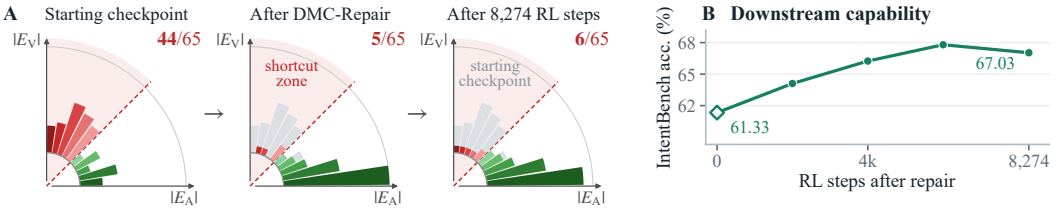


Figure 6: Column area is proportional to the number of Audio families per 9° angle of $(|E_A|, |E_V|)$.

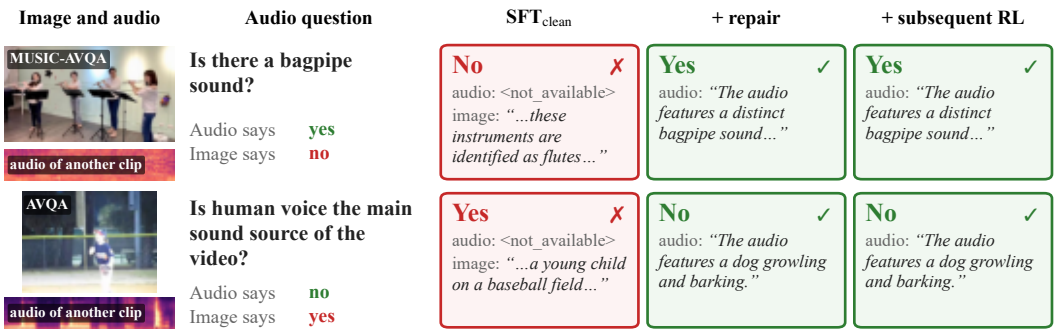


Figure 7: Generated answers on two conflict cells, colored by the modality that they follow.

of the input, and the three judge rewards evaluated on all families increase the image effect. GRPO with only format and accuracy rewards increases the image effect about twice as much as the audio effect. Mixing grid cells into the RL data lowers the index by less than a tenth of the repair's effect.

Robustness to subsequent RL. When GRPO follows the repair, the suppression persists through all 8,274 steps, over which IntentBench accuracy improves (Figure 6 and Appendix F). Of the 44 development Audio families in the shortcut zone (SI > 0.5) before the repair, 41 leave and 2 others enter, and 6 of the 65 lie in it at the RL endpoint. On the confirmation set, the index is unchanged and audio-following rises by another 9 points, which meets both pre-registered retention criteria.

Other abilities. The repair itself has little effect on other abilities. Before subsequent RL, IntentBench accuracy remains within one point of the starting checkpoint (61.33% versus 60.64%), and accuracy on Audio-Visual questions, for which we retain only the two original cells, does not decrease (Appendix E). Overall, RL does not replace the repair but can follow it without undoing it.

5.6 Case Study

Finally, Figures 1 and 7 show how the repair changes the generated text, and Appendix I adds eight more cells, including two on which all three checkpoints follow the image. The starting checkpoint leaves its audio description empty and bases its answer on the image. In contrast, the repaired model and the RL endpoint describe the audio content, namely the bagpipe in row 1 of Figure 7 and the barking dog in row 2 from the unseen AVQA, and base their answers on the audio.

6 Conclusion

Omni-modal LLMs should answer audio questions from the audio, but their training data and rewards rarely separate the image from the audio. By swapping the two between real samples, our Factorized Modality Diagnostic shows that the tested models rely on the image about as much as on the audio, at every stage of SFT and RL. To suppress this cross-modal shortcut, DMC-Repair recombines real samples into grids, labels each cell by its designated modality, and supervises only the answer tokens. The repair reduces the Shortcut Index by 59.9% on a sealed confirmation set, generalizes across two model families and zero-shot to an unseen dataset and benchmark, and persists through subsequent RL that improves task accuracy. Future work can extend the relabeling rule of Section 4.4 to questions that designate several modalities and to open-ended answers.

AI USE STATEMENT

We used generative AI coding and writing assistants to implement author-specified methods, clean and reformat datasets, review conceptual framing and experimental design, assess pre-registered hypotheses and criteria, and assist with result interpretation. The one model-assisted change to training data was the split of the cold-start context annotation into separate audio and visual descriptions, where the model could only assign sentences of the original annotation and a coverage check rejected any split that dropped content. All research decisions were made by the authors. We did not use generative AI to synthesize datasets or labels or to formulate or prove mathematical claims. Translation and qualitative or thematic data analysis are not applicable to this work.

AI tools also assisted with drafting and editing the paper, drawing and editing figures, creating and editing code, finding and summarizing related work, and formatting references. The authors reviewed all AI-assisted work. Code was reviewed and tested, analysis scripts were checked against independent computations, and reported numbers were recomputed from raw outputs. Confirmation results were computed with the frozen analysis script described in Appendix F. Text was checked against the results and citations against the original publications. The authors take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

The study uses publicly released benchmarks and involves no new human-subject data collection. MUSIC-AVQA is used under its release license. We are not aware of ethical concerns specific to this evaluation and training-data construction study.

REPRODUCIBILITY STATEMENT

Section 3.2 defines the diagnostic and its statistics, and Section 4 specifies how training cells are built and supervised. Appendix A gives training and evaluation settings, and Appendix F gives the pre-registered confirmation protocol and its criteria. The diagnostic sets use MUSIC-AVQA and AVQA, the training cells are built from MUSIC-AVQA, and the external evaluations use IntentBench and AVHBench, all of which are publicly released. The core scripts for the diagnostic, cell construction, training, and analysis are available at the anonymous repository linked in the abstract, together with the family lists of the diagnostic sets, and the complete code will be released upon acceptance.

REFERENCES

Ami Baid, Zihui Xue, and Kristen Grauman. Don't let the video speak: Audio-contrastive preference optimization for audio-visual language models. In *ECCV*, 2026.

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. MoD-DPO: Towards mitigating cross-modal hallucinations in omni LLMs using modality decoupled preference optimization. In *CVPR*, 2026.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *ICASSP*, 2020a.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Chao Sun, Rongzhou Zhang, Guanyu Zhou, and Lijie Wen. OmniDPO: A preference optimization framework to address omni-modal hallucination. In *AAAI*, 2026a.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, 2020b.

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. In *CVPR Findings*, 2026b.

Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, Biao Wang, Qinglin Lu, and Ruqi Huang. OmniVideo-R1: Reinforcing audio-visual reasoning with query intention and modality attention. In *ICML*, 2026c.

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 *CVPR*, 2026.

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

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *NeurIPS*, 2025.

Ronald A. Fisher. *The Design of Experiments*. Oliver & Boyd, 1935.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.

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 *ICLR*, 2022.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. AVCD: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In *NeurIPS*, 2025.

