

DOES LOCATION MATTER? A BLACK-BOX CAUSAL AUDIT REVEALS MODEL-DEPENDENT AUDIO GROUNDING IN LALM HALLUCINATION

Mohammad Khalooei^{1,2} Mohammad Sabokrou^{3,4}

¹Sharif University of Technology, Iran ²Amirkabir University of Technology, Iran
³New Uzbekistan University, Uzbekistan ⁴Okinawa Institute of Science and Technology, Japan

ABSTRACT

Large audio-language models (LALMs) frequently *hallucinate* sounds that are not present. Audio-Aware Decoding (AAD) [1] contrasts logits under the real recording against an all-zero “blank” waveform, which is out-of-distribution and conflates “evidence absent” with “input malformed.” We introduce a black-box causal test: localize the region a claim would rely on with CLAP, remove only that region via spectral gating, and compare against an equally-sized random window, on three LALMs. On Qwen2-Audio-7B ($n=4355$), localized ablation’s raw AUROC trails blank’s (0.613 vs. 0.721); we show this is a *duration* confound: a duration-only regression predicts blank’s mean effect almost exactly (0.067 vs. 0.071), and holding duration fixed, localized ablation beats random removal at every grid cell, on both models. On Audio Flamingo 3 (AF3) it flips 62.5% of true positives vs. 41.5% for random, and a synthetic oracle control disentangles model dependence from CLAP localization error. At decoding time, a language-prior contrast lifts Qwen2-Audio’s accuracy from 57.1% to 68.2%, beating AAD, and to 71.9% (cross-validated) once combined and gated. A unified causal profile does *not* transfer between models; Qwen2.5-Omni-7B replicates Qwen2-Audio’s pattern, not AF3’s. Code: github.com/khalooei/causa.

Index Terms— audio-language models, hallucination detection, causal ablation, black-box interpretability, model comparison

1. INTRODUCTION

Large audio-language models (LALMs) such as Qwen2-Audio [2] and SALMONN [3] answer free-form questions about audio content but frequently exhibit *object hallucination*: confidently claiming a sound event is present when it is not [4, 5]. The current inference-time mitigation, Audio-Aware Decoding (AAD) [1], contrasts the model’s logits with real audio against logits from an all-zero “blank” waveform, almost certainly out-of-distribution for the audio encoder: the contrast conflates “evidence absent” with “input malformed.”

We address this limitation with a more surgical intervention: localize the spectro-temporal region a claim’s evidence would occupy with sliding-window CLAP (Contrastive Language-Audio Pretraining) similarity, then physically remove *only* that region via spectral gating, leaving the rest of the real, in-distribution recording untouched. This is a *necessity* test in the sense of Pearl’s Probability of Necessity [6, 7]: if a claim survives removal of its own best-matching evidence, it was not grounded in that evidence. We complement it with a *sufficiency* test (inserting a real exemplar into a genuine negative-control clip) that validates the pipeline’s sensitivity.

Unlike concurrent causal work that needs gradient or internal-activation access [8–11], our test is fully *black-box* and training-

free. On Qwen2-Audio-7B-Instruct, localized ablation’s raw AUROC trails blank’s; we show this is a duration confound that reverses once duration is held fixed (Section 5.1). On NVIDIA Audio Flamingo 3 [12] and Qwen2.5-Omni-7B [13], raw detection value remains model-dependent.

2. RELATED WORK

Object hallucination in LALMs. Kuan et al. [4] introduce the object-existence discriminative-QA benchmark used here; a follow-up [5] broadens this to temporal-order/attribute hallucinations. AAD [1] is the closest prior mitigation, our baseline.

Causal sufficiency/necessity, and white-box audio grounding. COPO [8] trains a causal-completeness reward via RL on image patches, needing training access; concurrent audio work traces causal effects layer-by-layer [9], reads activations under silence/unrelated-audio substitution [10], or trains a frame-level head [11]. Our tests need no training and intervene on raw audio only.

