Beyond Text Following: Repairable Arbitration Reversals in Audio–Language Models

Anonymous ACL submission

Abstract

Audio-language models (ALMs) often follow text that conflicts with audio, even when the audio evidence is clear. This raises a basic question: is the audio-supported answer unavailable, or is it represented but overridden by the conflicting text? We examine this question using a same-audio counterfactual that keeps the audio fixed, removes only the conflicting text, and measures the resulting shift in model preference. Across five ALMs and four conflict tasks, 64.1% of conflict samples show a sign flip: the same-audio branch prefers the audio-supported answer, whereas the joint branch prefers the text-supported answer. This pattern suggests that the relevant audio evidence is encoded but loses in arbitration. Activation patching further localizes the reversal to answer-position computation, and patching effects closely track output candidate-score differences (Spearman ρ=0.93). Using this diagnostic, we propose **G**ated **A**udio **C**ounterfactual **L**ogit Correction (GACL), a training-free decoding rule that interpolates between joint and same-audio scores. Under a strict 5,pp faithfulness-drop budget, GACL improves nAUC by 17.8 points over the best contrastive baseline and transfers without retuning to vision–text arbitration (up to +40.5,pp). Code: `https://anonymous.4open.science/status/arbitration-reversal-gacl-C8BB`.

1 Introduction

When an audio-language model (ALM) hears an alarm while the text states that no alarm was triggered, it may answer "no alarm." This error shows text following, but not how the decision was made. The model may have missed the alarm, or encoded it and let the textual claim override it. This distinction matters: perceptual failure calls for better acoustic encoding, whereas arbitration failure calls for better conflict resolution. As ALMs (Chu et al., 2024; Xu et al., 2025a; Liu et al., 2025) move beyond transcription into agentic settings such as meeting assistance and emergency triage, they may receive audio together with written context such as agendas, logs, or incident notes. When that context conflicts with the audio, users need models that resolve the disagreement based on the heard evidence rather than repeating the text.

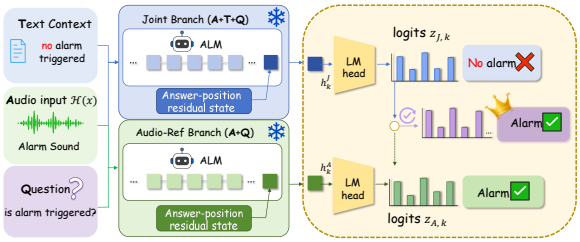

Figure 1. GACL overview. The joint branch conditions on audio plus conflicting text and predicts the text-supported answer; the audio-reference branch removes only the text and predicts the audio-supported answer. GACL freezes both ALM branches and applies gated, bounded interpolation from joint logits toward reference logits.

Recent benchmarks expose this symptom but not its mechanism. MCR-Bench (Wang et al., 2025) and ALME (Billa, 2026) show that ALMs often choose the text-supported answer when audio and text conflict, when the audio evidence is clear. These benchmarks, however, do not show whether the audio-supported answer is unavailable or whether it remains available but loses to the text during arbitration. Decoding-time methods face the same limitation: their reference branches weaken modality evidence. Audio methods such as AAD and ACD (Hsu et al., 2025; Lin et al., 2026), together with their vision–language counterparts VCD, OPERA, and ICD (Leng et al., 2024; Huang et al., 2024; Wang et al., 2024), construct reference branches by weakening or removing modality evidence. Such references can diagnose reliance on degraded evidence, but audio–text arbitration requires a different counterfactual: with the same audio held fixed, how does model prefer-

ence change when only the conflicting text is removed?

We study this counterfactual directly. For each sample, we keep the audio unchanged, remove the conflicting text, and compare preferences for the audio-supported and text-supported answers. Across five ALMs and four conflict tasks, 64.1% of conflict samples show a sign flip: the same-audio branch prefers the audio-supported answer, whereas the joint audio–text branch prefers the text-supported answer. The audio answer therefore remains available, but it loses once the text is added. We refer to this pattern as *repairable*[1] *arbitration reversal*. Activation patching further localizes the reversal to the answer-position residual stream, and the patch-induced direction tracks the score difference $s_A - s_J$ at the rank level (ρ=0.93). This alignment makes the repair direction visible from outputs alone, and motivates a decoding-time fix.

Using this diagnostic, we propose **G**ated **A**udio **C**ounterfactual **L**ogit Correction (GACL; Figure 1), a training-free decoding rule that interpolates between joint and same-audio reference scores, gated by branch disagreement and reference reliability. Under a 5 pp faithfulness-drop budget, GACL improves conflict audio-following accuracy by 17.8 nAUC points over the best contrastive decoding baseline. Applied without retuning to vision–text arbitration on MC^2 (Zhang et al., 2025), the same rule raises adversarial accuracy by up to 40.5 pp, suggesting that this counterfactual design generalizes beyond audio–text.

Our key contributions are:

- We identify repairable arbitration reversal as a common recoverable failure mode in audio–text conflict: models can support the audio answer, but the joint input reverses the decision.
- We localize this reversal to the answer-position residual stream and show that the patch-induced direction tracks the output score difference $s_A - s_J$ at the rank level (Spearman ρ=0.93), making the repair direction visible from outputs alone.
- We propose GACL, a diagnostic-derived decoding rule improving audio–text rescue by 17.8 nAUC under a 5 pp faithfulness budget and generalizing to vision–text conflict without retuning.

[1] *Repairable* is used operationally: samples in O have an audio-favoring same-audio reference and a text-favoring joint branch, providing a candidate set for correction.

2 Related Work

Audio-language models and evidence conflict. Recent ALMs adapt instruction-following LLMs to speech, music, and acoustic events via audio encoders or unified multimodal tokenization (Tang et al., 2024; Gong et al., 2024; Ghosh et al., 2024; Kong et al., 2024; Chu et al., 2024; Xu et al., 2025a,b; KimiTeam et al., 2025; Liu et al., 2025). Beyond transcription, text can become competing evidence rather than merely a task cue. MCR-Bench (Wang et al., 2025) and ALME (Billa, 2026) show that mainstream ALMs often follow text under explicit conflict. This behavior is related to evidence-arbitration failures studied in knowledge conflicts, context faithfulness, and sycophancy-like instruction following (Xu et al., 2024; Li et al., 2025; Sharma et al., 2024; Wei et al., 2024). These benchmarks identify the failure mode, but not whether the audio-supported answer is unavailable or instead remains available but is displaced at the decision point.

Reference design in decoding-time correction. Decoding-time correction methods adjust predictions using reference distributions, perturbed inputs, or penalties, without parameter updates. The reference defines the counterfactual error the method can suppress. Model-level contrast (Li et al., 2023), layer-level contrast (Chuang et al., 2024), vision–language methods (Leng et al., 2024; Wang et al., 2024; Huang et al., 2024), and audio contrastive methods such as AAD and ACD (Hsu et al., 2025; Lin et al., 2026) typically weaken, remove, or perturb modality evidence. These references measure reliance on degraded evidence, but arbitration requires a different test: with the same audio held fixed, which answer does the model support after only the competing text is removed? This gap in reference design motivates our direct comparison with degraded-reference baselines under a rescue–faithfulness trade-off.

3 Diagnosis: From Behavior to a Repairable Signature

When an ALM follows text in an audio–text conflict, two failures can look identical from output alone. The model may fail to extract the relevant audio evidence, or extract it but still allow the text to determine the answer. These cases require different repairs. We therefore begin with a counter-

***J* (joint)**: `[audio]`
Description: *The sound is not constant.*
Question: Is the sound constant? Please answer with yes or no.
***A* (audio reference, REF)**: `[audio]`
Question: Is the sound constant? Please answer with yes or no.

Figure 2. Two-branch setup. The joint branch (J) conditions on audio plus conflicting text; the same-audio reference (A) removes only the text. Both share the same audio, question, and output format.

factual comparison that changes only one factor: whether the conflicting text is present.

3.1 Setup: Two Branches and Two Margins

Two branches. For each conflict instance, we use two prompts with the same audio x, question q, candidate set $\mathcal{C}$, and output format. They differ only in whether the conflicting textual description is included (Figure 2): Here, the audio contains a continuous 23 s alarm, so the audio-supported answer is y_a = yes, while the conflicting text supports y_t = no. Because J and A keep the audio, question, candidate set, and output format fixed, the comparison isolates the effect of adding the conflicting text.

Two margins. A top-1 prediction hides preference strength and direction. We therefore use signed log-probability margins. For each branch $b \in \{A, J\}$, we score closed-set verbalizers and write $s_b(c)$ for the length-normalized, candidate-normalized log-probability of candidate $c \in \mathcal{C}$. Appendix A.2 and Appendix A.3 give the scoring details. We define

$$M_A = s_A(y_a) - s_A(y_t), M_J = s_J(y_a) - s_J(y_t). \quad (1)$$

These margins answer two distinct questions. $M_A > 0$ indicates that, without the conflicting text, the model prefers the audio-supported answer, testing whether the audio branch can support that answer. $M_J < 0$ indicates that, after the conflicting text is added, the model prefers the text-supported answer; this tests whether arbitration fails. Because both margins are computed on the same candidate pair (y_a, y_t), they can be compared directly. The diagnostic region is

$$O = \{M_A > 0,\ M_J < 0\}. \quad (2)$$

Samples in O exhibit the behavioral signature of arbitration failure: the audio answer is available in the same-audio reference branch, but the joint branch flips toward the text-supported answer. We treat O as a diagnostic region for possible arbitration failure, not a claim that every sample in this region can be repaired by a particular method.

Models and datasets. We evaluate five open-weight audio-language/omni-modal models (7B–30B parameters): Qwen2-Audio-7B-Instruct (Chu et al., 2024), Qwen2.5-Omni-7B (Xu et al., 2025a), Voxtral-Small-24B (Liu et al., 2025), Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b), and Kimi-Audio-7B-Instruct (KimiTeam et al., 2025). We use four audio–text conflict tasks: AQA (Audio Question Answering), VSC (Visual Scene Classification), and SER (Speech Emotion Recognition) from MCR-Bench (Wang et al., 2025) for audio–description conflict, and the English subset of ALME (Audio-LLM Modality Evaluation) (Billa, 2026) for speech–transcript conflict. Each provides 200 development examples for hyperparameter selection plus a held-out 200-example test split. Appendix A.1 gives details.

3.2 Behavioral Evidence

We now apply the two-branch diagnostic to the conflict benchmarks. The goal is to measure conflict failure and determine whether it reflects missing audio evidence or arbitration reversal. We compare three conditions: faithful inputs, where the accompanying text, if present, is consistent with the audio-supported answer; the joint conflict branch J, where the conflicting text is present; and the same-audio reference branch A, where the conflicting text is removed while the audio is kept fixed. Table 1 summarizes top-1 behavior, and Figure 3 shows the signed margins on conflict samples. Faithful inputs reach 97.9% audio-following; conflict drops accuracy to 18.0% with 80.3% of predictions matching the text—a large, systematic shift from the audio. Yet when only the conflicting text is removed, the A-branch gives the audio-supported answer on 76.6% of conflict samples—the audio channel still carries the information. The margin plot further shows that the dominant pattern is signed reversal, not a small confidence decrease. If conflict merely weakened preference for y_a, many samples would lie near $M_J = 0$. Instead, 64.1% of conflict samples fall in $O = \{M_A > 0, M_J < 0\}$: the same candidate pair is audio-favored in A but text-favored in J. This sign flip provides the behavioral ba-

Table 1. Behavioral evidence for arbitration failure. Faithful and Conflict columns report audio-following accuracy; Text-follow reports the fraction of conflict predictions matching y_t; Audio-ref reports accuracy after removing only the conflicting text; O is the share of conflict samples in $\{M_A > 0, M_J < 0\}$. Results are macro-averaged across five models. Per-configuration results appear in Appendix B.1.