Chaeyoung Jung, Youngjoon Jang, Jongmin Choi, and Joon Son Chung. Fork-merge decoding: Enhancing multimodal understanding in audio-visual large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026.

Sung-Bin Kim, Hyun-Bin Oh, Jung-Mok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. AVHBench: A cross-modal hallucination benchmark for audio-visual large language models. In *ICLR*, 2025.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, 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*, 2026a.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan Wen, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. OmniBench: Towards the future of universal omni-language models. In *NeurIPS Datasets and Benchmarks Track*, 2025.

Zongxia Li, Wenhao Yu, Zhenwen Liang, Chengsong Huang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. Vision-SR1: Self-rewarding vision-language model via reasoning decomposition and multi-reward policy optimization. In *ICLR*, 2026b.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *COLM*, 2025.

Jie Ma, Min Hu, Pinghui Wang, Wangchun Sun, Lingyun Song, Hongbin Pei, Jun Liu, and Youtian Du. Look, listen, and answer: Overcoming biases for audio-visual question answering. In *NeurIPS*, 2024.

OpenBMB. MiniCPM-o 2.6: A GPT-4o level MLLM for vision, speech and multimodal live streaming on your phone. `https://github.com/OpenBMB/MiniCPM-o`, 2025.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-R1: Do you really need audio to fine-tune your audio LLM? In *ASRU*, 2025.

Ramaneswaran Selvakumar, Kaousheik Jayakumar, S Sakshi, Sreyan Ghosh, Ruohan Gao, and Dinesh Manocha. Do audio-visual large language models really see and hear? In *CVPR Findings*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xiaofei Wen, Wenjie Jacky Mo, Xingyu Fu, Rui Cai, Tinghui Zhu, Wendi Li, Yanan Xie, Muhao Chen, and Peng Qi. When vision speaks for sound. *arXiv preprint arXiv:2605.16403*, 2026.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-R1: Mitigating shortcuts in visual reasoning with reinforcement learning. *Transactions on Machine Learning Research*, 2026.

Cihan Xiao, Yiwen Shao, Chenxing Li, Xiang He, Zhenwen Liang, Steve Yves, Sanjeev Khudanpur, and Liefeng Bo. Escape the language prior: Mitigating late-stage modality collapse in audio reasoning via modality-aware policy optimization. *arXiv preprint arXiv:2605.27741*, 2026.

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*, 2025.

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 *ACM MM*, 2022.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. HumanOmniV2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Ronghua Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Efficient GPT-4V level multimodal large language model for deployment on edge devices. *Nature Communications*, 16:5509, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *NeurIPS*, 2025.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-Omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025.

Overview of the Appendices

- **Setup.** Appendix A gives the training and evaluation details, and Appendix L lists the prompts.
- **Diagnostic.** Appendix B extends the pilot study to every configuration, Appendix D examines the interaction term, and Appendix E breaks the results down by question type.
- **Repair.** Appendix C shows that none of the six RL variants of Section 5.5 removes the shortcut, Appendix F reports the pre-registered confirmation tests, Appendix J ablates the loss and the parameterization, and Appendix K gives the cross-model diagnostic on MiniCPM-o-2.6.
- **Zero-shot generalization.** Appendices G and H give the AVQA and AVHBench protocols.
- **Case study and related work.** Appendix I adds eight conflict cells to the case study, and the related work continues in Appendix M.

A Training and Evaluation Details

This appendix gives the settings behind Sections 3.3 and 5 and traces how the two diagnostic sets are built.

SFT. $\mathrm{SFT}_{\mathrm{clean}}$ is obtained by training the LLM backbone for two epochs with a learning rate of 2×10^{-5}, a batch size of 1 per GPU, and no gradient accumulation. The vision and audio encoders remain frozen. All training runs use four NVIDIA A800-SXM4-80GB GPUs.

HumanOmniV2. We adopt HumanOmniV2 (Yang et al., 2025) as the baseline model because it is an open omni-modal reasoning model obtained from Qwen2.5-Omni by SFT and GRPO post-training, and our starting checkpoint is trained on its cold-start data.

Repair recipe. The fully fine-tuned model (Full) is the repaired model used for confirmation and subsequent RL, and the LoRA model is the reference for the construction controls. The training pool is built from MUSIC-AVQA after every question family and every media file of the diagnostic sets has been removed, so that no training cell shares content with the evaluation. It comprises 461 Audio pairs and 591 Visual pairs with four cells each and 984 Audio-Visual pairs with two cells each (Section 4.2), and its answers take 42 distinct values. In the LoRA recipe, rank-16 adapters (α=32, dropout 0.05) are trained on the LLM and both encoders for three epochs over the 6,176-cell pool, which takes about four hours. This recipe uses a learning rate of 1×10^{-4}, gradient accumulation over 4 steps, and bf16 precision. The hinge threshold is $\tau = 1.0$ over length-normalized candidate scores, and the base weights of the vision and audio encoders remain frozen. Full fine-tuning follows the same schedule with a learning rate of 1×10^{-5} and AdamW with bf16 optimizer states, and it takes about three hours. The LoRA adapters are merged before evaluation.

Construction controls. The controls use the LoRA recipe of the complete-pool reference, with changes to the training cells described in Section 5.3. The single-direction control contains 5,124 cells and is trained for 4 epochs, which roughly matches the number of cells that the complete pool provides in 3 epochs. The other two controls each use 6,176 cells for 3 epochs.

Subsequent RL. GRPO training starts from the repaired model and runs for the configured budget of 8,274 steps on the RL training data of HumanOmniV2, filtered as described below. It uses 2 rollouts per prompt, a learning rate of 5×10^{-6}, a maximum sequence length of 2,048, and gradient checkpointing. As in HumanOmniV2, the KL coefficient decays linearly from 0.04 to 0.01 over the first half of training. Rollout filtering retains prompts with accuracy in $(0, 0.75]$.

Cross-model adaptation. MiniCPM-o-2.6 is evaluated with a neutral prompt, and its teacher-forced yes/no margin is read at the first generated token position as $m = \log p(\texttt{yes}) - \log p(\texttt{no})$. The LoRA adapters are injected only into the LLM through the `inject_adapter_in_model` interface of the PEFT library with the same rank, α, and schedule as the Qwen recipe.

Statistics. Confidence intervals are 95% family-cluster bootstrap intervals over 10,000 resamples unless stated otherwise.

Diagnostic set construction. Table 3 traces both diagnostic sets from the source annotations to the retained families. Among the candidate families, almost all removals result from the requirement of at least two samples for each answer, and the media-uniqueness constraint of Appendix G removes only one AVQA family.

Table 3: Construction of the two diagnostic sets, from source annotations to retained families. The first two data rows count annotation rows, and the next five rows count families. Samples that share a question form a family, and we retain only families with at least two `yes` and two `no` samples. One Audio pair of the MUSIC-AVQA confirmation set was later dropped for missing media, giving 773 pairs rather than 774. The media-uniqueness constraint of Appendix G leaves six AVQA families with a single pair, giving 202 pairs rather than 208.