Content-free calibration and adaptive contrastive decoding. Zhao et al. [14] calibrate against a content-free input; M3ID [15] contrasts an image-absent baseline, structurally identical to our Section 3.2 analysis (adapted to audio, not novel). VCD [16]/ICD [17] contrast a *distorted* input, closer to our localized/random ablation. Concurrently, TCD [18] gates a temporally-blurred contrast by uncertainty/audio-reliance and TAD [19] gates a silence contrast by token-level confidence; both gate *when* to apply one fixed contrast, where Section 5.5 instead scales *how much* of a language-prior contrast to apply.

3. METHOD

3.1. Necessity: localized spectral-gate ablation

For a claim “yes” to “is there a sound of o in the audio?”, we slide a window (2 s or 4 s, 0.5 s hop) over the clip and score each against “a sound of o ” using LAION-CLAP [20], taking the highest-similarity window as $R(o)$. We attenuate energy within $R(o)$ by -40 or -100 dB via Short-Time Fourier Transform (STFT) masking with a 50 ms raised-cosine edge fade, leaving the rest of the waveform untouched, and re-query the model (e.g. a true-positive “laughter” claim: gating the CLAP-localized region flips it, $P(\text{yes}): 0.905 \rightarrow 0.133$, while an equal-size random window does not).

3.2. Controls

Random-window. Same ablation, same size/depth, on a uniformly random window instead of CLAP’s choice. **Blank-audio** (AAD’s

baseline [1]). An all-zero waveform, scored identically. **Language-only**. No audio content block at all (not silence), giving the object-conditioned “yes” prior independent of any clip. **Real-unrelated-audio substitution**. Following [10], the clip is replaced entirely with a different real, verified-negative recording, reproduced here fully black-box.

Sufficiency. For “no” claims on a negative-control clip, we mix a real exemplar of the target class (CLAP-localized from a separate positive clip) in at matched loudness and re-query, checking whether the claim flips. **Scoring**. Every condition is scored identically: a single forward pass yields next-token log-probabilities for “yes”/“no”, renormalized over the forced choice to give $P(\text{yes})$, deterministic, with no sampling.

3.3. Oracle-localization control

The grid above ablates only CLAP’s *guess* at the evidence location, so a null result is consistent with two mechanisms: the claim does not depend on any specific region, or CLAP’s guess is wrong enough that ablating it resembles ablating random. We disentangle these by inserting a real exemplar into a negative-control clip at a random but *recorded* offset, keeping only mixtures the model affirms, and ablating three same-size windows on the same mixture: the true insertion window (**oracle**), CLAP’s own top pick (**clap**), and a uniform random window (**random**), all at -100 dB, also reporting the temporal Intersection-over-Union (IoU) between CLAP’s window and the true one: a pure localization-accuracy number real clips cannot provide.

3.4. From detection to mitigation: contrastive decoding

Contrastive decoding [21] (and its instantiations AAD [1] and VCD/ICD/M3ID [15–17]) uses a contrast to *correct* the answer at decoding time:

$$d_{\text{adj}} = (1 + \alpha) \logit(p_{\text{yes}}^{\text{real}}) - \alpha \logit(p_{\text{yes}}^{\text{contrast}}) \quad (1)$$

with $\logit(p) = \log \frac{p}{1-p}$, answer “yes” iff $d_{\text{adj}} > 0$. Reusing already-computed logits, we evaluate three contrasts: **AAD-CD** (blank audio), **LangPrior-CD** (Section 3.2, M3ID-style), **CAUSA-CD** (the grid’s strongest localized cell), sweeping $\alpha \in [0, 3]$ and reporting the literature-standard $\alpha=1$ to avoid cherry-picking. Our best rule (Section 5.5) needs two forward passes per query (real, blank); the language prior is a precomputed per-*class* constant.

3.5. Theoretical framing: from a binary PN to a graded curve

The necessity-flip rate over true positives estimates Pearl’s Probability of Necessity, $PN = P(C(\text{do}(R(o):=\emptyset)) = \text{no} \mid C(X) = \text{yes}, o \text{ present})$; the sufficiency-flip rate over correct negatives estimates Pearl’s Probability of Sufficiency, PS [6, 7]. Unlike purely observational PN/PS (generally unidentifiable without strong assumptions [7]), intervening on the same physical unit is closer to true counterfactual evidence under monotonicity: removing genuine evidence cannot newly cause “yes,” nor can inserting it decrease detection.