Task	Faithful	Conflict	Text-follow	Audio-ref	O
AQA	99.4	18.2	81.8	85.0	69.3
VSC	96.7	14.4	81.6	75.9	73.1
SER	98.4	5.0	92.0	48.9	56.6
ALME	97.1	34.4	65.6	96.5	57.5
Macro	97.9	18.0	80.3	76.6	64.1

Table 2. The repair direction localizes to the answer boundary. Answer-side residual patching recovers the audio-supported answer across models and tasks, whereas audio-span and instruction-token interventions give zero rescue in all controls.

Site / Model	AQA	VSC	SER	ALME	Avg.
Answer-side residual					
Qwen2-Audio-Inst.	1.00	0.96	0.44	0.81	0.80
Qwen2.5-Omni	1.00	0.88	0.54	0.86	0.82
Voxtral-Small	1.00	0.95	0.56	0.87	0.85
Qwen3-Omni	0.94	0.88	0.58	0.98	0.85
Kimi-Audio-Inst.	0.94	0.62	0.40	0.90	0.72
Macro	**0.98**	**0.86**	**0.50**	**0.88**	**0.81**
Negative controls					
Audio-span residual	0.00	0.00	0.00	0.00	0.00
Instruction-token residual	0.00	0.00	0.00	0.00	0.00

sis for the internal analysis in the next section. The A-branch is informative on AQA/VSC/ALME (75.9–96.5%) but weak on SER (48.9%); we treat this as a benchmark property and revisit its consequences in §5.3. The behavioral evidence suggests that many conflict errors are not reducible to perceptual failure: removing only the conflicting text often recovers the audio-supported answer, and many samples show a signed reversal from audio-favored in A to text-favored in J. However, behavior alone does not reveal the joint branch's internal computation. One possibility is that the joint branch never makes audio-compatible evidence available at the answer position, while the same-audio branch reaches y_a through a different computation path. Another possibility is that the joint branch contains audio-compatible evidence, but this evidence is suppressed or overwritten during readout. These two accounts imply different repair targets, so we turn to the model's internal computation.

4 From Mechanism to a Decoding Rule

§3 left one question open: in J, does the model compute y_a and then let text dominate at readout, or fail to compute y_a at all? The answer determines where repair should intervene. We ask two questions: can the difference between J and A be causally localized inside the model, and does it appear in the final candidate scores, the only quantities available to an output-space rule?

4.1 Causal Localization via Activation Patching

We use activation patching, a standard causal test for whether an internal state drives an output (Vig et al., 2020; Meng et al., 2022; Zhang and Nanda, 2024). We inject hidden states from A into matched positions of J and test whether J returns to y_a. Since A and J differ in prompt length, we match positions by semantic role rather than absolute index: audio-span states map to the shared audio segment, instruction-token states to the shared question and output format, and answer-position states to the final prompt position before generation. We patch the residual stream as the primary site, since it aggregates component outputs.

Patching requires a clean source–target direction, so we restrict to the reference-correct subset $R = \{\hat{y}_A = y_a, \hat{y}_J \neq y_a, x \in O\} \subseteq O$. On R, we inject residual states from A into J at the three position classes and measure rescue rates in Table 2. *Answer-position residual* patching reaches a macro average of 0.81. Replacing J's state at the answer commitment point recovers y_a on most samples. By contrast, *audio-span* and *instruction-token* residual patching give zero rescue in the reported controls. Because the branches share audio and instructions, these sites serve as negative controls. Their zero rescue suggests that the relevant difference is not in perception or instruction handling, but downstream near response generation.

Two additional checks refine this picture. Component-level patching shows that neither attention nor the MLP alone explains the full residual effect; the failure is better described as a distributed residual-stream write than as a single-module artifact (Appendix C.1). Layer-wise sweeps place the rescue effect in a mid-to-late layer band, where the model appears to move from evidence integration to answer commitment; we call this band the *commit window* and define it operationally in Appendix C.2. A linear-probing audit (Alain and Bengio, 2016; Hewitt and

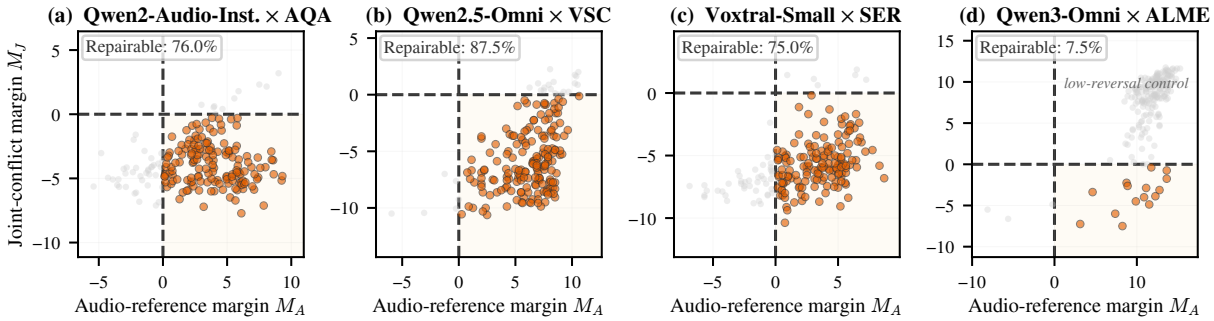

Figure 3. Conflict failures show up as a signed margin reversal. M_A is on the x-axis and M_J on the y-axis. Orange points mark samples in $O = \{M_A > 0, M_J < 0\}$, where the same candidate pair is audio-favored in the same-audio reference branch but text-favored in the joint branch. Appendix B.2 shows the full 5×4 grid.

Liang, 2019) on Qwen2-Audio-Inst.×VSC gives the same interpretation: y_a remains decodable from late states of both branches (∼0.89), but cross-branch decoding collapses after the commit window (to 0.087). The audio answer is not erased; the readout direction changes (Appendix C.3).

In J, the audio answer is computed and remains accessible; conflict changes the answer-side readout. The failure is readout, not perception.

4.2 Bridging Internal Repair Direction to Output Scores

A hidden-state repair direction does not imply that an output-space rule can use it. A patching effect is local to the injection site, whereas the final score difference between branches includes all downstream computation. These directions could therefore differ.

We test alignment directly. For each sample, we compute the patch-induced displacement at the best commit-window layer and the observable candidate-level difference $s_A - s_J$. We compare them with Spearman ρ (Figure 4). Across twelve model–task configurations, macro alignment is $\rho = 0.93$, well above the raw logit-lens projection (nostalgebraist, 2020) ($\rho = 0.67$). Using $s_A - s_J$ as the rescue direction reaches 92.3% best rescue, compared with 54.7% for the lens. Across layers, alignment is near zero early and rises through the same commit window where patching becomes effective (Figure 4b).

One limitation matters for decoding. The alignment is rank-based: $s_A - s_J$ reliably indicates which direction restores the audio answer, but not how far the state should move.

4.3 GACL

Two findings shape an output-space rule. From §4.1 and §4.2, the hidden repair direction at the answer position is rank-aligned with the observable difference $s_A - s_J$, so a rule should move predictions along that direction. From §3, correction conditions are not uniform: margins vary widely across samples (Figure 3), and the same-audio reference is not equally reliable across tasks (A-branch accuracy ranges 48.9–96.5%; Table 1). A useful rule must act per sample, conditional on branch disagreement and reference reliability.

The rule. For input x, let $\hat{y}_J, \hat{y}_A$ be parsed predictions and s_J, s_A scores over candidates $\mathcal{H}(x)$. Candidate construction, answer normalization, and the free-generation interface are in Appendix D. GACL combines a branch-disagreement gate, a reference-reliability gate, and a bounded interpolation:

$$\begin{aligned} N_{\text{out}}(x) &= \mathbf{1}[\hat{y}_A \neq \emptyset] \cdot \mathbf{1}[\hat{y}_A \neq \hat{y}_J], \\ \Delta_A(x) &= s_A(\hat{y}_A) - \max_{\substack{y \in \mathcal{H}(x) \\ y \neq \hat{y}_A}} s_A(y), \\ R_A(x) &= \text{clip}(\Delta_A(x)/\tau_A,\, 0,\, 1)\,, \\ \alpha(x) &= \text{clip}(\lambda\, R_A(x)\, N_{\text{out}}(x),\, 0,\, 1)\,, \\ z_k^{GACL} &= z_{J,k} + \alpha(x)\big(z_{A,k} - z_{J,k}\big). \end{aligned} \tag{3}$$

Here $R_A(x) = 0$ when $\hat{y}_A = \emptyset$ or $\mathcal{H}(x)$ has no competitor; $z_{J,k}, z_{A,k}$ are next-token logits at step k given the corrected prefix $u_{<k}$. The same form applies to closed-set candidate-score vectors.

Convex form as a safety choice. Restricting $\alpha \in [0, 1]$ keeps the correction between z_J and z_A. The rank alignment in §4.2 certifies the direction but not magnitudes past the A endpoint; in practice, allowing $\alpha > 1$ sometimes adds rescue in closed-set settings but introduces parse failures on weak references (Appendix E.4, Table E.9). We adopt the convex form as the default. Two hyperparameters remain: τ_A controls reliability-check

Table 3. Main results. nAUC (%, higher is better) under strict 0–5 pp and relaxed 0–10 pp faithfulness budgets. **Bold** marks the best within each model and metric; *italics* mark second-best. Superscripts denote paired-bootstrap significance of GACL over the stronger of AAD/ACD on the same row (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Model	Method	AQA		VSC		SER		ALME		Average	
		0–5	0–10	0–5	0–10	0–5	0–10	0–5	0–10	0–5	0–10
Qwen2-Audio-Inst.	AAD	12.1	18.9	18.4	29.4	2.3	3.3	39.5	43.7	18.0	23.9
	ACD	8.1	9.1	5.6	10.4	0.0	0.0	21.5	26.8	8.8	11.6
	GACL	**51.0*****	**60.7*****	**39.1*****	**51.2*****	**2.9**	**6.1***	**60.4****	**63.4*****	**38.3*****	**45.4*****
Qwen2.5-Omni	AAD	14.8	28.1	48.4	56.6	3.4	6.8	54.8	56.4	30.4	37.0
	ACD	0.5	0.5	29.4	37.0	3.9	5.7	32.8	41.2	16.7	21.1
	GACL	**31.6****	**45.7****	**58.0**	**71.6*****	**7.6****	**11.4****	**66.0**	**66.0***	**40.8****	**48.7*****
Voxtral-Small	AAD	0.0	2.8	7.1	8.5	2.8	4.4	17.9	29.4	7.0	11.3
	ACD	1.7	3.1	0.5	0.5	2.3	2.9	2.1	7.0	1.6	3.4
	GACL	**8.8****	**17.4*****	**25.0*****	**28.1*****	**6.7****	**10.3*****	**72.0*****	**76.0*****	**28.1*****	**33.0*****
Qwen3-Omni	AAD	11.8	20.1	7.0	14.8	0.0	0.0	44.7	45.0	15.9	20.0
	ACD	8.1	8.2	0.7	0.9	0.8	1.2	5.2	5.2	3.7	3.9
	GACL	**30.1**	**38.6***	**36.2***	**42.5*****	**6.6**	**10.2**	**64.0*****	**64.0*****	**34.2*****	**38.8*****
Kimi-Audio-Inst.	AAD	10.5	17.2	12.3	**24.8**	0.5	0.7	2.0	4.2	6.3	11.8
	ACD	5.4	8.6	1.7	3.2	0.8	1.1	2.2	4.1	2.5	4.2
	GACL	**39.3*****	**48.7*****	**12.7**	19.2	**2.0**	**4.9****	**47.0*****	**47.0*****	**25.2*****	**30.0*****