	MUSIC-AVQA	AVQA
Source annotation rows	42,492	6,402
usable answer format	19,478	2,343
Candidate families	2,521	846
– fewer than two `yes` or two `no`	−2,134	−741
– media already claimed by another family	—	−1
– overlapping the frozen evaluation manifest	−0	—
Retained families	**387**	**104**
Pairs / cells	773 / 3,092	202 / 808
Question types	193 Audio, 80 Visual, 114 A-V	104 Audio

B THE DIAGNOSTIC ACROSS MODEL FAMILIES AND PROMPT FORMATS

This appendix shows that the shortcut appears in every tested configuration and that the diagnostic measures modality use (Section 3.3). Figure 3 shows only the Qwen checkpoints under the native prompt, so that the change produced by the repair is easy to follow. Figure 8 adds the neutral-prompt configurations and MiniCPM-o-2.6, whose repair is the exploratory cross-model generalization experiment of Section 5.2 (Appendix K). Table 4 lists the pilot-study diagnostic for every configuration other than $\text{SFT}_{\text{clean}}$, which Table 1 reports. Every baseline configuration lies near the $\text{SI} = 0.5$ diagonal regardless of model family or prompt, and both repairs move toward $E_{\text{V}} = 0$. Under the native prompt, HumanOmniV2 lies within the paired equivalence margin $|\Delta\text{SI}| < 0.1$ of its base model (Table 4). The shortcut thus predates the post-training of HumanOmniV2 and persists through it.

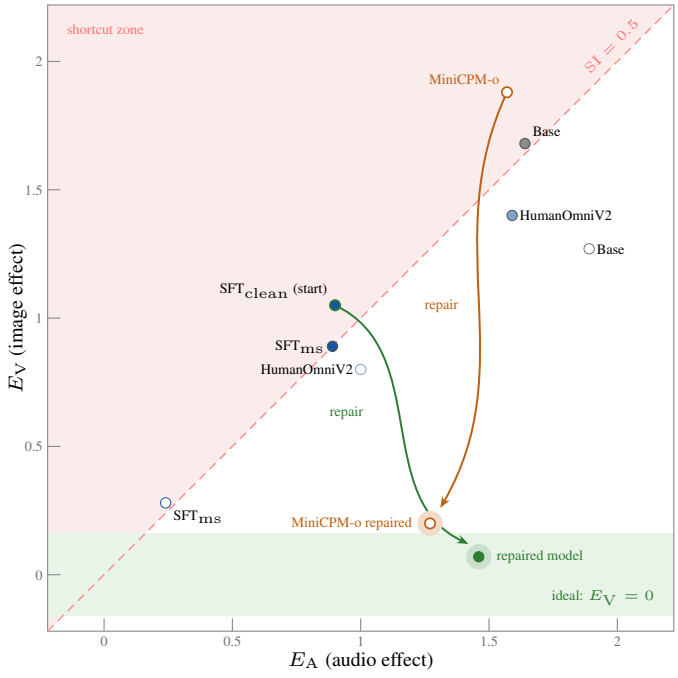


Figure 8: The diagnostic of Figure 3 with all seven configurations of Table 4 and both repairs. Filled markers are native-prompt runs and open markers neutral-prompt runs. The MiniCPM-o arrow starts at its base model, since no SFT stage precedes it.

Table 4: Audio-question diagnostic on 65 development families, where the ideal SI is 0. Base is Qwen2.5-Omni, HumanOmniV2 is the released checkpoint of Yang et al. (2025), and SFT_{ms} is an earlier version of $\text{SFT}_{\text{clean}}$. The neutral prompt omits the system prompt and output format. The paired native-prompt change in SI from Base to HumanOmniV2 is $+0.021$ $[-0.019, +0.061]$.

Model	Prompt	E_{A}	E_{V}	SI [95% CI]
Base	native	+1.64	+1.68	0.478 [0.416, 0.538]
Base	neutral	+1.89	+1.27	0.419 [0.366, 0.475]
HumanOmniV2	native	+1.59	+1.40	0.498 [0.441, 0.553]
HumanOmniV2	neutral	+1.00	+0.80	0.486 [0.431, 0.539]
SFT_{ms}	native	+0.89	+0.89	0.484 [0.424, 0.543]
SFT_{ms}	neutral	+0.24	+0.28	0.565 [0.504, 0.621]
MiniCPM-o-2.6	neutral	+1.57	+1.88	0.554 [0.501, 0.608]

Validity checks. Several checks show that the diagnostic measures modality use. Shuffling the family-to-media mapping 1,000 times centers E_{A} and E_{V} on zero, with the observed effects outside the permutation nulls ($p < 0.001$), so the effects require media that carry the family's answers. Answer-level readouts agree with the index. Swaps pull answers toward the incoming sample's ground truth at similar rates for audio and images (25.0% versus 21.2% for the base model). On conflict cells HumanOmniV2 follows the audio in exactly half of the direct answer comparisons, matching its SI ≈ 0.5. On Visual questions, the diagnostic gives $E_{\text{V}} > 2E_{\text{A}}$ (Appendix E), which rules out an index that returns one half by construction. Total sensitivity $|E_{\text{A}}| + |E_{\text{V}}|$ varies more than sixfold across the configurations of Table 4 while SI remains nearly constant, so the index does not simply track the size of the margins.

C Modality Reward Designs

This appendix details the six RL variants of Section 5.5 and shows that none of them removes the cross-modal shortcut. Every run retains the standard format and accuracy rewards and comprises 1,000 GRPO steps on rollouts restricted to prompts of intermediate difficulty. The four judge designs add a modality reward. In each judge design, a judge scores the response's own description of the input, and the score is added to the reward with weight w. The four judge runs start from SFT_{ms}, and the two runs without a modality judge start from $\text{SFT}_{\text{clean}}$. The two checkpoints share their training samples and hyperparameters, and both are trained on cold-start data whose description is split into separate audio and visual parts so that a judge can score each modality separately. SFT_{ms} uses the first version of this split and $\text{SFT}_{\text{clean}}$ a revised version.

The four judge designs differ in how the description is scored. One design (graded description score) uses an external model to score the description on an open scale. In two others, a frozen judge model answers the question from the description alone, with the media hidden. One of them (description gain) bases the reward on how much the description improves the judge's answer. The other (description and answer) combines the correctness of the judge's answer with the correctness of the answer in the response. In the last (yes/no description score), a model that hears the audio or sees the frames returns a yes/no verdict on whether the description is grounded and sufficient. Each change in Table 5 is measured against that run's own starting checkpoint.

Table 5: Changes in the Audio-question diagnostic after RL, paired against each run's own starting checkpoint (65 families unless noted). The Start column gives the starting checkpoint (ms for SFT_{ms}, clean for $\text{SFT}_{\text{clean}}$), and w is the weight of the modality reward. Positive ΔE_{V} indicates a larger image effect. Bold marks the design that increases only the image effect.