Ranking a clip’s windows by CLAP similarity $w_{(1)}, \dots, w_{(K)}$ and nesting the intervention $\text{do}(R_r)$ (remove the top $\lceil rK \rceil$ windows, $r \in [0, 1]$), monotonicity makes the flip event itself monotone in r , licensing

$$R^* = \inf\{r \in [0, 1] : C(\text{do}(R_r)) = \text{no}\}, \quad (2)$$

the minimal necessary removal fraction (censored at 1 if never flipped). This turns binary PN into $PN(r) := P(R^* \leq r \mid \cdot)$, its

CDF, with $PN(1)$ the classical case. Since $R^* \in [0, 1]$,

$$\mathbb{E}[R^*] = \int_0^1 (1 - PN(r)) dr =: \text{NAUC}, \quad (3)$$

the Normalized AUC, giving the deletion metric [22, 23] an exact causal reading as the mean of a latent counterfactual. Searching windows in *random* order gives R_{rand}^* ; paired per clip, the *localization efficiency* $\Delta_{\text{loc}} = \mathbb{E}[R_{\text{rand}}^* - R_{\text{clap}}^*]$ tests “does localization help” more powerfully than independent grid cells.

Estimation and monotonicity. We use 6 checkpoints (1, 2, 4, 7, 11, 17 cumulative windows, $\approx 6\%$ to full coverage of a 10s clip) at -100 dB, stopping at the first flip (biasing $\hat{R}^*$ conservatively upward). Raw $P(\text{yes})$ is *not* monotone (it locally increases for 90%/54% of Qwen-Audio/AF3 claims), but R^* needs only the *thresholded decision* never to revert to “yes”, which holds almost exactly: 0.02% reversals for Qwen-Audio (CLAP order), 0% elsewhere.

3.6. A causal grounding profile, and its identifiability

Every quantity above answers a different question about the same claim: blank-drop asks whether *any* audio matters; Δ (Section 3.2) asks whether audio matters *beyond the language prior*; the matched-control contrast asks whether the *located* region matters more than an equally-sized *arbitrary* one. We define the last as a specificity statistic, $\text{CS}_i = \logit(p_i^{\text{rand}}) - \logit(p_i^{\text{loc}})$, averaged over the four grid cells, and its graded analogue Δ_{loc} (Section 3.5), and collect all five signals into a per-claim profile $z_i = [\text{blank-drop}_i, \text{clap-sim}_i, \Delta_i, \text{CS}_i, \Delta_{\text{loc},i}]$.

What this can and cannot identify. Waveform deletion implements $\text{do}(E_o:=0)$ only if the removed span is the *only* causal channel disturbed. Under a *window-exchangeability null* (the CLAP span carries no more claim-relevant information than a size-matched random span), $\mathbb{E}[\text{CS}_i] = 0$; a nonzero CS rejects it but does not prove identification (CLAP bias toward salient spans would also give $\text{CS} > 0$), hence Section 5.3’s synthetic control with known true location.

4. EXPERIMENTAL SETUP

Models. Qwen2-Audio-7B-Instruct [2] is the primary subject; NVIDIA Audio Flamingo 3 [12] (Whisper-style encoder, Qwen2.5-7B decoder) and Qwen2.5-Omni-7B [13] (Qwen2-Audio’s lineage, newer/larger) are cross-model checks. All three queried identically black-box.

Data. The object-existence split of [5]¹, 10,800 rows, 53 classes from AudioCaps/ESC-50/VocalSound. Qwen2-Audio ($n=5000$, seed 42) yields 4355 affirmative claims (2488 correct, 1867 hallucinated, 43% overclaim rate). AF3 affirms far less (needed an explicit “answer only Yes or No” suffix; even so 11.0% vs. Qwen2-Audio’s 87.1%), so we draw the full $n=10,800$, yielding 1192 affirmative claims (1109 correct, 83 hallucinated, 93.0% precision). Qwen2.5-Omni ($n=2000$) yields 1774 affirmative claims (1022 correct, 752 hallucinated). The oracle control (§3.3) reuses the same benchmark’s negative-control rows and exemplar bank; no new data source.