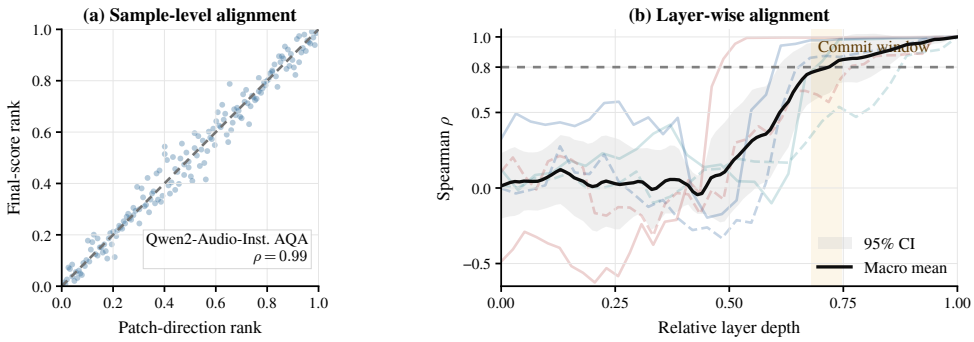

Figure 4. The internal repair direction is visible in final scores. (a) Per-sample rank alignment between the patch-induced direction and $s_A - s_J$ on Qwen2-Audio-Inst.×AQA (macro across twelve configurations ρ=0.93), supporting $s_A - s_J$ as a training-free proxy for the hidden repair direction. **(b)** Layer-resolved alignment starts near zero and rises through the commit window (shaded). Appendix C gives per-configuration curves.

strictness, and λ caps correction strength. Both are tuned on the development set for each model–task pair and then fixed for test. A single global setting, $\lambda = 0.5$ and $\tau_A = 0.5$, still retains 91.2% of tuned nAUC@0–10 (see (4) for the nAUC definition) and keeps 19/20 model–task settings within the 5 pp budget (Appendix D.4). At inference, GACL requires only two branch forward passes and changes neither parameters nor hidden states.

5 Experiments

The previous section justified GACL mechanistically. We now evaluate whether it works in practice: how it compares to contrastive-decoding baselines under a faithfulness budget, when rescue becomes reference-limited, which components are necessary, and how it relates to fine-tuning and other modality pairs.

Setup. We use the same five ALMs and four tasks as in §3 (AQA, VSC, SER, ALME; details in Appendix A.1, Appendix A.2). Because deployed systems trade conflict accuracy against consistent-input accuracy, we evaluate methods on a *rescue–faithfulness frontier*: for each faithful-drop cap K (in pp), $R(K)$ is the highest conflict gain within that cap. We report normalized AUC (nAUC) over $K \in [0, 5]$ pp (strict) and $K \in [0, 10]$ pp (relaxed):

$$\mathrm{nAUC}(K_{\max}) = \frac{1}{K_{\max}} \int_0^{K_{\max}} R(K)\, dK, \qquad R(K) = \max_{\theta:\, \Delta_{\mathrm{faith}}(\theta) \le K} \Delta_{\mathrm{rescue}}(\theta), \tag{4}$$

where θ denotes method hyperparameters, Δ_{rescue} is conflict-accuracy gain over the joint baseline,

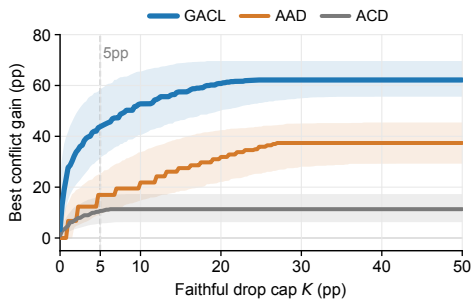

Figure 5. GACL dominates in the low-drop regime. Max conflict gain versus faithful-drop cap K, macro-averaged across five models and four tasks; bands are 95% bootstrap CIs. Dashed line: main 5 pp budget.

Table 4. Versus fine-tuning. Retention is (GACL − Joint)/(SFT − Joint), averaged across tasks.

Task	SFT adv. acc.↑	SFT faith. drop↓	GACL adv. acc.↑	Retention↑
AQA	92.1	10.3	86.0	93%
VSC	63.4	1.0	55.5	85%
SER	40.5	21.7	20.5	51%
Macro	65.3	11.0	54.0	76%

and Δ_{faith} is faithful-accuracy drop, both in percentage points. We integrate at 0.5 pp increments via the trapezoidal rule. Higher nAUC means more conflict-sample rescue per unit of faithful-accuracy cost.

Baselines are the joint model, AAD (Hsu et al., 2025) (no-audio reference), and ACD (Lin et al., 2026) (perturbed-audio reference). At each budget K, hyperparameters are selected on dev to maximize rescue under the K-pp faithful-drop constraint, then frozen for test.

5.1 Main results

GACL is best on 39 of 40 model–task–budget combinations (Table 3). Across model–task pairs, GACL exceeds the stronger of AAD/ACD by 17.8 nAUC at 5 pp and 18.4 at 10 pp; it outperforms hard budgeted reference selection on 14/20 settings and yields larger matched-drop gains than audio-priority prompting (+34.1 vs. +6.7 pp; Appendix E.5–Appendix E.6).

The advantage appears early on the frontier. At $K = 5$ pp, GACL has already surpassed AAD's eventual plateau near $K \approx 27$ pp; the gain is not that GACL extends farther, but that it rescues before the budget is exhausted. Figure 5 shows this low-drop advantage; Figure 6 gives the geometric view: bounded interpolation moves repairable samples toward the audio-supported side while leaving the rest of the joint distribution intact.

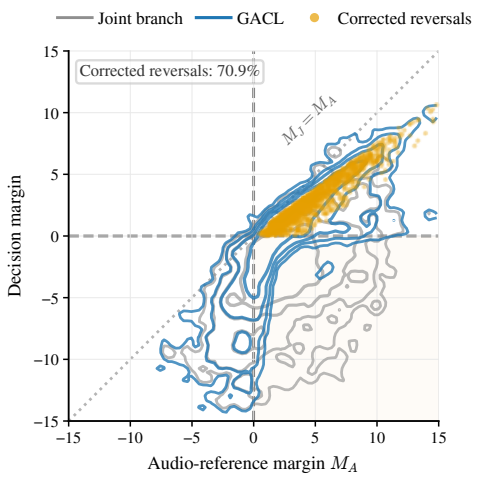

Figure 6. GACL moves repairable samples toward the audio side. Joint distribution in gray, GACL in blue; yellow points have decision margins crossing zero. Bounded interpolation pulls $M_J < 0$ samples toward M_A without overshooting.

5.2 Reference-limited rescue

Each rescued sample passes through three stages: it must offer a repairable opportunity ($M_A > 0, M_J < 0$), the gate must fire, and gated interpolation must flip the top-1 prediction. Decomposing rescue along this chain separates missed opportunities from weak reference signals.

Outside SER, all three stages hold: opportunity, gate activation, and top-1 conversion are 66.3 / 99.3 / 90.9%, respectively. The gate captures almost every opportunity, and most fired cases convert. On SER, opportunity (56.6%) and gate activation (94.2%) remain high, but conversion drops to 48.6%. The reason is that the A-branch is reliable on only 35.3% of SER inputs, versus 88.3% elsewhere. SER is therefore reference-limited, the same regime that produced 0.51 rather than 0.93 answer-side rescue in §4.1: when the reference is weak, R_A dampens correction, preventing GACL from amplifying an unreliable signal.

ALME is the opposite extreme: the A-branch is reliable (96.5%) and the joint branch heavily suppressed (34.4%), so GACL recovers more than 60 nAUC. Together, SER and ALME define the operating range: GACL converts repairable opportunities when the reference supports an answer, and stays conservative when it is weak.

5.3 Ablation

GACL has three mechanism-derived choices: same-audio direction $s_A - s_J$, gates $N_{\text{out}} \cdot R_A$, and convex bound $\alpha \leq 1$. Table 5 shows each guards against a distinct failure mode. Naive in-

Table 5. Each component of GACL guards against a distinct failure mode. Removing any single component degrades at least one stress metric by more than $1.5\times$ relative to the full method (red shading), whereas GACL satisfies all budget constraints. Macro-averaged across the 5×4 model–task grid.

Variant	*Budget feasibility*		*Low-R_A stability*	*Surface stability*	*Generation stability*	Avg.
	MCQ@5 pp drop ↓	MCQ@2.5 pp drop ↓	Faith. drop ↓	FF preserve ↑	SER parse fail ↓	gain ↑
Joint (no correction)	0.0	0.0	0.0	100.0	0.0	0.0
Hard switch (endpoint)	13.3	*infeasible*	—	—	—	66.5
Naive interp. (fixed α)	4.6	3.9	4.7	70.0	7.0	38.4
w/o bound ($\alpha \leq \infty$)	2.1	1.8	2.1	90.0	**27.0**	46.4
w/o N_{out}	2.1	1.8	2.1	**22.0**	0.0	46.1
w/o R_A	2.3	**2.3**	**3.2**	90.0	0.0	47.4
GACL (full)	**2.1**	**1.8**	**2.1**	**90.0**	**0.0**	**45.8**

Table 6. Cross-modal transfer. GACL applied to vision–text arbitration with fixed hyperparameters. Gate column reports firing rate on conflict / faithful inputs.

Model	Method	Adversarial		Faithful		Gate
		Acc ↑	Δ ↑	Acc ↑	Drop ↓	conf.↑/faith.↓
Qwen3-VL-2B	Joint baseline	43.0	–	100.0	0.0	–
Qwen3-VL-2B	**GACL (fixed)**	**83.5**	**+40.5**	**100.0**	**0.0**	56.5%/0.5%
Qwen3-VL-8B	Joint baseline	55.5	–	100.0	0.0	–
Qwen3-VL-8B	**GACL (fixed)**	**82.0**	**+26.5**	**99.5**	**0.5**	42.5%/2.0%

terpolation already captures most rescue headroom (38.4 avg.); the gates and bound (Table 5) turn this raw signal into a safe rule satisfying every stress constraint while retaining comparable rescue. The targeted stress columns identify which component controls each failure surface. R_A suppresses correction when the reference is unreliable; removing it raises the low-R_A faithful drop from 2.1 to 3.2 pp and violates the tighter 2.5 pp budget. N_{out} blocks corrections that rewrite surface form rather than the answer; removing it drops free-form preservation from 90.0% to 22.0%. The convex bound controls the opposite failure: allowing $\alpha > 1$ slightly increases average rescue but produces 27.0% SER parse failures, indicating that unbounded interpolation can push generations off the candidate manifold. Hard switching to the endpoint rescues more samples but violates the 5 pp budget and is infeasible under 2.5 pp. Bounded interpolation with both gates is the only configuration satisfying all stress constraints.

5.4 Relation to fine-tuning and cross-modal transfer

Versus task-specific fine-tuning. A natural concern is whether supervised fine-tuning could obtain the same effect. Against a LoRA baseline (Hu et al., 2022), GACL retains 76% of the adversarial gain in macro without parameter updates (Table 4); lower SER retention again reflects its weak A-branch. This suggests that supervised gain on these tasks largely comes from evidence already in the audio branch, which GACL reads out at inference when the reference is reliable.

Cross-modal transfer. The diagnostic motivating GACL is not audio-specific: a same-modality reference can be held against a joint input, and a reliability gate can decide when to trust it. To test this, we apply GACL unchanged to MC^2 vision–text arbitration, using the image-only branch as reference and fixed defaults $(\lambda, \tau_A) = (0.5, 0.5)$ without tuning on MC^2 (Table 6). GACL achieves +40.5 pp for Qwen3-VL-2B and +26.5 pp for Qwen3-VL-8B at near-zero faithful cost. Gate rates are low on faithful inputs, indicating the gain comes from conditional intervention rather than always prioritizing the reference branch. This transfer suggests that the failure structure documented for audio–text arbitration may extend to other modality pairs.