Reward design	Start	ΔE_{A} [95% CI]	ΔE_{V} [95% CI]	ΔSI [95% CI]
Graded description score (w=1.0)	ms	-0.012 $[-0.03, +0.01]$	$+\mathbf{0.053}$ $[+0.03, +0.08]$	$+0.019$ $[+0.005, +0.033]$
Description gain (w=0.3)	ms	$+0.051$ $[+0.03, +0.07]$	$+0.069$ $[+0.04, +0.09]$	-0.002 $[-0.018, +0.015]$
Description and answer (w=0.3)	ms	$+0.028$ $[+0.02, +0.04]$	$+0.035$ $[+0.02, +0.05]$	$+0.005$ $[-0.010, +0.020]$
Yes/no description score (w=0.3)	ms	-0.029 $[-0.28, +0.22]$	-0.178 $[-0.40, +0.04]$	inconclusive (31 fam.)
GRPO, no modality reward	clean	$+0.038$ $[+0.02, +0.06]$	$+0.074$ $[+0.05, +0.10]$	$+0.003$ $[-0.011, +0.016]$
GRPO + grid cells	clean	$+0.104$ $[+0.07, +0.14]$	$+0.058$ $[+0.03, +0.09]$	-0.021 $[-0.036, -0.007]$

The three judge rewards evaluated on all 65 families increase the image effect. The graded score increases the image effect while leaving the audio effect unchanged. Although its visual and audio scores enter the reward with equal weight, this judge awards the visual score roughly three times as often as the audio score, so the reward favors visual descriptions. The description gain and the combined description-and-answer reward increase both effects, with a larger point estimate for images. The yes/no design was evaluated on 31 families and is inconclusive. Without a modality reward, GRPO increases the image effect about twice as much as the audio effect. Mixing grid cells into the RL data lowers the index by less than a tenth of the repair's effect in Table 1.

Swap-margin reward. We also screened a swap-margin reward but did not train it, because it would be zero for every completion. The reward is granted only when the answer margin changes sign after the designated modality is swapped and keeps its sign after the non-designated modality is swapped. Before training, we checked whether this reward varies across completions. Across 179 prompts with four completions each, the margins obtained by rescoring the same completion under different media are almost perfectly correlated, and no completion satisfies the gating conditions. After a completion, the answer margin depends far more on the completion text than on the media.

D Interaction Terms

This appendix examines the interaction term of Section 3.2 and shows that the repair lowers the image effect for either audio clip (Section 5.2). The interaction ψ measures how the effect of swapping one modality changes with the source of the other modality. The sign of ψ is defined only once the two items of a pair are ordered, so we use the orientation fixed in Section 3.2, in which the own item is always the one whose ground-truth answer is `yes`. Under that convention, signed ψ values are averaged over a family's two pairs and then across families, exactly like the main effects. A negative ψ means that swapping the audio has a smaller signed effect when the image is the own item's than when it is the partner's. Equivalently, the mean margin of the two mismatched cells lies above that of the two matched cells. Every configuration in Table 6 shows this pattern, as expected if the margin saturates when both modalities support `yes`. Since $|\psi|$ is smaller than both main effects of Table 4 in every configuration, swapping either modality moves the mean margin in the same direction at both levels of the other modality.

Table 6: Interaction terms ψ for Audio questions, under the orientation and aggregation stated above.

Configuration	ψ [95% CI]
Base, native	-0.482 [-0.718, -0.267]
Base, neutral	-0.556 [-0.813, -0.327]
HumanOmniV2, native	-0.513 [-0.769, -0.277]
HumanOmniV2, neutral	-0.353 [-0.511, -0.201]
SFT_{ms}, native	-0.359 [-0.505, -0.229]
SFT_{ms}, neutral	-0.052 [-0.113, $+0.005$]
MiniCPM-o-2.6, neutral	-0.445 [-0.693, -0.218]

Conditional Image Effects

Equation 1 is a statement about each condition separately, so we report the image effect at each level of the audio as well as its average. Writing the image effect at each level of the audio as $\delta_{\text{V}}(o) = m_{oo} - m_{xo}$ and $\delta_{\text{V}}(x) = m_{ox} - m_{xx}$, we have $E_{\text{V}} = \frac{1}{2}[\delta_{\text{V}}(o) + \delta_{\text{V}}(x)]$ and $\delta_{\text{V}}(o) - \delta_{\text{V}}(x) = 2\psi$, so ψ measures how far the two conditional effects differ. Table 7 reports both conditional effects and ψ for the starting checkpoint, the repaired model, and the RL endpoint, together with the mean over families of the absolute image effect $|E_{\text{V}}|$ of each family.

Both conditional effects are positive before the repair and decrease afterward, and the confidence interval of the own-audio effect then includes zero. In the last row of Table 7, we take the larger absolute value of the two conditional image effects in each family and average these values across families. The repair reduces this mean by more than half, so it lowers the image effect within families, not only on average. At the RL endpoint, the partner-audio effect and both per-family means remain well below their pre-repair values.

Table 7: Image effect on Audio questions decomposed by the audio held fixed, 65 development families. Own audio comes from the item whose answer is `yes`, and partner audio from the item whose answer is `no`. The last two rows average over families the absolute image effect and the larger absolute conditional effect of each family, so effects of opposite sign cannot cancel. Intervals use 8,000 resamples.

Statistic	$\text{SFT}_{\text{clean}}$	Repaired (full)	RL endpoint
$\delta_{\text{V}}(o)$, own audio	$+0.682$ [$+0.414$, $+0.966$]	-0.027 [-0.149, $+0.083$]	$+0.019$ [-0.113, $+0.146$]
$\delta_{\text{V}}(x)$, partner audio	$+1.417$ [$+0.905$, $+1.970$]	$+0.165$ [$+0.016$, $+0.319$]	$+0.449$ [$+0.228$, $+0.684$]
ψ	-0.368 [-0.521, -0.233]	-0.096 [-0.181, -0.015]	-0.215 [-0.338, -0.096]
mean $\lvert E_{\text{V}}\rvert$	1.275 [0.939, 1.651]	0.343 [0.278, 0.413]	0.506 [0.417, 0.602]
mean $\max\lvert\delta_{\text{V}}\rvert$	1.687 [1.220, 2.191]	0.629 [0.524, 0.742]	0.928 [0.775, 1.100]

E Per-Question-Type Breakdown

This appendix breaks the development results down by question type to support the validity check of Appendix B and the accuracy statements of Sections 5.2 and 5.5. On Visual questions, where the image is designated, every model in Table 8 gives an image effect more than twice its audio effect.

Table 8: Diagnostic by question type on the development set. For Audio questions, the image is non-designated (SI ideal $= 0$). For Visual questions, the audio is non-designated (SI ideal $= 1$).

Model	Q-type	E_{A}	E_{V}	SI
Base	Audio (65)	+1.635	+1.682	0.478
	Visual (27)	+1.565	+5.014	0.747
HumanOmniV2	Audio (65)	+1.591	+1.403	0.498
	Visual (27)	+1.856	+4.744	0.716
MiniCPM-o-2.6	Audio (65)	+1.565	+1.883	0.554
	Visual (27)	+1.363	+7.104	0.802

The repair improves matched-cell accuracy on Audio questions, and the changes on the other two question types have intervals that include zero (Table 9). A matched cell carries the media of a single real sample, so its ground-truth answer is that sample's own answer. Accuracy is the rate at which this answer receives the higher score under direct scoring, averaged within a family and then across families. Audio-Visual questions contribute only their two original cells to training (Section 4.2), yet their accuracy does not fall after repair and rises after subsequent RL. Because matched cells are the original samples, the repair also improves Audio-question answers on unmodified inputs.