Metrics. AUROC/AUPRC for hallucinated-vs-correct separation; empirical $\hat{PN}/\hat{PS}$; 2000-resample bootstrap 95% confidence

¹HF: kuanhuggingface/Audio-Hallucination_Object-Existence_AudioCaps-ESC50-VocalSound.

Table 1. Hallucination-detection AUROC, $n = 4355$ (1867 hallucinated, 2488 correct).

Method	AUROC
CAUSA causal grounding ensemble (5-feature, §3.6)	0.743
AAD blank-audio ablation (baseline) [1]	0.721
CAUSA localized, 4 s / -100 dB (best single cell)	0.613
CAUSA random-window control, 4 s / -100 dB	0.577
CAUSA localized, 2 s / -40 dB (original config)	0.563
CAUSA random-window control, 2 s / -40 dB	0.550

intervals (CIs). AF3’s small hallucinated count is well-powered at full-benchmark scale ($n=83$ hallucinated, 1109 true positives) but its per-cell AUROC ranges widely (0.37–0.58); flip-rate is the reliable metric for this model.

5. RESULTS

5.1. Blank’s edge is a duration confound

Table 1: every *single* localized cell trails blank-audio; its edge over random-window is inconsistent (-1.0 to $+3.6$ AUROC points, significant only at -100 dB depth). Localized ablation flips only $P_N = 1.0\%$ of true positives vs. 10.8% for blanking, an $\sim 11\times$ gap, though blanking is the out-of-distribution intervention we set out to improve on. Window size and depth both push AUROC monotonically toward blank’s, consistent with *how much* is removed dominating *where*. Sufficiency works as designed ($\hat{P}S = 0.731$): the weak link *appears* to be localized removal alone.

We test the duration explanation directly: the same mask-size confound vision-XAI’s deletion metrics are built to integrate away rather than report at one size [22, 23] (§3.5), diagnosed here directly instead. Blank silences 100% of a clip (median 10.0 s) vs. the grid’s 2–4 s, an unmatched comparison by construction. Fitting $P(\text{yes})$ -drop as a linear function of duration-fraction-removed, using only random-window cells (zero localization signal by construction), predicts blank’s mean drop almost exactly: 0.067 vs. 0.071 actual on Qwen2-Audio ($n=4355$, $R=0.21$, $p=2 \times 10^{-210}$); 0.710 vs. 0.734 on AF3 ($n=1192$, $R=0.73$, $p<10^{-300}$). Blank sits almost exactly on the duration-only line, so its AUROC edge is largely a duration artifact, not evidence localization fails. Holding duration fixed instead (localized vs. random-window removal of *identical* size, on the *identical* clip, paired), CLAP localization wins at every grid cell, on both models, overwhelmingly: mean drop 0.041 vs. 0.016 at 4 s / -100 dB on Qwen2-Audio (Wilcoxon $p=4.8 \times 10^{-46}$); 0.393 vs. 0.233 on AF3 ($p=5.7 \times 10^{-43}$). Pooled over all four cells, localized ablation’s residual *above* the duration-only line is large while random-window’s is ≈ 0 by construction: $+0.014$ vs. -0.001 on Qwen2-Audio ($t=19.9$, $p=1.4 \times 10^{-87}$, $n=17,420$ paired cells), $+0.096$ vs. -0.006 on AF3 ($t=23.1$, $p=2.2 \times 10^{-112}$, $n=4768$). Localization does real, statistically dominant work once amount-removed is controlled for; Table 1’s raw ranking measures a duration confound, not localization validity.

5.2. Graded necessity: thresholding costs the detector

The adaptive search (§3.5) on all 4355 Qwen2-Audio examples confirms the localization theory: paired localization efficiency is real, $\hat{\Delta}_{\text{loc}} = 0.028$ (Wilcoxon $p=4.1 \times 10^{-27}$, CI $[0.023, 0.033]$): CLAP’s ranking reliably finds sufficient evidence with less removal than a blind search. This hides heterogeneity: splitting classes into acoustically *transient* vs. *continuous/ambient*, Δ_{loc} is several times larger for transient (0.116 vs. 0.017, Mann–Whitney $p=3.1 \times 10^{-15}$), independently reproduced in the oracle control

Table 2. Oracle control, synthetic Qwen2-Audio positives with known evidence location ($n=608$, -100 dB, matched size).

