CORRUPTED SIGNAL, SILENT VERDICT

Anonymous authors
Paper under double-blind review

ABSTRACT

AV evaluation benchmarks construct negatives by running source clips through an editing pipeline and use unedited originals as clean controls, so negatives and controls differ in both acoustic content and processing history. We introduce a *register-only* control: a clip that traverses the full pipeline — encode, decode, re-encode at matching settings — but receives no acoustic corruption. A commercial API judge flags register-only clips above clean ($\delta = +5.0$ pp, $[+0.7, +9.3]$, $p = 0.014$) yet flags corrupted clips (four of five types) no higher than register-only ($p = 0.394$). Verdicts from a local judge, Qwen2.5-Omni-3B, do not distinguish the three conditions (clean $72.0\,\%$, $n = 582$; register-only $71.6\,\%$; corrupted $74.2\,\%$; all $p > 0.5$), and its continuous decision logit is near chance (corrupted vs. register-only AUC 0.54 $[0.48, 0.59]$), while linear probes on the same model's hidden states separate matched items at $d' = 0.77$–2.05 (AUC 0.71–0.93) across all five corruption types. Across both judges, verdicts track processing provenance or nothing — not the acoustic manipulation — while the manipulation is linearly decodable from the representations. Benchmarks of this construction should report clean false-alarm rates and include a register-only arm.

1 INTRODUCTION

AV evaluation benchmarks test whether multimodal models detect acoustic corruptions in video. Three of the four benchmarks surveyed here construct their negatives by running source clips through an editing pipeline; unedited originals serve as clean controls (Table 1). Negatives and clean controls therefore differ on two dimensions simultaneously: acoustic content (a corruption is present) and processing history (the edited clip has traversed an encoding pipeline that the unedited clip has not). Whether multimodal judge responses track the acoustic manipulation or the processing history is untested.

A judge that responds to *processed* rather than *corrupted* would score well on every benchmark type at once without encoding anything about the corruption. The question is not robustness but confound: accuracy measurements on this benchmark design are upper bounds on acoustic sensitivity, not measurements of it.

Register-only control. We construct a *register-only* control (C2): a clip that has traversed the full editing pipeline — encode, decode, re-encode at the same codec settings and bitrate — but received no acoustic corruption and is perceptually indistinguishable from the original. A register-only flag-rate elevation above the clean baseline is attributable to pipeline provenance alone, not acoustic content. The register-only paradigm is after Chen et al. (2026a), which deploys it to separate MLLM judge response to perceptual change from code-path confounds; porting to acoustic AV corruption is our extension. We apply it to a commercial API judge (J1) and a local judge (J2, Qwen2.5-Omni-3B (Xu et al., 2025)), instrumenting J2 with linear probes at 36 hidden-state layers: across both judges, verdicts track processing provenance or nothing, while the acoustic manipulation remains linearly decodable from J2's representations (§5).

Relation to prior work. Senses Wide Shut Quang et al. (2026) and Represented but Ignored Peng & Nuchged (2026) documented representation–action gaps in omni-modal LLMs under unedited stimuli; neither addresses AV corruption detection or a processing-pipeline confound. Our dissociation has a different structure: the verdict responds to provenance rather than acoustic content, while

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

Table 1: Negative-item construction and processing-control status across AV evaluation benchmarks. C1 = same-pipeline clean control; C3 = source-dataset probe gate; entries state whether each is applied or released.

Benchmark	Negative construction	C1	C3	Item-level release
AVID Chen et al. (2026b)	5 editing injectors over LongVALE video; frames, resolution, duration preserved	No	No	Not released
AV-SyncBench Zhou et al. (2026)	Offset, jitter, speed, voice/instrument conversion; originals not re-encoded	No	No	Code/data public; model outputs unspecified
OmniHalluc-L Dong et al. (2026)	"Minimally edited" counterfactuals rebinding evidence to wrong speaker/moment/scene	No	No[a]	Withheld (documented)
AV-Phys-Bench Cui et al. (2026)	Human-annotated anti-physics prompts; not pipeline-edited	N/A	No	Public sample split
VidAudit Cakiroglu et al. (2026)	AI-generated vs. real video; 14 detectors	Yes	Yes	Yes
This work	Acoustic corruption pipeline vs. original source	Yes (n=498)	Yes (fired; 2 types withdrawn)	Yes

[a]Text-only language-prior sanity checks only.

the content is present in the representations. We establish which input feature drives the behavioural response, not the mechanism of the gap.