Table 9: Matched-cell accuracy by question type on the development set, under the direct answer scoring of Section 3.2. Values are percentages, and the final column gives the repaired model's paired change against $\text{SFT}_{\text{clean}}$ with 95% family-cluster bootstrap intervals over 8,000 resamples.

Q-type	$\text{SFT}_{\text{clean}}$	Repaired (full)	RL endpoint	Δ (full)
Audio (65)	64.2	72.3	73.8	$+8.1$ [$+2.3$, $+13.5$]
Visual (27)	83.3	80.6	81.5	-2.8 [-10.2, $+4.6$]
Audio-Visual (38)	50.7	53.9	61.8	$+3.3$ [-4.6, $+11.2$]

F Pre-Registered Confirmation Protocol

This appendix reports the pre-registered tests behind the held-out results of Sections 5.2, 5.3, and 5.5. The confirmation set contains 257 families and was constructed together with the 130-family development set. The confirmation set was held out from model selection, hyperparameter tuning, and analysis, and it was used once to evaluate the frozen checkpoints. The seven hypotheses are reported in a pre-specified order, and if one of them does not meet its criterion, the subsequent results are reported descriptively.

Primary results. The repair effect holds on the confirmation set, where all seven pre-registered tests meet their criteria (Table 10). These results confirm the direction and the statistical reliability of the

Table 10: Results for the pre-specified confirmation hypotheses on the 257-family confirmation set. All seven tests meet their criteria, and H2 is reported in two rows. The Audio-question SI tests use 128 families, and H3 evaluates the 53 Visual families. SFT, full, and RL denote the starting checkpoint, the repaired model, and the RL endpoint. H4 uses the repair gain $B = \mathrm{SI}_{\mathrm{SFT}} - \mathrm{SI}_{\mathrm{full}}$ and the change through RL $D = \mathrm{SI}_{\mathrm{RL}} - \mathrm{SI}_{\mathrm{full}}$, observed as 0.000 $[-0.024, +0.023]$, and G2 uses the audio-following rates p_S, p_F, and p_R of the three checkpoints. H2 tests whether the repaired image effect lies within $\pm 10\%$ of $|E_{\mathrm{V}}(\mathrm{SFT})|$ around zero, and the interval of its upper row must lie below zero and that of its lower row above zero. The lower block compares each construction control with the complete-pool LoRA reference as control minus reference, and its last column states whether the result goes in the pre-registered direction.

Test	Statistic	Criterion met
H1 (primary): repair ΔSI	-0.335 $[-0.390, -0.279]$	Yes
G1: repair audio-following Δ (points)	$+13.7$ $[+7.4, +19.7]$	Yes
G2: RL endpoint generation retention (points)	$+11.9$ $[+8.1, +15.6]$	Yes
H4: SI retention through RL ($\geq 80\%$)	$D - 0.2B = -0.067$ $[-0.089, -0.046]$	Yes
H3: Visual-question ΔSI	$+0.176$ $[+0.110, +0.241]$	Yes
H5: LoRA variant ΔSI	-0.352 $[-0.407, -0.294]$	Yes
H2 (upper): $E_{\mathrm{V}}(\mathrm{full}) - 0.1\,\lvert E_{\mathrm{V}}(\mathrm{SFT})\rvert$	-0.088 $[-0.158, -0.017]$	Yes
H2 (lower): $E_{\mathrm{V}}(\mathrm{full}) + 0.1\,\lvert E_{\mathrm{V}}(\mathrm{SFT})\rvert$	$+0.154$ $[+0.061, +0.248]$	Yes
Single-direction cells: ΔSI	$+0.069$ $[+0.019, +0.119]$	Yes
Mismatch labels: ΔSI	$+0.213$ $[+0.163, +0.261]$	Yes
Modality dropout: audio-following Δ (points)	-17.6 $[-23.2, -11.9]$	Yes
Modality dropout: ΔSI (descriptive)	$+0.195$ $[+0.134, +0.255]$	—

repair. Because the confirmation families played no role in selecting the recipe, the 59.9% reduction in Section 5.2 is not an artifact of tuning on the development set.

Retention criteria. Subsequent RL leaves the index unchanged and further improves audio-following, which meets both retention criteria (Table 10). H4 requires the RL endpoint to retain at least 80% of the repair gain B in the index, which holds when the interval of $D - 0.2B$ lies below zero. G2 applies the same 80% criterion to the audio-following rates p_S, p_F, and p_R over all conflict cells and holds when the lower confidence bound of $p_R - 0.8\,p_F - 0.2\,p_S$ exceeds zero. The development index remains near its repaired value throughout the RL run of Section 5.5 (Figure 9), while IntentBench accuracy improves (Figure 6). The suppression therefore does not depend on stopping training at the repaired checkpoint.

Equivalence band of H2. H2 confirms that the repair shrinks the mean image effect below a tenth of its starting magnitude. The pre-registration text specifies a $\pm 20\%$ equivalence band for H2, whereas the frozen analysis script implements the stricter $\pm 10\%$ band that is reported in Table 10. The confidence intervals satisfy both bands.

Construction controls. In all three pre-registered exploratory comparisons, each control underperforms the complete-pool LoRA reference, as specified before the evaluation (lower block of Table 10). For the modality-dropout control, the pre-registered comparison concerns the generated answers, and its change in the index is a descriptive result. For the mismatch-label control, the pre-specified evaluation of the generated answers finds an explicit mismatch verdict in about a third of the conflict-cell outputs, and nearly all of these verdicts are among the outputs that cannot be parsed as `yes` or `no`. Counting these outputs as failures places the control below the starting checkpoint in audio-following (Table 11). This is the behavior anticipated in Section 4.2 for a label that does not depend on the designated modality.

Unparseable outputs. The lead of the complete grid in free generation (Section 5.3) holds under all three treatments of unparseable outputs (Table 11). Most unparseable outputs of the other LoRA-trained models end in a malformed answer tag such as `<yes>Yes</answer>`, which the pre-specified parser rejects, whereas most of those of the mismatch-label control are mismatch verdicts. When these outputs count as failures, the single-direction control trails the reference both because more of its outputs are unparseable and because more of its answers follow the image. When they are excluded, every control remains below the reference in audio-following, with intervals that exclude zero. As a third treatment, we read the text after `</think>` of an unparseable output and accept it as an answer if it contains `yes` but not `no`, or `no` but not `yes`. This recovers most unparseable

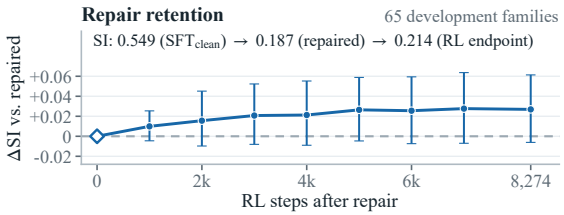


Figure 9: Change in Audio-question SI from the repaired model during subsequent RL post-training, with pointwise 95% paired family-cluster bootstrap intervals.

outputs of the reference and of the single-direction control, and every control again remains below the reference. This analysis of the unparseable outputs was not pre-registered and is descriptive.