	Oracle	CLAP	Random
Flip rate	16.8%	10.5%	3.1%
Mean $P(\text{yes})$ drop	0.126	0.079	0.016

Mean CLAP–true-location IoU: 0.45 (25% zero-overlap).

below, and Table 1’s headline gap splits the same way, narrower on transient classes ($+8.1$ AUROC points) than continuous ($+16.8$): a real, substantial, but partial closing of the gap.

$\hat{R}^*$ itself is, surprisingly, a chance-level detector (AUROC = 0.491): 89% of true positives/hallucinations never flip even after near-total coverage, so the binary statistic discards the graded confidence change along the way. The *continuous* $P(\text{yes})$ drop at the final checkpoint recovers it (AUROC = 0.651): a clean causal quantity and a good detector are not automatically the same once censored. On AF3, efficiency is an order of magnitude larger ($\hat{\Delta}_{\text{loc}}=0.195$), censoring nearly eliminated, yet $\hat{R}^*$ stays chance-level (AUROC = 0.52): removal amount needed and early resistance to removal are related but distinct signals, unreliable for AF3 at any scale.

5.3. Oracle-localization control: is the null a localizer artifact?

Table 2 answers the question Section 5.1 leaves open: is Qwen2-Audio’s near-flat gap because CLAP finds the wrong region, or because the model does not depend on any specific region even when known exactly? Both. CLAP’s guess only moderately overlaps the true region (mean IoU 0.45; 25% zero overlap), and ablating the true location beats ablating CLAP’s guess by a wide, significant margin (16.8% vs. 10.5%, CI $[4.1, 8.4]$), which beats random (10.5% vs. 3.1%, CI $[4.9, 9.9]$): a real share of the near-null result is a localizer-accuracy artifact. But even the *true* location flips only 16.8% of these claims, far below blank’s near-total effect: genuine, partial location-insensitivity survives this control too.

5.4. Language-prior correction: ensemble significant, alone it isn’t

Querying each class with no audio block gives mean language-only $P(\text{yes}) = 0.935$ for true positives, 0.928 for hallucinations, nearly identical (AUROC = 0.490, chance): the model leans “yes” $\approx 93\%$ of the time regardless of object. Correcting for this, $\Delta = P(\text{yes} | \text{real}) - P(\text{yes} | \text{no audio})$ gives AUROC = 0.727; a 3-feature ensemble (blank-drop, CLAP, Δ ; 5-fold cross-validation (CV)) reaches 0.743. Δ alone doesn’t beat blank-audio (CI $[-0.009, 0.020]$); the *ensemble* does per-example ($+0.022$, CI $[0.011, 0.033]$) but not per-class (CI $[-0.043, 0.046]$). Adding substitution [10] as a fourth feature does not help (AUROC = 0.646 alone; 4-feature ensemble statistically flat at 0.743), redundant rather than complementary; the full 5-feature profile (Table 1) plateaus at the same 0.743.

5.5. Contrastive decoding: a real accuracy gain

Unlike Δ ’s non-significant detection AUROC, the same contrast as a decoding correction gives a substantial, monotone gain: $+11.1$ at $\alpha=1$ (Table 3, Fig. 1), with no sign reversal in $[0, 3]$; detection and correction can disagree. AAD gives $+7.5$; CAUSA-CD alone the smallest gain; combined and gated (below) reaches $+14.3$. On AF3, whose language prior is near zero per class, LangPrior-CD/AAD-CD give $+0.0$ and CAUSA-CD *hurts* (-3.4). Aggregate accuracy hides a fix/break trade-off: on Qwen2-Audio, LangPrior-CD fixes 40.7% of hallucinations at the cost of breaking 11.1% of correct claims;