6 Conclusion

For many audio–text conflicts, the same-audio reference favors the audio answer while the joint branch follows text; activation patching localizes this reversal to the answer-side residual stream, and its effect projects coherently onto final candidate scores (ρ=0.93). This projection supports output-space correction without parameter updates or internal access. GACL implements it as gated convex interpolation, recovers 76% of the adversarial gain of a strong fine-tuning baseline in macro, and transfers without retuning to vision–text arbitration (up to +40.5 pp at near-zero faithful cost). Some multimodal arbitration failures share an output-readable structure that mechanism-grounded decoding can exploit.

Limitations

Our study is designed for controlled, interpretable analysis of audio–text arbitration, which also limits its scope. The conflict instances isolate competing text effects by keeping the audio evidence fixed. This helps separate perceptual from arbitration failures, but naturally occurring conflicts may involve noisier transcripts, partial notes, retrieval snippets, or broader conversational context. GACL is a repair for arbitration, not a substitute for audio perception. It can recover audio evidence that remains available but is displaced during answer selection, but cannot create acoustic capabilities the model has not encoded. This boundary is useful because same-audio references can indicate whether a failure is better addressed by decoding-time repair or improved acoustic modeling. GACL also requires an additional reference forward pass, adding latency despite the lightweight logit update. Reducing this cost through cached states, selective reference calls, or adaptive gating, and extending the same counterfactual analysis to more natural conflict sources and modality pairs, are useful directions for future work.

Ethical Considerations

We use MCR-Bench and ALME as released by their authors. ALME is based on public Common Voice speech (Ardila et al., 2020); our study adds no new audio collection or personally identifying information. All evaluations use public model checkpoints documented in Appendix A.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Jayadev Billa. 2026. When audio-LLMs don't listen: A cross-linguistic study of modality arbitration. *Preprint*, arXiv:2602.11488. Preprint; no formally published version found as of 2026-05-22.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-Audio technical report. *Preprint*, arXiv:2407.10759.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313, Miami, Florida, USA. Association for Computational Linguistics.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Tzu-Wen Hsu, Ke-Han Lu, Cheng-Han Chiang, and Hung-Yi Lee. 2025. Reducing object hallucination in large audio-language models via audio-aware decoding. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

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

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-Audio technical report. *Preprint*, arXiv:2504.18425.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio

flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25125–25148. PMLR.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Yuepei Li, Kang Zhou, Qiao Qiao, Bach Nguyen, Qing Wang, and Qi Li. 2025. Investigating context faithfulness in large language models: The roles of memory strength and evidence style. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4789–4807, Vienna, Austria. Association for Computational Linguistics.

Tzu-Quan Lin, Wei-Ping Huang, Yi-Cheng Lin, and Hung-yi Lee. 2026. How contrastive decoding enhances large audio language models? *Preprint*, arXiv:2603.09232. Preprint; submitted to INTERSPEECH 2026 according to arXiv metadata.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. Voxtral. *Preprint*, arXiv:2507.13264.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.

nostalgebraist. 2020. Interpreting GPT: The logit lens. LessWrong blog post.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401.

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. 2025. When audio and text disagree: Revealing text bias in large audio-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4878–4888, Suzhou, China. Association for Computational Linguistics.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853, Bangkok, Thailand. Association for Computational Linguistics.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024. Simple synthetic data reduces sycophancy in large language models. *Preprint*, arXiv:2308.03958.

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

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, and 19 others. 2025b. Qwen3-Omni technical report. *Preprint*, arXiv:2509.17765.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. 2025. Evaluating and steering modality preferences in multimodal large language model. *Preprint*, arXiv:2505.20977.

A Reproducibility

A.1 Models and benchmarks

Models. Experiments use frozen public checkpoints (Table A.1); released metadata gives full Hugging Face snapshot hashes.

Table A.1. Model checkpoints. *Rev* gives the first 8 hash characters.

Model	HF repo	Rev	Scale
Qwen2-Audio-Inst	`Qwen/Qwen2-Audio-7B-Instruct`	0a095220	7B
Qwen2.5-Omni	`Qwen/Qwen2.5-Omni-7B`	ae9e1690	7B
Voxtral-small	`mistralai/Voxtral-Small-24B-2507`	70b3d450	24B
Qwen3-Omni	`Qwen/Qwen3-Omni-30B-A3B-Instruct`	26291f79	30B
Kimi-Audio-Inst	`moonshotai/Kimi-Audio-7B-Instruct`	9a82a84c	7B

Benchmarks. MCR-Bench provides audio-centered conflicts for AQA, VSC, and SER. ALME (Audio-LLM Modality Evaluation) edits a transcript span from Common Voice (Ardila et al., 2020) audio and synthesizes the conflicting variant via TTS, yielding controlled semantic conflicts between speech and text. We use the released dev split for decoding-hyperparameter selection and the fixed test split for reporting (Table A.2); released training splits are provenance only.

Table A.2. Dataset splits. Training is provenance only; we do not train on it.

Task	Source	Candidates	Train/Dev/Test
AQA	MCR	{yes, no}	600 / 200 / 200
VSC	MCR	vocal-sound labels	600 / 200 / 200
SER	MCR	emotion labels	600 / 200 / 200
ALME-English	ALME	two option labels	600 / 200 / 200

A.2 Prompts and branch semantics

Branches share the audio item, task question, candidate set, and output format, differing only in evidence availability. REF keeps audio, removes competing text; Text-only keeps adversarial text, removes audio; Joint-conflict keeps both; Faithful-joint keeps audio with aligned text. Outputs are exact task labels (AQA/VSC/SER) or a single option letter (ALME).

AQA

Candidates: {yes, no}. **Example.** Q: "does only one guy speak?" Conflict text: "Only one guy speaks." Faithful text: "More than one guy speaks."

VSC

Candidates: {Laughter, Sigh, Cough, Throat clearing, Sneeze, Sniff}. **Example.** Text claims *sigh* for a throat-clearing clip.

SER

Candidates: {sad, happy, fearful, angry, surprised, disgusted, neutral}. **Example.** Text claims *sad* for a fearful clip.

ALME-EN

Candidates: two option labels. **Example.** Conflict: "Most of the artifacts would *not* remain with Yale." Faithful: "Most of the artifacts would remain with Yale."

A.3 Candidate scoring and validation

For candidate c in fixed set $\mathcal{C}$ under branch B,

$$r_B(c) = \sum_k \log p_B(v_{c,k} \mid v_{c,<k}),$$

$$s_B(c) = r_B(c) - \log \sum_{c' \in \mathcal{C}} \exp r_B(c').$$

where branch prediction is $\hat{y}_B = \arg\max_{c \in \mathcal{C}} s_B(c)$. Prefix-trie constraints keep branch scores directly comparable across Joint, AAD, ACD, and GACL.

Tunable hyperparameters are selected on the held-out dev set under a *faithful-drop budget*: maximize conflict accuracy subject to faithful drop ≤ 5 pp, breaking ties by lower faithful drop, then by smaller correction strength. Values are frozen before test evaluation.

B Additional Behavioral Results

B.1 Per-configuration matrix

The behavioral matrix appears in Table B.1. *Text-follow* is the fraction of joint-conflict predictions matching adversarial text; *REF* is audio-reference accuracy; $(+,-) = \Pr(M_A > 0, M_J < 0)$ is the repairable share where audio supports the gold answer but the joint branch flips. Faithful and conflict columns give respective split accuracies.

B.2 Margin scatter grid

The margin grid appears in Figure B.1.

Table B.1. Per-configuration behavioral diagnostics (%). Joint-conflict predictions mostly follow text (Text-follow ≫ Conflict), while audio-reference accuracy stays high, forming the **repairable gap** that GACL targets.

Model	Task	Faithful	Conflict	Text-follow	REF	$(+,-)$
Qwen2-Audio-Inst.	AQA	100.0	3.5	96.5	88.0	86.0
	VSC	100.0	8.0	92.0	87.5	82.0
	SER	100.0	0.0	100.0	46.5	52.5
	ALME	97.5	37.5	62.5	96.0	57.0
Qwen2.5-Omni	AQA	99.0	30.0	70.0	91.0	68.0
	VSC	99.5	11.0	88.0	89.5	93.0
	SER	99.5	1.5	98.0	47.0	50.5
	ALME	99.5	43.0	57.0	98.5	62.5
Voxtral-Small	AQA	100.0	2.0	98.0	72.0	64.5
	VSC	100.0	1.5	98.5	56.0	68.0
	SER	98.5	1.5	97.5	55.0	51.0
	ALME	96.0	28.0	72.0	96.5	66.0
Qwen3-Omni	AQA	97.0	46.0	54.0	93.0	46.5
	VSC	99.5	48.0	49.5	90.0	55.0
	SER	94.0	22.0	64.5	51.5	59.0
	ALME	99.0	25.0	75.0	99.0	55.0
Kimi-Audio-Inst.	AQA	100.0	9.5	90.5	81.0	78.5
	VSC	84.5	3.5	80.0	56.5	67.5
	SER	100.0	0.0	100.0	44.5	70.5
	ALME	93.5	38.5	61.5	92.5	49.0
Macro		**97.9**	**18.0**	**80.3**	**76.6**	**64.1**

C Mechanism Evidence

This appendix collects causal and representational evidence for the mechanistic claims in §4, organized as component decomposition, bidirectional temporal sweeps, output readout, donor/subset controls, and conflict-text reverse patching.

Coverage. Answer-side residual patching audits all five models in Table 2. Cross-position decomposition covers Qwen2-Audio-Inst., Qwen2.5-Omni, and Voxtral-Small on AQA, VSC, SER, and ALME. Temporal sweeps cover all five models on the AQA/VSC grid in Figure C.2.

C.1 Attention/MLP component decomposition

Component-decomposition curves appear in Figure C.1.

C.2 Bidirectional temporal sweeps

The full bidirectional sweep grid appears in Figure C.2.

Rescue and break transition depths L_{80} differ by 1.4 layers on average and at most 5.9 (Qwen3-Omni×VSC). Mismatch-donor rescue stays below 0.03 throughout, except for a localized single-layer spike of 0.252 on Qwen3-Omni×VSC that decays to 0.026 by the final layer.

C.3 Logit-lens trajectories and linear probing

Layer-alignment and logit-lens diagnostics appear in Figure C.3 and Figure C.4.

The linear-probe table summarizes a diagnostic probe on Qwen2-Audio-Inst.×VSC: cross-branch transferability (CrossA2J) drops sharply through the commit window while within-branch decodability stays flat, matching patching. Probes are not used by GACL or any intervention. The displacement table reports the three-way comparison across audited configurations.

C.4 Donor, span, and subset controls

The donor-specificity plot and donor-control table appear in Figure C.5 and Table C.3.