Table 11: Generated answers on the 510 confirmation conflict cells, as family-averaged shares in percent of answers that follow the audio, answers that follow the image, and unparseable outputs. The first numeric column is the pre-registered audio-following rate, in which unparseable outputs count as failures. The parseable columns exclude unparseable outputs, pooled over cells, and the lenient column accepts an unparseable output whose text after `</think>` contains only one of `yes` and `no`, and counts the other unparseable outputs as failures. The complete grid is the LoRA reference of Figure 5, and differences to it are paired, with 95% family-cluster bootstrap intervals.

	All conflict cells			Parseable outputs only		Lenient
Training cells	Audio	Image	Unparseable	Audio	Δ to grid	Δ to grid
SFT$_{\text{clean}}$ (no training)	42.8	56.6	0.6	43.0	−24.0 [−30.7, −17.3]	−22.1 [−28.5, −15.6]
Complete 2 × 2 grid	61.3	30.5	8.2	67.0	—	—
Single-direction cells	45.1	36.9	18.0	55.1	−11.8 [−17.3, −6.3]	−12.3 [−17.8, −7.0]
Mismatch labels	33.6	29.9	36.5	52.8	−14.2 [−19.7, −8.6]	−31.2 [−36.1, −26.4]
Modality dropout	43.8	41.6	14.6	51.1	−15.8 [−22.0, −9.6]	−14.5 [−20.5, −8.6]

G Zero-Shot AVQA Protocol

This appendix supports the zero-shot result of Section 5.4 by showing that all three pre-registered criteria are met on AVQA. For the AVQA diagnostic set, we apply the construction rules of Section 3.2 to the items of `OmniInstruct_v1` (Li et al., 2025) whose source field is AVQA. The questions of these items originate from AVQA (Yang et al., 2022) and their clips from VGGSound (Chen et al., 2020a). We add a global media-uniqueness constraint, because AVQA reuses source clips across questions and a single connected component of shared media would otherwise span a large share of the families. Enforcing this constraint removes one family and leaves no media shared across families (Table 3). None of these clips appears in any training configuration. Three criteria and their thresholds were registered before the frozen checkpoints were evaluated, and Table 12 reports them.

Before the repair, the index is lower on AVQA than on MUSIC-AVQA because the audio effect of SFT$_{\text{clean}}$ is larger ($+1.56$ versus $+0.90$) and a similar image effect accounts for a smaller share (Table 13). The image effect is nevertheless significantly positive before the repair, and both the image effect and the index decrease afterward. The repair reduces the same reliance on the image for questions about everyday sounds, outside the music domain of its training cells.

Table 12: Pre-registered AVQA criteria, evaluated once on 104 families after the checkpoints were frozen. Intervals are 95% family-cluster bootstrap CIs over 8,000 resamples, and C3 is paired against $\mathrm{SFT}_{\mathrm{clean}}$. The per-pair ΔSI row follows the registered definition, which averages the signed ratio $E_{\mathrm{V}}/(|E_{\mathrm{A}}| + |E_{\mathrm{V}}|)$ over pairs. The family-level SI of Section 3.2, which Table 13 reports, changes by -0.174 [-0.210, -0.140]. The generation rows use 40 of the 104 families (160 conflict cells) on AVQA and the 65 development Audio families on MUSIC-AVQA.

	Statistic	Result	Met
C1 answerable	$\mathrm{SFT}_{\mathrm{clean}}$ matched-cell accuracy $>$ chance	68.3% [64.2, 72.4]	Yes
C2 replicates	$\mathrm{SFT}_{\mathrm{clean}}$ $E_{\mathrm{V}} > 0$	$+0.852$ [$+0.694$, $+1.008$]	Yes
C3 repair transfers	$\Delta E_{\mathrm{V}} < 0$	-0.529 [-0.664, -0.399]	Yes
	per-pair ΔSI < 0	-0.104 [-0.167, -0.040]	Yes
	matched-cell accuracy drop ≤ 2 points	68.3% $\rightarrow$ 82.2%	Yes
descriptive	conflict audio-following (gen.)	44.4% $\rightarrow$ 62.5%	—
	same quantity on MUSIC-AVQA	43.1% $\rightarrow$ 61.2%	—

Table 13: Audio-question diagnostic on the 104 AVQA families under the native prompt. Intervals are 95% family-cluster bootstrap CIs over 8,000 resamples.

Model	E_{A} [95% CI]	E_{V} [95% CI]	SI [95% CI]
Base	$+2.911$ [$+2.564$, $+3.286$]	$+0.950$ [$+0.764$, $+1.149$]	0.292 [0.253, 0.330]
HumanOmniV2	$+3.381$ [$+2.970$, $+3.785$]	$+1.001$ [$+0.801$, $+1.199$]	0.302 [0.258, 0.347]
$\mathrm{SFT}_{\mathrm{clean}}$	$+1.556$ [$+1.337$, $+1.776$]	$+0.852$ [$+0.694$, $+1.008$]	0.392 [0.349, 0.436]
Repaired (full)	$+1.522$ [$+1.347$, $+1.693$]	$+\mathbf{0.323}$ [$+0.247$, $+0.402$]	**0.218** [0.186, 0.251]
RL endpoint	$+2.044$ [$+1.806$, $+2.276$]	$+0.466$ [$+0.359$, $+0.576$]	0.238 [0.201, 0.274]

H ZERO-SHOT AVHBENCH PROTOCOL

This appendix details the AVHBench evaluation behind Sections 5.2 and 5.4 and traces the audio-task gain to audible objects. We evaluate the three yes/no tasks of AVHBench (Kim et al., 2025), which comprise 5,302 questions about AudioCaps and VALOR clips in their original and audio-swapped versions. These questions reference 2,092 distinct video identifiers in our evaluation files, while the benchmark reports 2,136 videos across its four tasks. Video-driven audio hallucination asks whether a visible object is audible (2,290 questions). Audio-driven video hallucination asks whether an audible object is visible (1,136 questions), and audio-visual matching contributes another 1,876 questions. Each task has balanced labels.

We use the IntentBench evaluation pipeline for the Qwen checkpoints. Video and audio are supplied as separate inputs, with the HumanOmniV2 system prompt and output format. The predicted answer is the first yes/no token within the `<answer>` tags. Unparseable outputs count as errors and make up at most 0.2% of outputs for every Qwen checkpoint. MiniCPM-o-2.6 is scored as in its diagnostic, with the neutral prompt and the first-token yes/no margin (Appendix A). Its diagnostic interface takes a single image, so the model receives the middle frame of each video together with the full audio track. Accuracy and the paired difference of Section 5.2 are averaged over questions. The interval of this difference uses a paired bootstrap over video clusters with 10,000 resamples, retaining the same per-question weighting, so the gain equals the difference between the unrounded accuracies.

The repair improves the two tasks that require listening, and subsequent RL extends this gain only on video-driven audio hallucination (Table 14). On audio-driven video hallucination and audio-visual matching, the RL endpoint remains within a few tenths of a point of the repaired model. The gains of both stages concentrate on the listening tasks, where the diagnostic locates the shortcut.