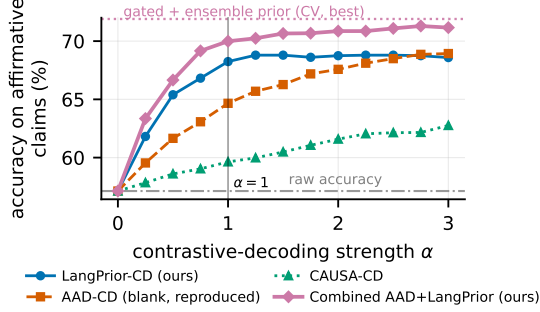


Fig. 1. Qwen2-Audio contrastive-decoding accuracy vs. α : combining AAD+LangPrior (ours) sits above every single contrast across the whole range, and confidence-gating it (dotted line, CV) higher still (Audio Flamingo 3’s sweep is flat at a 93.0% ceiling for every condition and is omitted; see text).

Table 3. Contrastive-decoding accuracy at $\alpha=1$, on each model’s own affirmative claims. Best is swept over the literature-standard $\alpha \in [0, 3]$ (Section 3.4).

Model (raw acc.)	Contrast	Acc.	Δ	Best
Qwen2-Audio (57.1%)	LangPrior-CD	68.2%	+11.1	68.8%
	AAD-CD	64.7%	+7.5	68.9%
	CAUSA-CD (loc. only)	59.6%	+2.5	62.8%
	AAD+LangPrior	70.0%	+12.9	71.3%
	+ gated	71.4%	+14.3	—
	+ ensemble prior (ours)	71.9%	+14.8	—
Audio Flamingo 3 (93.0%)	LangPrior-CD	93.0%	+0.0	93.0%
	AAD-CD	93.0%	+0.0	93.0%
	CAUSA-CD	89.6%	−3.4	93.0%

AAD-CD is more conservative (22.4%/3.6%); CAUSA-CD fixes fewest (9.1%/2.5%). On AF3, CAUSA-CD’s loss is fully explained: it fixes only 4 hallucinations while breaking 45 correct claims, 11:1. The gain is broad-based on Qwen2-Audio: of 48 classes with ≥ 15 examples, 42 improve, 5 worsen (drinking/sipping breaks twice as many as it fixes).

Since AAD-CD and LangPrior-CD correct different, only partially overlapping errors, we combine them:

$$d = (1 + \alpha_b + \alpha_l) \logit(p_{\text{real}}) - \alpha_b \logit(p_{\text{blank}}) - \alpha_l \logit(p_{\text{lang}}) \quad (4)$$

at $\alpha_b = \alpha_l = 1$: accuracy reaches 70.0%; its own swept best (over $\alpha_b, \alpha_l \in [0, 3]$) reaches 71.3%, +2.4 over either single contrast’s own best: substitutable, not redundant. 5-fold CV (fit (α_b, α_l) per training fold) gives $71.0\% \pm 0.7\text{pp}$, *confirming* the gain.

Confidence-gating and a less-noisy prior help further. Scaling LangPrior’s weight by p_{lang} itself ($w = \alpha_l p_{\text{lang}}$ replacing α_l , as TCD [18]/TAD [19] gate a temporal contrast) gives $71.4\% \pm 0.8\text{pp}$ CV, beating flat weighting in 79% of 20 splits (paired t -test, $p = 5.2 \times 10^{-9}$). Averaging 5 paraphrasings of p_{lang} instead of one gives $71.9\% \pm 0.7\text{pp}$ (+14.8 over raw), beating single-phrasing in 86% of folds (paired t -test, $p = 3.6 \times 10^{-12}$), 265 extra passes, nearly free. On AF3, combining or gating changes nothing.

5.6. Cross-model generalization: a third model, opposite pattern on AF3

The localized-vs-random gap is 7–21 points for AF3 against 0–4 for both Qwen-family models. Blanking flips *every* AF3 true positive vs. 10.8% for Qwen2-Audio; AF3 affirms only 11.0% of queries (vs. 87.1%) but at 93.0% precision (vs. 57.1%), plausibly tied to

Table 4. Necessity flip rate, localized vs. random, by model ($n=510/1022/1109$ Qwen2-Audio/Qwen2.5-Omni/AF3 true positives).