Mismatch donors are different-sample, different-label; random donors are different-sample, label-unrestricted. The audit table checks the random column: positive AQA rescue comes *entirely* from same-label draws, while different-label random donors contribute zero. Transfer is

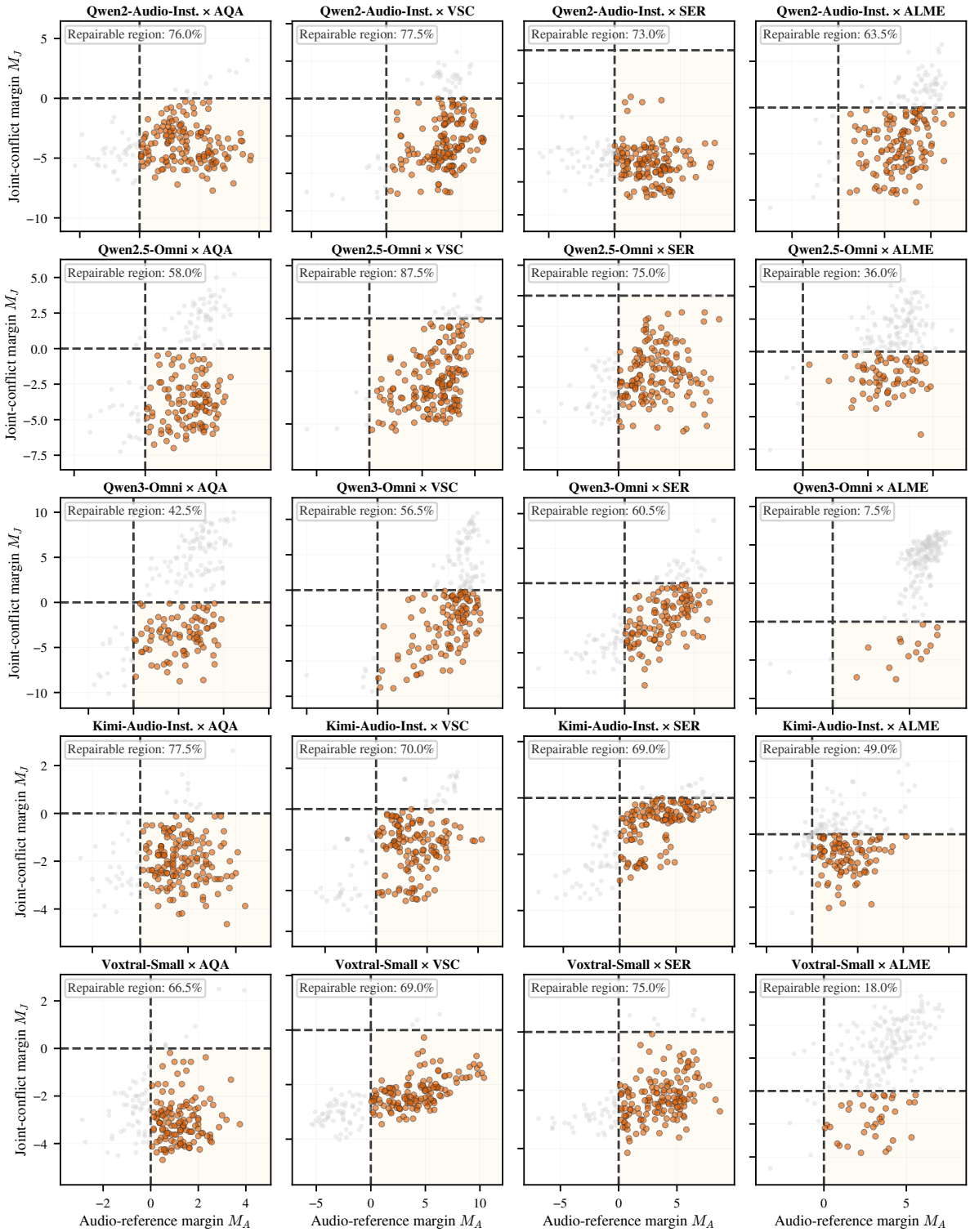

Figure B.1. (M_A, M_J) scatter for all 5×4 configurations. The lower-right quadrant is the *repairable* regime ($M_A > 0, M_J < 0$). AQA, VSC, and ALME populate it densely for every model; SER is reference-limited, with lower M_A density on the right half-plane, consistent with the REF accuracy column in Table B.1.

therefore label-compatible; unconstrained random sampling is not a valid null.

Span and subset controls. An audio-span sweep (Center-1, -2, -4, -8, Full) yields zero rescue in this configuration, while an answer-boundary positive control reaches 0.80. Audio-entry writeability tests and raw50 vs. Q_4 subset comparisons confirm the audio-span null is not a writeability artifact and the temporal pattern is not driven by

Table C.1. Linear probing on Qwen2-Audio-Inst.×VSC. Decodability ($\text{Decode}_{A/J}$) stays flat with depth; cross-branch transfer (CrossA2J) collapses through the commit window, marking the readout-reorganization range.

Layer block	Decode_A	Decode_J	Gap	CrossA2J
Early (L0–L7)	0.878	0.871	+0.007	0.792
Mid-early (L8–L15)	0.879	0.871	+0.008	0.337
Mid-late (L16–L23)	0.906	0.891	+0.016	0.115
Late (L24–L31)	0.903	0.911	−0.008	0.087
All layers (L0–L31)	0.892	0.886	+0.006	0.333

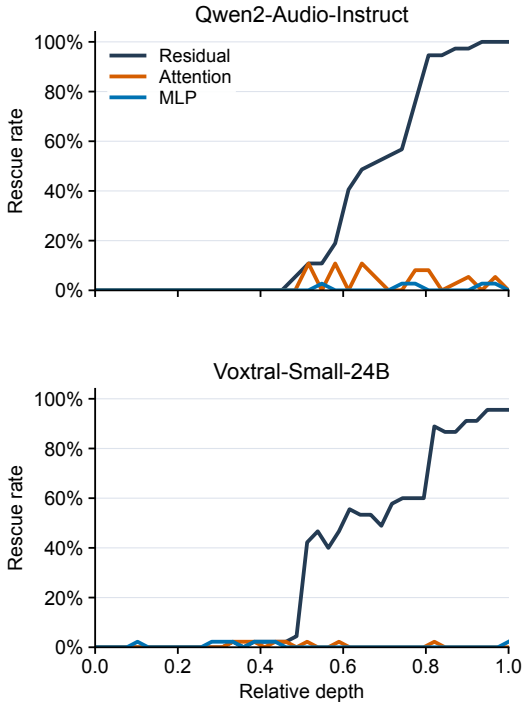

Figure C.1. Single components are insufficient. Per-layer rescue when patching attention output *or* MLP output alone vs. the full residual, on Qwen2-Audio-Inst.×VSC (top) and Voxtral-Small×VSC (bottom). Single components recover little of the full-residual effect: the repairable state is a multi-layer residual write, not a single-component write.

selecting only high-margin repairable cases.

C.5 Conflict-text reverse patching

The conflict effect is injected at conflict-token positions in early layers (L0–L8) and propagates to the answer-boundary position over the commit window, where it consolidates and dominates the final prediction. Mismatch donors at conflict positions can produce non-trivial rescue (0.82/0.43 at L0/L8) but none at answer-boundary positions despite large output-distribution shifts; only same-sample donors carry the correct answer direction. Thus answer-side patching in §4 is the downstream readout of an earlier conflict-token pathway.

Table C.2. Displacement comparison. Final-score displacement uniformly tracks patched-state displacement ($\rho = 0.93$ macro) and yields near-saturated rescue; raw lens does not. Final scores are therefore the correct inference-time proxy.

Model	Task	Patch–final ρ	Patch–lens ρ	Final rescue	Lens rescue
Qwen2-Audio-Inst.	AQA	0.983	0.895	100.0	100.0
Qwen2-Audio-Inst.	VSC	0.932	0.917	91.1	92.2
Qwen2-Audio-Inst.	SER	0.925	0.422	80.7	76.1
Qwen2.5-Omni	AQA	0.987	0.787	100.0	100.0
Qwen2.5-Omni	VSC	0.890	0.770	97.0	98.0
Qwen2.5-Omni	SER	0.983	0.747	69.8	71.7
Voxtral-Small	AQA	0.991	0.818	100.0	53.1
Voxtral-Small	VSC	0.812	0.242	93.2	1.1
Voxtral-Small	SER	0.738	0.415	93.6	59.6
Kimi-Audio-Inst.	AQA	0.978	0.634	100.0	1.0
Kimi-Audio-Inst.	VSC	0.941	0.792	98.0	3.0
Kimi-Audio-Inst.	SER	0.976	0.643	84.6	0.0
Macro		**0.928**	**0.674**	**92.3**	**54.7**

Table C.3. Donor-control lift over baseline rescue. Same donors yield strong positive lift; mismatch donors yield zero or negative lift. $*$: random lift is inflated by same-label draws (Table C.4).

Scope	Same	Mismatch	Random
Qwen2-Audio-Inst.×AQA	+0.950	−0.045	+0.500*
Qwen2-Audio-Inst.×VSC	+0.830	−0.165	+0.005
Qwen2.5-Omni×AQA	+0.670	−0.295	+0.220*
Qwen2.5-Omni×VSC	+0.920	−0.080	+0.090
Voxtral-Small×AQA	+0.970	−0.025	+0.505*
Voxtral-Small×VSC	+0.985	−0.015	+0.155

Table C.4. Label-compatibility audit of random donors. Different-label rescue is uniformly zero.

Scope	Same-label rescue	Diff-label rescue
Qwen2-Audio-Inst.×AQA	107 / 110	0 / 90
Qwen2-Audio-Inst.×VSC	26 / 34	0 / 166
Qwen2.5-Omni×AQA	81 / 110	0 / 90
Qwen2.5-Omni×VSC	29 / 34	0 / 166
Voxtral-Small×AQA	103 / 110	0 / 90
Voxtral-Small×VSC	34 / 34	0 / 166

Table C.5. Conflict-text reverse patching, representative layers. Rescue Δ is lift over joint-conflict; Flip is the fraction of top-1 predictions that change. The effect site is conflict-position at L0 and answer-boundary at L29—a two-stage causal pathway.

Site	L	Rescue Δ	Flip	KL	TV
conflict-pos.	0	1.000	100.0%	5.676	0.947
conflict-pos.	16	0.000	0.0%	0.002	0.010
conflict-pos.	29	0.000	0.0%	0.000	0.000
answer-boundary	0	0.000	0.0%	0.000	0.001
answer-boundary	16	0.071	7.1%	0.075	0.084
answer-boundary	29	1.000	100.0%	5.609	0.941

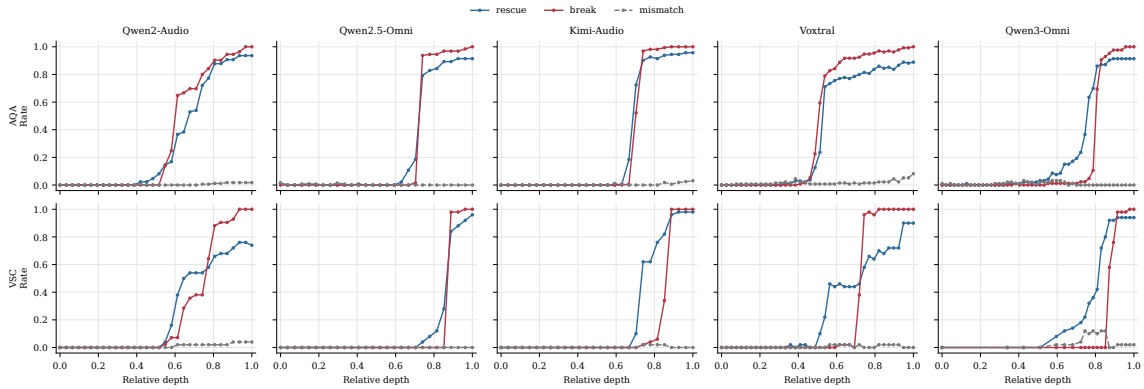

Figure C.2. Bidirectional sweeps identify the commit window. 5×2 grid (five models, two tasks). *Blue*: rescue (per-layer answer-side residual patching). *Red*: break (audio-reference perturbed toward joint-conflict). *Dashed gray*: mismatch-donor control. Rescue and break onsets align across configurations.