The audio-task gain primarily reflects improved detection of previously missed audible objects. Accuracy on audible objects rises at each stage (Section 5.4), whereas accuracy on silent objects falls by less than four points. Within the audio task, answers that change from `no` at the starting checkpoint to `yes` at the RL endpoint outnumber the reverse changes by almost three to one. Because most changed answers are corrections in both directions, the rising yes-rate in Table 14 mainly reflects the detection of audible objects rather than a general preference for `yes`.

Answer accuracy improves even though most outputs still omit an explicit audio description. Empty audio descriptions (`<not_available>`) occur in nearly all outputs of the starting checkpoint and

Table 14: Zero-shot AVHBench accuracy (%) by task, with the yes-rate (%) in parentheses. Tasks are video-driven audio hallucination (audio hall.), audio-driven video hallucination (video hall.), and audio-visual matching. The table covers all 5,302 questions, and "Overall" is the question-weighted mean over the three tasks. The two model families use different answer readouts and are not directly comparable to each other, so each is compared with its own starting checkpoint. Of the audio-hallucination answers that change from `no` to `yes` and from `yes` to `no` between the starting checkpoint and the RL endpoint, 69% and 64%, respectively, are correct at the RL endpoint.

Model	Readout	Audio hall.	Video hall.	AV matching	Overall
Qwen2.5-Omni $\text{SFT}_{\text{clean}}$	gen.	61.9 (37.2)	81.5 (48.6)	55.4 (62.3)	63.8
+ repair (full)	gen.	67.7 (44.1)	82.0 (47.9)	62.6 (58.6)	69.0
+ subsequent RL	gen.	**73.2** (52.1)	**82.2** (54.2)	**62.7** (65.8)	**71.4**
MiniCPM-o-2.6 (base)	margin	78.7 (44.2)	76.0 (31.8)	64.0 (19.1)	72.9
+ repair (LoRA)	margin	**80.2** (41.1)	**78.0** (35.2)	**64.8** (16.6)	**74.3**

still in more than four in five after the repair and after RL. This is expected under answer-token supervision, which sets no target for the generated description (Section 4.3).

The LoRA repair also improves the accuracy of MiniCPM-o-2.6 over its base model, most on the two hallucination tasks. In this model family, the yes-rate falls on the audio task and on audio-visual matching as accuracy rises (Table 14). The cross-model gain therefore does not stem from a general shift toward `yes`.

Table 2 in Section 5.4 compares the Qwen results with published methods on the same backbone.

I Generated Answers on Conflict Cells

This appendix adds eight conflict cells to the case study of Section 5.6. All outputs use greedy decoding under the native prompt (Appendix L), and the text in each box is quoted from the model's output and truncated. Questions keep the wording of the datasets.

Figure 10 shows two cells in the format of Figure 7 on which all three checkpoints follow the image. On the MUSIC-AVQA cell, the audio is an ensemble piece, and the repaired model and the RL endpoint describe its violin melody but still answer from the image. On the AVQA cell, the RL endpoint describes the audio through the beach scene in the image.

Figure 11 adds six successful cells, four from development pairs and two, in rows three and four, from the confirmation set. On 67 of the 260 development conflict cells, the starting checkpoint follows the image, whereas both the repaired model and the RL endpoint follow the audio. The four development rows are drawn from these, and the two confirmation rows follow the same pattern. In Figure 11, the audio description of $\text{SFT}_{\text{clean}}$ is empty on every cell and its answer follows its image description, whereas the repaired model and the RL endpoint state what the audio contains.

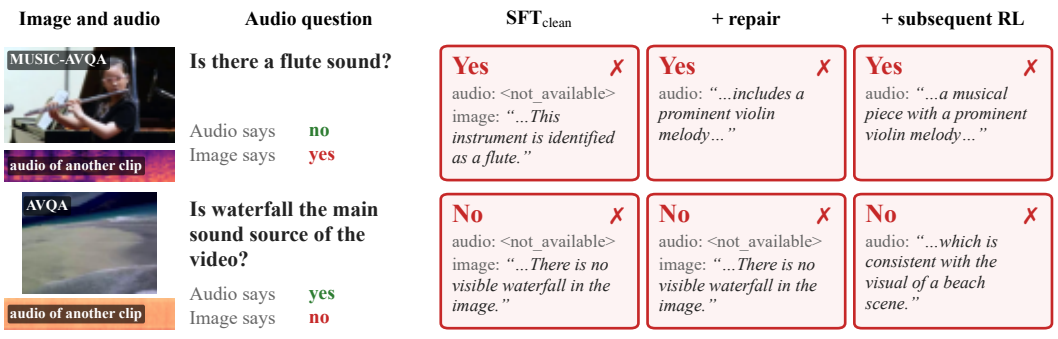


Figure 10: Two conflict cells in the format of Figure 7. All three checkpoints follow the image.



Figure 11: Six conflict cells in the format of Figure 7, four from development pairs and two, in rows three and four, from the confirmation set. Rows one and two are the two directions of the same pair, and the remaining rows show cells from four other pairs. Green boxes indicate outputs that follow the audio, and red boxes indicate outputs that follow the image.

J Loss and Parameterization Ablations

This appendix supports the attribution of the gain to the training cells in Section 5.3. Table 15 lists the variants that underlie the loss and parameterization paragraph of Section 5.3, together with the matched-cells-only control. Every run is evaluated on the 65 development Audio families and follows the recipe of Appendix A, except the two runs marked † in Table 15, which are trained with cross-entropy on an earlier and smaller pool. ΔSI is the paired difference from the starting checkpoint under direct scoring, with a family-cluster bootstrap interval. For audio-following, we compute within each family the proportion of all conflict cells whose generated answers follow the audio, and then average these proportions across families.

The loss and the parameterization have little effect on the index reduction. Cross-entropy achieves 89% of the index reduction obtained with the hinge loss under the same recipe, and full fine-tuning and LoRA reach almost the same ΔSI.

Removing the counterfactual cells has a much larger effect. The matched-cells-only control achieves less than a third of the index reduction of the reference. Relative to cross-entropy training on all cells of the same earlier pool, it achieves about a third of the reduction, as stated in Section 5.3. The control's audio-following does not differ detectably from that of the starting checkpoint (Table 15). Without counterfactual cells, the image predicts the label as well as the audio does (Section 4.2).

Table 15: Loss and parameterization variants, development families. The reference row is the LoRA recipe on the complete pool, which is also the reference of Figure 5. The seed row reports the range over four seeds of that reference, each with an interval that excludes zero, and full fine-tuning differs from the reference by -0.007 [-0.053, $+0.042$] in paired ΔSI. The matched-cells-only control differs from $\text{SFT}_{\text{clean}}$ by -4.6 [-12.3, $+3.5$] points in audio-following. †Trained with cross-entropy on an earlier and smaller pool.