Model	Setting	Localized	Random	Gap
Qwen2-Audio	4 s/−100 dB	3.3%	2.2%	+1.2
Qwen2.5-Omni	4 s/−100 dB	7.4%	3.9%	+3.5
Audio Flamingo 3	4 s/−100 dB	62.5%	41.5%	+21.0
Qwen2-Audio	2 s/−40 dB	1.0%	0.8%	+0.2
Qwen2.5-Omni	2 s/−40 dB	1.5%	0.8%	+0.7
Audio Flamingo 3	2 s/−40 dB	26.0%	18.8%	+7.1

its claims being more localizable. A third model tests whether the Qwen2-Audio/AF3 split is a two-point coincidence: Qwen2.5-Omni-7B ($n=1774$, Qwen2-Audio’s lineage but newer/larger) replicates Qwen2-Audio’s pattern, not AF3’s (best localized cell AUROC 0.609 vs. blank’s 0.766, +15.6pp, CI [+12.5, +18.6]), suggesting (with only three models, a hypothesis) that the split tracks encoder lineage more than scale. Decoding-side, its overclaiming is far less language-prior-driven (AAD-CD/LangPrior-CD move accuracy by only +0.7/+0.3pp), but CAUSA-CD, weakest for Qwen2-Audio, gives this model’s *largest* single-contrast gain (+1.6pp).

5.7. Cross-model transfer of the causal grounding profile

If grounding failure has a common causal structure, a detector *trained* on one model’s profile z (§3.6) should transfer to another’s. In-distribution, a 5-fold CV ensemble over z only modestly beats its best single signal on Qwen-Audio (0.743 vs. 0.727, CI crosses zero by-class) and *underperforms* on AF3 (−9.9 pp), with only 83 AF3 hallucinations, so it overfits noise. The sharper test: standardize z within each model, fit the logistic router on one, score the other *without retraining*. Both directions lose to the target’s own best fixed contrast (Qwen→AF3: −1.4 pp; AF3→Qwen: −13.1 pp), a genuine negative result: the two failure modes differ qualitatively (language-prior override vs. coarse localization sensitivity).

6. DISCUSSION, LIMITATIONS, AND CONCLUSION

Limitations. Qwen2-Audio’s weak localized-vs-random gap (§5.3) is only *partly* a CLAP-accuracy artifact; the language-prior ensemble’s significance holds per-example, not per-class; we do not correct for multiple comparisons. Qwen2.5-Omni has only headline numbers; Phi-4-multimodal and FLAM [24] remain non-functional (dependency incompatibilities). On the shorter HalluAudio benchmark [25] (3.8 s vs. 10.0 s median), §5.1’s matched-duration effect attenuates sharply, though a small residual survives on Qwen2-Audio ($p = 8.9 \times 10^{-5}$). Combined-decoding (§5.5) replicates cleanly regardless ($81.2\% \pm 2.4$ vs. $80.1\% \pm 2.1$).

Conclusion. Localization’s raw detection value is model- and class-dependent, but its weakness on Qwen2-Audio is a duration confound: holding duration fixed, localized removal decisively beats a matched random window on both models (§5.1). A single causal grounding profile does not transfer across models (§5.7). Together with a validated decoding-time correction, this is our contribution.

Compliance with Ethical Standards

This study uses only publicly released benchmark audio (AudioCaps, ESC-50, VocalSound, via [5]) and openly released pretrained models; it involves no new data collection from human or animal subjects and required no institutional ethical review.