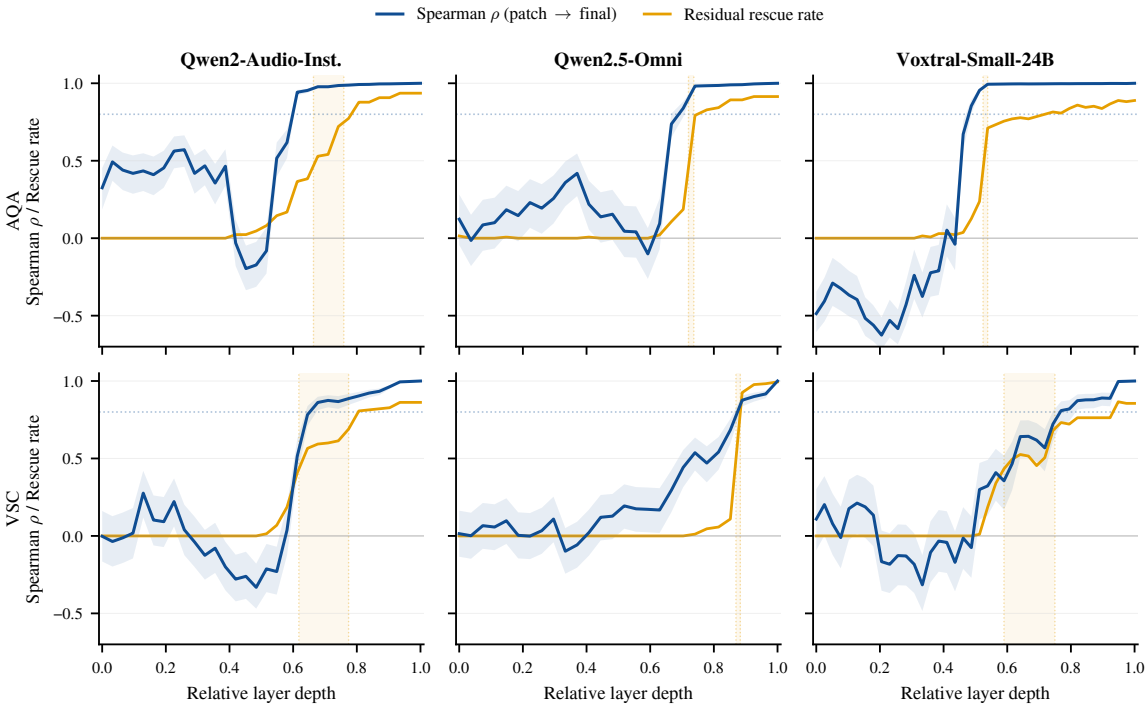

Figure C.3. Per-layer alignment between patch displacement and final-branch displacement. Correlation rises where answer-side patching becomes effective, supporting final candidate scores as the inference-time proxy for the localized internal state.

D GACL Implementation Details

D.1 Default protocol and pseudocode

Defaults. Reference branch: same audio with conflict text removed or neutralized. Update: bounded interpolation between joint and audio-reference logits/scores with $\alpha_{\max} = 1$. Gate: sequence-level $N_{\text{out}} \cdot R_A$, computed once per example. Tuned parameters: λ, τ_A, selected on dev under a 5 pp faithful-drop budget, then frozen for test.

Table D.1. GACL closed-set inference.

Input: x, candidate set $\mathcal{C}$, λ, τ_A	
1.	Compute $s_J(c)$ for all $c \in \mathcal{C}$ (joint-conflict).
2.	Compute $s_A(c)$ for all $c \in \mathcal{C}$ (audio-reference).
3.	Compute $N_{\text{out}}(x)$ and $R_A(x)$ from branch outputs/scores.
4.	$\alpha(x) \leftarrow \text{clip}\big(\lambda\, R_A(x)\, N_{\text{out}}(x),\ 0,\ 1\big)$.
Output: $\arg\max_{c \in \mathcal{C}} \big[s_J(c) + \alpha(x)(s_A(c) - s_J(c))\big]$	

Pairwise flip condition. For audio-supported y_a and text-supported y_t, bounded interpolation gives $M_G = M_J + \alpha(x)(M_A - M_J)$. When $M_A > 0$

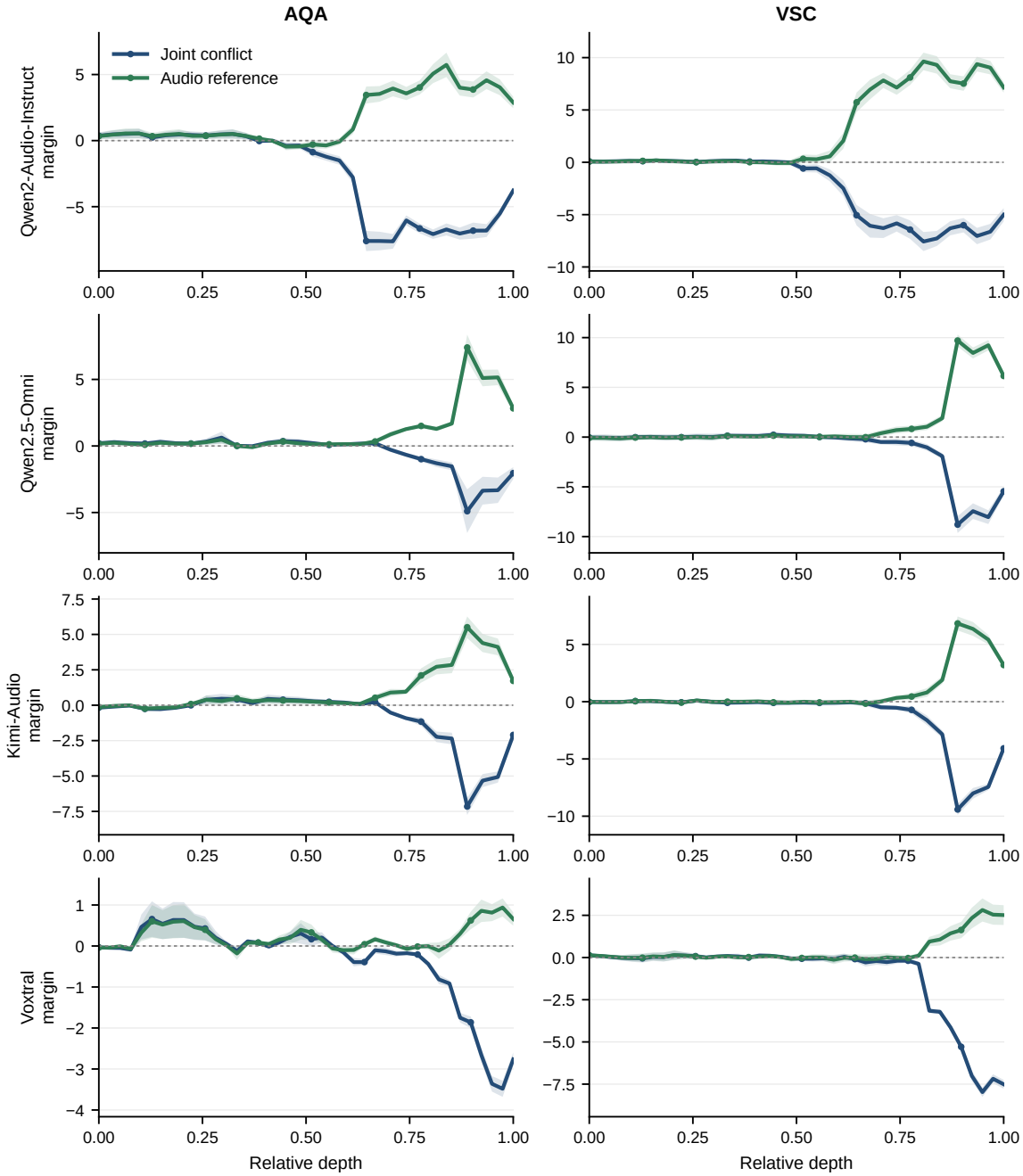

Figure C.4. Logit-lens trajectories. The audio answer remains decodable through late layers, but the branch-specific readout subspace reorganizes after the commit window. The lens is diagnostic; final-score displacement is the stable proxy (Table C.2).

and $M_J < 0$ the pairwise margin flips iff

$$\alpha(x) > \frac{|M_J|}{M_A + |M_J|}.$$

D.2 Interface-specific details

Answer normalization π. For constrained short-label generation, π maps valid generated labels to canonical task labels; unparsable outputs map to $\emptyset$. For free short-answer generation, π applies evaluator canonicalization: Unicode normalization, case folding (where labels are case-insensitive), whitespace/quote/markdown stripping, trailing-punctuation removal, and alias folding. Exact-surface preservation in Appendix E.4 is computed *before* alias folding, so it is stricter than canonical accuracy. For MC^2, π maps to one of two candidates or returns $\emptyset$.

Hypothesis set $\mathcal{H}(x)$. For constrained labels, $\mathcal{H}(x)$ is the task label set (and trie). For MC^2,

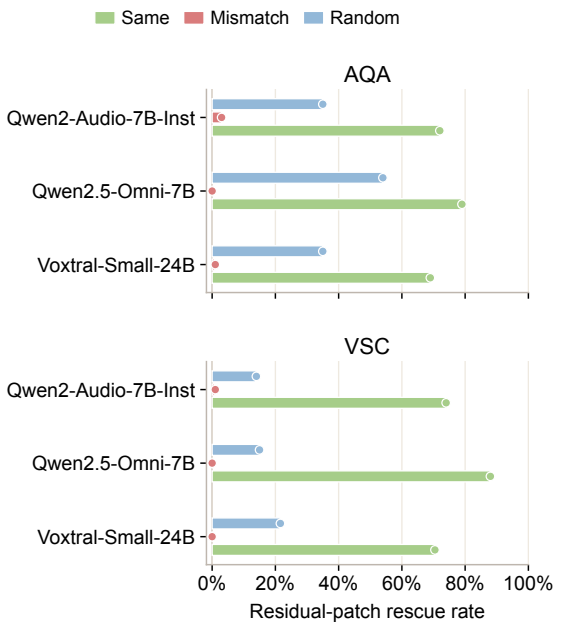

Figure C.5. Donor specificity. Same-sample donors produce high rescue; mismatch and (filtered) random donors do not. Rescue depends on answer-compatible arbitration states, not nonspecific injection.

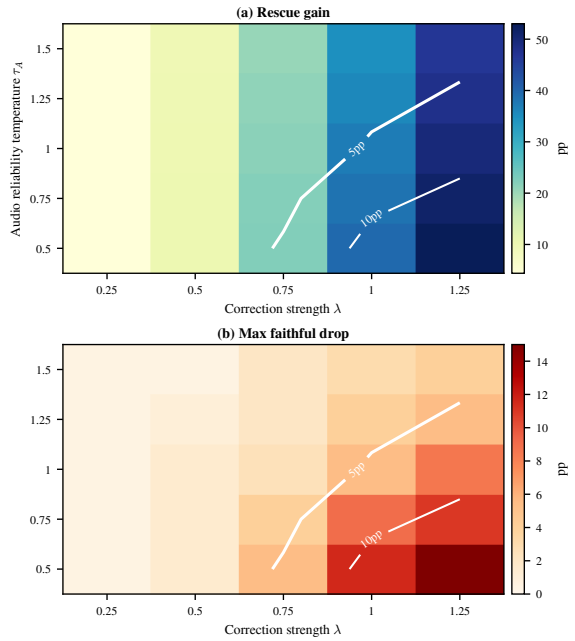

Figure D.1. $\lambda \times \tau_A$ landscape for rescue and faithful drop on a representative configuration.