Variant	Training cells	ΔSI [95% CI]	Audio-Following, % (gen.)
$\text{SFT}_{\text{clean}}$ (no training)	—	—	43.1
Hinge, LoRA (reference)	complete 2×2	-0.356 [-0.427, -0.285]	60.8
Hinge, full fine-tuning	complete 2×2	-0.362 [-0.430, -0.293]	61.2
Cross-entropy, LoRA	complete 2×2	-0.317 [-0.377, -0.258]	61.9
Hinge, LoRA, four seeds	complete 2×2	-0.356 to -0.296	—
Cross-entropy, LoRA†	complete 2×2	-0.297 [-0.364, -0.232]	—
Matched cells only, LoRA†	two original cells	-0.101 [-0.166, -0.038]	38.5 [32.3, 44.6]

K Cross-Model Generalization on MiniCPM-o-2.6

Table 16 gives the Audio-question diagnostic behind the cross-model result of Section 5.2. The LoRA repair of MiniCPM-o-2.6 uses the same training cells as the Qwen recipe (Appendix A) and reduces the index by about one half on both the development and the confirmation set. The image effect falls to about one tenth of its value before the repair on both sets, whereas the audio effect decreases by about one fifth. The lower index reflects the removal of most of the image effect, while most of the audio effect remains.

Table 16: Audio-question diagnostic of MiniCPM-o-2.6 before and after the repair, under the neutral prompt and the first-token margin. Arrows go from the base model to the repaired model, and ΔSI is the paired change with its 95% family-cluster bootstrap interval over 10,000 resamples.

Set (Audio families)	E_{A}	E_{V}	SI	ΔSI [95% CI]
Development (65)	$+1.57 \to +1.27$	$+1.88 \to +0.20$	$0.554 \to 0.261$	-0.293 [-0.361, -0.223]
Confirmation (128)	$+1.17 \to +0.92$	$+1.68 \to +0.17$	$0.604 \to 0.297$	-0.307 [-0.365, -0.248]

L Prompts

This appendix lists the prompts of all experiments. Braces mark fields that are filled per sample, and the original spelling of each prompt is preserved.

System prompt. All Qwen checkpoints under the native prompt, the subsequent GRPO runs, and the Qwen benchmark evaluations use the system prompt of HumanOmniV2 (Yang et al., 2025).

```
You are a helpful assistant.  Your primary goal is to deeply
analyze and interpret information from available various
modalities (image, video, audio, text context) to answer
questions with human-like depth and a clear, traceable
thought process.

Begin by thoroughly understanding the image, video, audio or
other available context information, and then proceed with
an in-depth analysis related to the question.

In reasoning, It is encouraged to incorporate
self-reflection and verification into your reasoning
process.  You are encouraged to review the image, video,
audio, or other context information to ensure the answer
accuracy.

Provide your understanding of the image, video, and audio
between the <context> </context> tags, detail the
reasoning between the <think> </think> tags, and then give
your final answer between the <answer> </answer> tags.
```

Diagnostic and generation prompt. The diagnostic of Section 3.2 and free generation give the model the image, the audio, and the following text.

```
Here is the image, with the coresponding audio.
{question}
Possible answers:  yes, no
```

Direct scoring appends <answer> as the start of the assistant turn and reads $\log p(\texttt{yes}) - \log p(\texttt{no})$ at the next token. For the own-reasoning index of Table 1, the model first generates up to 512 tokens greedily on the matched cell (i_o, a_o). We keep the text before its first <answer>, or the full text if there is none, and append <answer> to obtain the prefix shared by all four cells. Free generation decodes greedily with up to 512 new tokens and takes the first yes or no after <answer>. If there is no such match, the parser reads the answer region, or the last two lines of the output when it has no answer region. It accepts this text if it contains affirmative words but no negative words, or negative words but no affirmative words. Any remaining output counts as unparseable. In the development evaluation, 1% of the generated outputs of $\text{SFT}_{\text{clean}}$, 8% of those of the LoRA model, and 2% of those of the fully fine-tuned model are unparseable and count as failures. The neutral prompt omits the system prompt and the answer prefix, and scoring uses the first assistant token. In both the diagnostic and AVHBench, MiniCPM-o receives the same text after its media markers (<image>./</image>)(<audio>./</audio>).

Training prompt. DMC-Repair uses the same system prompt and text, except that the answer list contains all 42 answers of the training pool.

```
Possible answers:  accordion, acoustic_guitar, bagpipe,
banjo, bassoon, cello, clarinet, congas, drum, eight,
electric_bass, erhu, five, flute, four, guzheng, indoor,
left, middle, more than ten, nine, no, one, outdoor, piano,
pipa, right, saxophone, seven, simultaneously, six, suona,
ten, three, trumpet, tuba, two, ukulele, violin, xylophone,
yes, zero
```

The score of a candidate is its mean token log-probability after <answer>, and the hinge loss of Equation 7 compares the two candidates of each cell. MiniCPM-o uses its own chat template with the media markers above and without a system prompt or answer prefix.

Benchmark prompts. For the Qwen checkpoints, AVHBench and IntentBench are evaluated with the code of Yang et al. (2025) and the system prompt above. The user turn contains the video, its audio track, and the following text.

```
Here is a {data_type}, with the audio from the video.
{question}{type_template}
```

AVHBench uses video as the data type and the yes/no template below.

```
Please answer Yes or No within the <answer> </answer>
tags.
```

Multiple-choice questions list their options after the word Options: and use this template.

```
Please provide only the single option letter (e.g., A, B, C,
D, etc.)  within the <answer> </answer> tags.
```

The other question types of IntentBench use the corresponding templates of the same code. The GRPO runs use the same system prompt and templates, with a line break between the question and the template.

M Extended Related Work

Modality-aware RL objectives. SFFL (Li et al., 2026a) pairs modality-specific reasoning traces with a modality-preference reward, OmniVideo-R1 (Chen et al., 2026c) modifies query grounding and modality-attentive fusion, and MAPO (Xiao et al., 2026) reweights the policy gradient on

modality-critical tokens alongside an attention-based auxiliary loss. Caption-based self-verification (Xia et al., 2026) and modality-separated self-reward (Li et al., 2026b) pursue the same goal for images. DMC-Repair differs in that it changes which input predicts the training label.

Diagnostics for modality dependence. Ma et al. (2024) introduce MUSIC-AVQA-R to test robustness to question rephrasing and propose logit-space debiasing, while Sensory PID (Fang et al., 2026) decomposes modality contributions through partial information decomposition. A probing study finds that audio information present in intermediate representations can be suppressed in later fusion layers (Selvakumar et al., 2026). The Factorized Modality Diagnostic complements these analyses with a measurement of each modality's effect on individual answers.

Hallucination mitigation and counterfactual training. Among preference methods, ACPO (Baid et al., 2026) penalizes captions that ignore a swapped audio track, MoD-DPO (Chaubey et al., 2026) encourages invariance to irrelevant-modality changes, and OmniDPO (Chen et al., 2026a) combines text and multimodal preferences. Training-free methods instead contrast or reweight modality branches at decoding time (Jung et al., 2025; 2026; Chung et al., 2026). Earlier counterfactual VQA work synthesizes examples by masking critical evidence and changing labels (Chen et al., 2020b). Table 2 compares DMC-Repair with the published AVHBench results of several of these methods.