7. REFERENCES

- [1] Tzu-wen Hsu, Ke-Han Lu, Cheng-Han Chiang, and Hung-yi Lee, “Reducing object hallucination in large audio-language models via audio-aware decoding,” *arXiv preprint arXiv:2506.07233*, 2025.
- [2] 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, “Qwen2-audio technical report,” *arXiv preprint arXiv:2407.10759*, 2024.
- [3] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang, “Salmonn: Towards generic hearing abilities for large language models,” in *Proc. ICLR*, 2024, arXiv:2310.13289.
- [4] Chun-Yi Kuan, Wei-Ping Huang, and Hung-yi Lee, “Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models,” in *Proc. INTERSPEECH*, 2024, arXiv:2406.08402.
- [5] Chun-Yi Kuan and Hung-yi Lee, “Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning,” in *Proc. ICASSP*, 2025, arXiv:2410.16130.
- [6] Judea Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000.
- [7] Jin Tian and Judea Pearl, “Probabilities of causation: Bounds and identification,” *Annals of Mathematics and Artificial Intelligence*, vol. 28, pp. 287–313, 2000.
- [8] Peizheng Guo, Jingyao Wang, Wenwen Qiang, Jiahuan Zhou, Changwen Zheng, and Gang Hua, “Copo: Causal-oriented policy optimization for hallucinations of mllms,” in *Proc. CVPR*, 2026, arXiv:2508.04182.
- [9] Wei-Chih Chen, Chien-yu Huang, and Hung-yi Lee, “Causal tracing of audio-text fusion in large audio language models,” *arXiv preprint arXiv:2603.13768*, 2026.
- [10] Hyebin Cho, Suho Yoo, Jihoo Jung, and Joon Son Chung, “Tracing audio grounding and answer selection in audio llms,” *arXiv preprint arXiv:2609.04637*, 2026.
- [11] Joseph An, Phillip Keung, Jiaqi Wang, Orevaoghene Ahia, and Noah A. Smith, “Encode once, decode never: Reusing audio lm internals for efficient temporal localization,” *arXiv preprint arXiv:2602.10230*, 2026.
- [12] Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sanggil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro, “Audio flamingo 3: Advancing audio intelligence with fully open large audio language models,” *arXiv preprint arXiv:2507.08128*, 2025, NVIDIA and University of Maryland.
- [13] 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.
- [14] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh, “Calibrate before use: Improving few-shot performance of language models,” in *Proc. ICML*, 2021, arXiv:2102.09690.
- [15] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto, “Multi-modal hallucination control by visual information grounding,” in *Proc. CVPR*, 2024, arXiv:2403.14003.
- [16] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing, “Mitigating object hallucinations in large vision-language models through visual contrastive decoding,” in *Proc. CVPR*, 2024.
- [17] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann, “Mitigating hallucinations in large vision-language models with instruction contrastive decoding,” *arXiv preprint arXiv:2403.18715*, 2024, *Proc. ACL Findings* 2024.
- [18] Yanda Li, Yuhua Liu, Zirui Song, Yunchao Wei, Martin Takáč, and Salem Lahlou, “Temporal contrastive decoding: A training-free method for large audio-language models,” *arXiv preprint arXiv:2604.15383*, 2026, *ACL 2026 Findings*.
- [19] Heyu Chang, Nianwen Si, Hao Zhang, Wenlin Zhang, and Dan Qu, “Tad: Token-adaptive contrastive decoding with confidence-guided gating for hallucination mitigation in large audio-language models,” *arXiv preprint arXiv:2609.07286*, 2026, *Interspeech* 2026.
- [20] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *Proc. ICASSP*, 2023, arXiv:2211.06687, *LAION-CLAP*.
- [21] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis, “Contrastive decoding: Open-ended text generation as optimization,” in *Proc. ACL*, 2023, arXiv:2210.15097.
- [22] Vitali Petsiuk, Abir Das, and Kate Saenko, “Rise: Randomized input sampling for explanation of black-box models,” in *Proc. BMVC*, 2018, arXiv:1806.07421.
- [23] Ruth Fong, Mandela Patrick, and Andrea Vedaldi, “Understanding deep networks via extremal perturbations and smooth masks,” in *Proc. ICCV*, 2019, arXiv:1910.08485.
- [24] Yusong Wu, Christos Tsirigotis, Ke Chen, Cheng-Zhi Anna Huang, Aaron Courville, Oriol Nieto, Prem Seetharaman, and Justin Salamon, “Flam: Frame-wise language-audio modeling,” in *Proc. ICML*, 2025, arXiv:2505.05335.
- [25] Feiyu Zhao, Yiming Chen, Wenhan Lu, Daipeng Zhang, Xianghu Yue, and Jiangguo Wei, “Halluaudio: A comprehensive benchmark for hallucination detection in large audio-language models,” *arXiv preprint arXiv:2604.19300*, 2026.