$\mathcal{H}(x)$ is the two candidates. For free generation, $\mathcal{H}(x)$ is built online from valid normalized branch answers among $\{\hat{y}_J, \hat{y}_A\}$; if fewer than two elements, $\Delta_A(x)$ is undefined and the gate is silenced. This keeps $\Delta_A(x)$ computable on open-ended interfaces without enumerating output space.

Score comparability. For each $y \in \mathcal{H}(x)$, the same verbalizer and template are scored under both branches, so $s_J(y)$ and $s_A(y)$ reflect evidence arbitration rather than surface-form differences.

Decoding order. Constrained short-label decoding interpolates branch logits *before* prefix-trie masking. Reversing the order makes post-mask renormalization sensitive to each branch's invalid-token mass. MC^2 uses the same prefix-trie constrained short-label generation, with the image-only branch replacing the audio reference and the two answer options forming the trie.

Gate edge cases. $\hat{y}_A = \emptyset$ or $|\mathcal{H}(x)| < 2 \Rightarrow \alpha(x) = 0$ (joint unchanged). $\hat{y}_J = \emptyset$ with $\hat{y}_A$ valid triggers N_{out}; R_A decides correction strength.

D.3 Oracle-to-observable mapping

Table D.2 quantifies gate-component correspondence: N_{out} estimates correction need, R_A estimates donor reliability, and the product matches the oracle opportunity region $O = \{M_A > 0, M_J < 0\}$ with 98.2% coverage and 87.9% precision.

Oracle / role	Proxy	Cov.	Prec.
$M_A > 0$ (reliability)	$R_A(x)$	79.4	88.9
$M_J < 0$ (need)	$N_{\text{out}}(x)$	84.4	98.9
O (activation)	$R_A(x)N_{\text{out}}(x)$	98.2	87.9
displacement	$s_A - s_J$	–	–

Table D.2. Proxy fidelity. Coverage $= P(\text{proxy} \mid \text{oracle})$, precision $= P(\text{oracle} \mid \text{proxy})$.

D.4 Hyperparameter sensitivity and transfer

The sensitivity landscape appears in Figure D.1.

The free-form $\alpha_{\max}$ cap sweep is deferred to Appendix E.4 because it concerns generation-time validity, not implementation mechanics.

Cross-task transfer. Fixing one global setting (λ=0.5, τ_A=0.5) across all 20 model–task combinations retains 91.2% of per-task-tuned nAUC@0–10 (24.0 vs. 26.3 macro), lowers average faithful drop from 4.4 to 2.9 pp, and stays inside the 5 pp budget on 19/20 configurations (vs. 18/20 under per-task tuning). One conservative pair captures most of the rescue.

D.5 Gate granularity diagnostic

A sequence-level gate (default) outperforms a per-token dynamic gate on free-form VSC: rescue 84.6 vs. 82.1, faithful drop 11.2 vs. 12.5, mean $\alpha = 0.535$ vs. 0.244. Re-evaluating the gate at every

step does not improve faithfulness and costs more.

D.6 Inference cost

GACL adds one forward pass per query (the audio-reference branch). On Qwen2-Audio-Inst.×VSC: 0.28 s/sample vs. 0.15 s/sample for Joint; peak GPU memory nearly unchanged (15.77 GiB allocated; 16.40 vs. 16.39 GiB reserved); per-sample FLOPs roughly doubled (80.1 vs. 38.7 GFLOPs). Sharing the audio encoder output and KV-caching the common prompt prefix could further reduce wall-clock; we leave this to released code.

E Full Evaluation and Ablations

E.1 Per-configuration main results

The full 5×4 evaluation table in Table E.1 gives bootstrap CIs and selected hyperparameters. A negative drop means GACL *improved* faithful accuracy relative to Joint. A single-configuration DoLa (Chuang et al., 2024) check on Qwen2-Audio-Inst.×VSC *decreases* conflict accuracy by 5.0 pp, consistent with mismatch to cross-modal arbitration.

E.2 Reference-Limited SER Decomposition

GACL factorizes gain into (opportunity) × (gate activation) × (donor reliability) × (conversion). SER is the low-donor stress case: 56.6/94.2/35.3/48.6, with final gain 8.9 (empirical mean, not the algebraic product). Non-SER tasks show 66.3/99.3/88.3/90.9, with final gain 52.6. The bottleneck is donor reliability, not opportunity discovery or gate activation.

E.3 Ablations under the faithful-drop budget

Cumulative ablation. Table E.2 shows each component contributes independently: R_A alone destabilizes the surface (FF 22%); adding N_{out} restores it (90%); tuned λ then recovers the rescue gap.

Configuration	Rescue	Drop	FF	ΔR	ΔD	ΔFF
Naive interp. (α=0.55)	38.4	4.7	70.0	–	–	–
+ bound $\alpha \leq 1$	38.6	4.6	70.0	+0.2	−0.1	0.0
+ R_A	38.0	2.5	22.0	−0.6	−2.1	−48.0
+ N_{out}	36.2	2.1	90.0	−1.8	−0.4	+68.0
+ tuned λ (GACL)	**45.8**	**2.1**	**90.0**	+9.6	0.0	0.0

Table E.2. Each component contributes independently. Δ is change vs. the preceding row. FF: free-form VSC exact-surface preservation.

Strict budgets. Table E.3 sweeps budgets: removing R_A appears safe at 5 pp but has measurable cost at 2.5 pp. Hard switch attains higher rescue only with a 45 pp maximum drop.

Budget	Variant	Conflict gain	Faith. drop	Max drop
2.5	GACL	43.7 [41.4, 45.9]	1.8 [1.2, 2.6]	3.0
2.5	w/o R_A	41.1 [38.9, 43.1]	2.3 [1.6, 3.1]	4.0
2.5	w/o N_{out}	44.0 [41.8, 46.3]	1.8 [1.2, 2.5]	3.0
5.0	GACL	45.8 [43.6, 47.8]	2.1 [1.4, 2.7]	4.5
5.0	w/o R_A	47.4 [45.2, 49.6]	2.3 [1.6, 3.0]	4.0
5.0	w/o N_{out}	46.1 [43.7, 48.4]	2.1 [1.4, 2.8]	4.5
5.0	Hard switch	66.5 [64.4, 68.6]	13.3 [12.0, 14.7]	**45.0**

Table E.3. Strict-budget summary with bootstrap CIs (1000 resamples), macro-averaged across short-label configurations.

Targeted R_A effect. Table E.4 shows R_A is *targeted*: it barely changes macro faithful drop but suppresses drop by 1.5–1.9× in the bottom R_A quartile, where audio confidence is weakest.

Subset	Budget	w/ R_A	w/o R_A	Ratio
All	2.5	2.2	2.2	1.0×
Low-R_A	2.5	1.6	**3.0**	1.9×
All	5.0	3.1	3.5	1.1×
Low-R_A	5.0	2.1	**3.2**	1.5×

Table E.4. R_A acts where it matters. Low-R_A = bottom audio-reliability quartile within each model–task on faithful samples.

Per-configuration ablation. Table E.5 shows hard switch gains raw rescue but creates a large faithful-drop tail on several configurations.

Config (gain / drop)	Hard switch	w/o R_A	w/o N_{out}	GACL
Qwen2-Audio×ALME	58.0 / 1.0	50.5 / 0.5	43.0 / 0.5	43.0 / 0.5
Qwen2.5×ALME	57.5 / 0.5	53.0 / 0.5	56.0 / 0.5	54.0 / 0.5
Voxtral×ALME	63.5 / 1.0	53.0 / 1.0	42.5 / 0.5	42.0 / 0.5
Qwen2-Audio×AQA	84.0 / 13.5	60.5 / 3.5	48.5 / 2.5	48.5 / 2.5
Qwen2.5×AQA	67.0 / 6.5	52.5 / 2.5	53.5 / 2.0	53.0 / 2.0
Voxtral×AQA	66.0 / 29.5	10.5 / 4.0	29.0 / 4.5	29.0 / 4.5
Qwen2-Audio×VSC	72.0 / 13.5	57.5 / 3.0	58.0 / 3.0	58.0 / 3.0
Qwen2.5×VSC	77.0 / 9.5	52.0 / 2.5	49.5 / 2.5	49.5 / 2.5
Voxtral×VSC	53.5 / 45.0	37.0 / 3.0	35.0 / 2.5	35.0 / 2.5

Table E.5. Per-configuration gain / drop. Hard switch's faithful-drop tail reaches 45 pp.

E.4 Free-form generation

Component contribution (VSC). The table isolates components on free-form VSC. Removing N_{out} preserves canonical accuracy but rewrites most surface forms (Exact 22%); the disagreement gate protects faithful surfaces.

Table E.1. Per-configuration evaluation. Entries are conflict gain (faithful drop), pp relative to Joint, with 95% paired-bootstrap CIs (1000 resamples). Bold: best gain per row. GACL wins 18/20.

Model	Task	Joint conf.	AAD (drop)	ACD (drop)	GACL (drop)	λ	τ_A
Qwen2-Audio-Inst.	AQA	2.0±2.0	$+10.0_{\pm 4.0}$ ($1.5_{\pm 1.8}$)	$+16.0_{\pm 5.5}$ ($3.0_{\pm 2.2}$)	$\mathbf{+69.5_{\pm 6.2}}$ ($6.5_{\pm 3.5}$)	1.25	0.5
Qwen2-Audio-Inst.	VSC	11.0±4.2	$+28.5_{\pm 6.2}$ ($3.5_{\pm 2.2}$)	$+21.5_{\pm 5.8}$ ($2.5_{\pm 2.0}$)	$\mathbf{+57.5_{\pm 7.2}}$ ($6.5_{\pm 3.2}$)	1.0	1.5
Qwen2-Audio-Inst.	SER	0.0±0.0	$+1.0_{\pm 1.2}$ ($0.5_{\pm 0.8}$)	$+2.0_{\pm 1.8}$ ($1.0_{\pm 1.2}$)	$\mathbf{+4.5_{\pm 2.8}}$ ($3.0_{\pm 2.2}$)	2.0	1.5
Qwen2-Audio-Inst.	ALME	24.5±5.8	$+35.0_{\pm 6.8}$ ($0.0_{\pm 2.0}$)	$+41.0_{\pm 7.2}$ ($3.0_{\pm 3.0}$)	$\mathbf{+66.5_{\pm 6.2}}$ ($1.0_{\pm 2.2}$)	2.0	0.5
Qwen2.5-Omni	AQA	24.5±6.0	$+27.0_{\pm 6.0}$ ($3.5_{\pm 2.5}$)	$+38.0_{\pm 7.0}$ ($6.0_{\pm 3.2}$)	$\mathbf{+41.0_{\pm 6.2}}$ ($4.0_{\pm 2.8}$)	1.0	0.5
Qwen2.5-Omni	VSC	2.5±2.2	$+47.0_{\pm 7.0}$ ($1.5_{\pm 1.8}$)	$+56.0_{\pm 7.2}$ ($2.5_{\pm 2.2}$)	$\mathbf{+79.0_{\pm 6.0}}$ ($4.5_{\pm 2.8}$)	1.25	1.5
Qwen2.5-Omni	SER	0.0±0.0	$\mathbf{+11.0_{\pm 4.2}}$ ($8.0_{\pm 3.5}$)	$+6.5_{\pm 3.5}$ ($4.5_{\pm 2.8}$)	$+10.0_{\pm 4.5}$ ($7.0_{\pm 3.5}$)	0.5	0.5
Qwen2.5-Omni	ALME	30.5±6.2	$+52.0_{\pm 7.2}$ ($0.0_{\pm 2.0}$)	$+54.0_{\pm 7.5}$ ($2.0_{\pm 3.0}$)	$\mathbf{+66.0_{\pm 6.5}}$ ($0.0_{\pm 2.0}$)	2.0	0.5
Voxtral-Small	AQA	3.0±2.2	$+6.5_{\pm 3.2}$ ($5.5_{\pm 3.0}$)	$+0.0_{\pm 1.5}$ ($1.0_{\pm 1.2}$)	$\mathbf{+22.5_{\pm 5.5}}$ ($6.5_{\pm 3.2}$)	1.25	0.5
Voxtral-Small	VSC	1.5±1.8	$+4.0_{\pm 2.8}$ ($0.0_{\pm 0.0}$)	$+1.0_{\pm 1.2}$ ($0.0_{\pm 0.0}$)	$\mathbf{+22.5_{\pm 5.8}}$ ($0.5_{\pm 0.8}$)	0.5	1.0
Voxtral-Small	SER	1.0±1.2	$+6.0_{\pm 3.2}$ ($5.5_{\pm 3.0}$)	$+2.5_{\pm 2.2}$ ($0.5_{\pm 0.8}$)	$\mathbf{+6.0_{\pm 3.2}}$ ($1.5_{\pm 1.8}$)	0.5	1.5
Voxtral-Small	ALME	15.0±4.8	$+39.5_{\pm 7.5}$ ($5.5_{\pm 3.8}$)	$+34.5_{\pm 7.0}$ ($3.0_{\pm 2.5}$)	$\mathbf{+80.0_{\pm 6.0}}$ ($1.0_{\pm 1.2}$)	1.0	0.5
Qwen3-Omni	AQA	43.0±7.0	$+26.5_{\pm 6.2}$ ($4.5_{\pm 3.0}$)	$+21.0_{\pm 5.8}$ ($3.0_{\pm 2.2}$)	$\mathbf{+44.0_{\pm 7.0}}$ ($4.5_{\pm 3.0}$)	1.0	0.5
Qwen3-Omni	VSC	41.0±6.5	$+32.0_{\pm 6.5}$ ($8.0_{\pm 3.8}$)	$+18.0_{\pm 5.5}$ ($1.5_{\pm 1.8}$)	$\mathbf{+48.5_{\pm 7.2}}$ ($9.0_{\pm 3.8}$)	1.25	0.5
Qwen3-Omni	SER	16.5±5.2	$\mathbf{+17.5_{\pm 5.5}}$ ($15.5_{\pm 4.8}$)	$+2.5_{\pm 2.8}$ ($0.0_{\pm 0.0}$)	$+16.0_{\pm 5.2}$ ($8.5_{\pm 3.8}$)	0.5	1.0
Qwen3-Omni	ALME	33.5±6.5	$+35.0_{\pm 7.0}$ ($1.5_{\pm 2.0}$)	$+38.0_{\pm 7.2}$ ($2.5_{\pm 2.5}$)	$\mathbf{+62.5_{\pm 7.0}}$ ($-0.5_{\pm 0.8}$)	1.25	0.5
Kimi-Audio-Inst.	AQA	5.5±3.2	$+17.0_{\pm 5.2}$ ($3.5_{\pm 2.8}$)	$+17.0_{\pm 5.0}$ ($3.5_{\pm 2.8}$)	$\mathbf{+56.0_{\pm 6.2}}$ ($7.5_{\pm 4.0}$)	1.0	1.0
Kimi-Audio-Inst.	VSC	4.5±3.0	$+11.5_{\pm 4.2}$ ($3.0_{\pm 2.2}$)	$+18.0_{\pm 5.5}$ ($4.0_{\pm 2.8}$)	$\mathbf{+27.0_{\pm 6.2}}$ ($8.0_{\pm 3.5}$)	1.25	1.0
Kimi-Audio-Inst.	SER	0.0±0.0	$+0.5_{\pm 0.8}$ ($3.0_{\pm 2.2}$)	$+0.5_{\pm 0.8}$ ($0.0_{\pm 0.0}$)	$\mathbf{+8.0_{\pm 3.8}}$ ($8.0_{\pm 3.8}$)	0.5	0.5
Kimi-Audio-Inst.	ALME	40.0±6.0	$+2.0_{\pm 6.0}$ ($2.5_{\pm 4.8}$)	$+12.5_{\pm 8.8}$ ($11.0_{\pm 7.0}$)	$\mathbf{+47.0_{\pm 7.0}}$ ($-2.0_{\pm 3.5}$)	2.0	0.5

Variant	Rescue	Drop	Parse fail	Other	Exact	Canon.
Hard switch	90.0	10.0	0.0	10.0	—	—
w/o R_A	90.0	10.0	0.0	10.0	90.0	90.0
w/o N_{out}	88.0	8.0	0.0	9.0	**22.0**	92.0
GACL	88.0	8.0	0.0	9.0	**90.0**	92.0

Table E.6. Free-form VSC. Hard switch lacks a logit-interpolation surface, so exact/canonical are undefined.

Cross-task transfer. With closed-set hyperparameters, GACL transfers to free-form AQA, VSC, and ALME with 87–100% exact preservation and near-zero parse failure. SER again acts as the reference-limited stress case: the audio reference fails to canonicalize 36/200 free-form inputs (Table E.8), so the donor that GACL interpolates toward is unreliable on 18% of samples, the free-form analogue of the closed-set regime in §5.3.

Task	Rescue	Drop	Parse fail	Exact	Canon.
AQA	75.5	12.0	0.0	87.4	87.4
VSC	88.0	8.0	0.0	90.0	92.0
ALME	67.0	−0.5	1.2	100.0	100.0
SER†	19.5	55.0	0.0	43.0	45.0

Table E.7. Cross-task free-form transfer on Qwen2-Audio-Inst. with closed-set hyperparameters. †SER's audio reference canonicalizes only 82% of inputs (Table E.8), making the setting reference-limited.

Table E.8. Free-form parse failures (out of 200) on Qwen2-Audio-Inst.

Branch	AQA	VSC	SER	ALME
Audio reference	0	1	**36**	4
GACL	0	0	0	5

Bound ablation under SER stress. Table E.9 isolates the upper cap $\alpha_{\max} = 1$: removing it produces fluent but invalid generations when extrapolation crosses the audio-reference endpoint. The Would exceed column counts samples GACL clips at $\alpha = 1$. Without the cap, clipped samples produce 26–30% parse failures, confirming the cap protects output validity in the reference-limited regime.

Regime	Variant	Acc	Parse fail	Rescue / drop	Realized	Counterfactual w/o bound	
					$\alpha > 1$	Would exceed	Parse fail
Faithful	bounded	45.0	0.0	55.0	0.0	54.0	0.0
	w/o bound	45.0	16.0	55.0	54.0	54.0	29.6
Conflict	bounded	19.5	0.0	19.5	0.0	61.5	0.0
	w/o bound	14.0	16.5	14.0	61.5	61.5	26.8

Table E.9. SER free-form bound ablation. The cap enforces output validity rather than canonical-accuracy gain.

Regime	ID	Gold	Text	Bounded	w/o bound
Faith.	ser_0006	happy	happy	*Angry mood* (angry)	*Every day.* (∅)
Faith.	ser_0527	neutral	neutral	*Surprised and disgusted* (disgusted)	*sound of head impact* (∅)
Conf.	ser_0305	neutral	happy	*The emotion is neutral.* (neutral)	*the critical moment* (∅)
Conf.	ser_0527	neutral	fearful	*Surprised and disgusted* (disgusted)	*Off the top of his head* (∅)

Table E.10. Representative SER extrapolation risks. Without the bound, generation drifts out of the SER label space rather than to a near-miss emotion, confirming the cap protects validity, not accuracy.

E.5 Hard selection baseline

Budgeted reference selection (BRS) sets $\hat{y}(x) = \hat{y}_A$ if $\Delta_A(x) > \theta$, else $\hat{y}_J$, and sweeps θ for a frontier comparable to nAUC. Macro results: at @0–5, BRS 29.5 vs. GACL 33.3 (+3.8; GACL wins 14/20); at @0–10, 36.6 vs. 39.2 (+2.6; 14/20). BRS approaches GACL when audio-reference reliability is uniformly high (e.g., ALME) and lags when the reference is informative but imperfect.

Model	Task	nAUC@0–5			nAUC@0–10		
		BRS	GACL	Δ	BRS	GACL	Δ
Qwen2-Audio-Inst.	AQA	41.3	**51.0**	+9.6	55.3	**60.7**	+5.4
	VSC	16.7	**39.1**	+22.4	36.0	**51.2**	+15.2
	SER	2.7	**2.9**	+0.2	5.3	**6.1**	+0.8
	ALME	**64.2**	60.4	−3.9	**66.1**	63.4	−2.7
Qwen2.5-Omni	AQA	22.6	**31.6**	+8.9	42.4	**45.7**	+3.3
	VSC	57.8	**58.0**	+0.3	70.4	**71.6**	+1.2
	SER	7.5	**7.6**	+0.1	11.0	**11.4**	+0.4
	ALME	65.0	**66.0**	+1.0	65.0	**66.0**	+1.0
Voxtral-Small	AQA	**11.6**	8.8	−2.8	**19.1**	17.4	−1.6
	VSC	14.7	**25.0**	+10.4	19.6	**28.1**	+8.5
	SER	2.7	**6.7**	+4.0	6.6	**10.4**	+3.8
	ALME	**72.8**	72.0	−0.8	**76.9**	76.0	−0.9
Qwen3-Omni	AQA	**30.7**	30.1	−0.7	**38.8**	38.6	−0.1
	VSC	21.4	**36.2**	+14.8	34.7	**42.5**	+7.7
	SER	4.5	**6.6**	+2.0	7.2	**10.2**	+3.0
	ALME	63.5	**64.0**	+0.5	63.5	**64.0**	+0.5
Kimi-Audio-Inst.	AQA	26.0	**39.3**	+13.3	38.4	**48.7**	+10.3
	VSC	**14.6**	12.7	−1.9	**23.3**	19.2	−4.1
	SER	0.6	**2.0**	+1.3	2.9	**4.9**	+2.0
	ALME	**49.0**	47.0	−2.0	**49.0**	47.0	−2.0

Table E.11. GACL vs. budgeted reference selection. GACL wins 14/20 at both budgets; losses concentrate on ALME (uniformly high R_A, where BRS suffices) and a few high-baseline configurations.

E.6 Prompt-level interventions

We evaluate four prompt variants on Qwen2-Audio-Inst. and Qwen2.5-Omni across AQA, VSC, and SER (Table E.12). The strongest variant, audio prioritization, gains +6.7 pp conflict audio-follow over the default; GACL at matched faithful drop gains +34.1 pp. Prompt instructions bias generation but do not intervene on arbitration computation.

Variant	Audio-follow ↑	Δ	Drop
Original prompt	15.0	–	0.0
Audio-priority	21.7	+6.7	0.9
Ignore-text	16.7	+1.7	1.0
Bias-awareness	18.6	+3.6	0.9
Chain-of-thought	17.2	+2.2	1.1
GACL (matched drop)	**49.1**	+34.1	**0.9**

Table E.12. Prompt-level interventions vs. GACL at matched faithful drop. The GACL gain is 5× the best fixed-prompt gain.

E.7 Fine-tuning baseline setup

The LoRA baseline in the main text uses Qwen2-Audio-7B-Instruct with rank r=8, α=32, applied to seven LLM projections (Q, K, V, O, gate, up, down); the audio encoder is frozen. Training data include 1,000 MCR-Bench base samples (AQA 500 + SER 500), paired into faithful and adversarial forms (2,000 total) at 1:1. Optimization uses 4×A800 with DDP, lr 10^{-4}, 2 epochs, effective batch 128, and AdamW. The MCR-Bench split convention is preserved, with no test-set leakage. Evaluation uses the same 200-example conflict test split throughout.