Contributions.

1. **Register-only dissociation, across both judges and within J2.** J1 flags register-only clips above clean ($p = 0.014$) but corrupted clips (four of five types) no higher than register-only ($p = 0.394$); J2's verdicts do not distinguish the three conditions at all, while the manipulation is linearly decodable from J2's hidden states at $d' = 0.77$–2.05 (all five types, 95 % CIs excluding zero). The behavioural measure tracks provenance, not manipulation, at this sample size.
2. **Processing-matched controls for MLLM AV judging, including the type-dependent pipeline footprint.** C1 (same-pipeline re-encode) and C3 (source-dataset probe) ported from VidAudit Cakiroglu et al. (2026); C2 (register-only) after Chen et al. (2026a), with physical calibration. C3 fired and withdrew two types before analysis (§3.1); the C1-vs-C2 footprint delta shows the pipeline inflates corrupted-vs-clean d' for two of five types (§5.3) — the outcome is not predictable a priori, so the controls must be run.

2 Related Work

AV evaluation benchmarks. AVID Chen et al. (2026b) constructs audio-visual inconsistencies via agent-driven editing over LongVALE video (11,200 videos, 39,400 annotated inconsistency events). AV-SyncBench Zhou et al. (2026) decouples temporal and semantic synchronisation (3,269 videos, 38,390 samples). Both construct negatives via a processing pipeline over real video (Table 1). OmniHalluc-L Dong et al. (2026) uses "minimally edited" counterfactuals with strict-pair verification; its release explicitly omits item-level data (raw video, source IDs, annotations, labels, and pair membership Dong et al. (2026)). AV-Phys-Bench Cui et al. (2026) uses human-annotated "Anti-AV-Physics" prompts over generated video — negatives are human-labelled, not pipeline-edited, so the processing-footprint argument does not apply to it. Table 1 summarises negative construction and control status.

Processing confounds in probing. VidAudit Cakiroglu et al. (2026) formalises a six-control protocol for AI-generated-video detection; C1 (same-pipeline re-encode) and C3 (source-dataset discriminability probe) are the two we apply. Our corrected transfer matrix (§5.4) is an instance of the

same confound in the acoustic domain: cross-artifact transfer AUC in $[0.75, 0.91]$ against shared clean controls. Annotation-artifact (Gururangan et al., 2018) and shortcut-learning (Geirhos et al., 2020) analyses apply the same logic to NLI benchmarks and vision respectively; the register-only control is the AV-judge analogue.

Representation–action gap. Senses Wide Shut Quang et al. (2026) showed that multimodal LLMs encode audio-visual information in hidden states but fail to express it behaviorally; the stimuli are presented with video and audio content unmodified Quang et al. (2026), so the gap is not attributable to stimulus alteration. Represented but Ignored Peng & Nuchged (2026) — covering Qwen2.5-Omni-7B in the speech prosody domain — established the F3 underuse pattern, in which the correct prosodic category is represented inside the model but is not sufficiently expressed in the final answer Peng & Nuchged (2026). Our F3 comparison (§G) is cross-domain: non-speech acoustic corruption magnitude, different model scale. PAL Alex et al. (2025) examines architectural strategies for audio-encoder–LLM transfer and is the closest prior analysis of the interface our T1 and T2 transitions measure (§G). Our addition is the register-only dissociation: the verdict tracks processing provenance, not the acoustic manipulation, while the acoustic content remains linearly decodable from the representations.

Judge evaluation. A Judge Should Know What Changed Chen et al. (2026a) deploys interventions against register-only controls to separate MLLM judge response to perceptual change from code-path confounds; our C2 design (§3.3) ports the register-only paradigm to acoustic AV corruption. Dark Current Usami et al. (2026) audits text LLM judges under a datasheet protocol; their *dark current* (FAR on null inputs) and severity ladder map directly to our d' and criterion c. Lee et al. Lee et al. (2026) formalise judge sensitivity ($q_1 = \text{HR}$) and specificity ($q_0 = 1 - \text{FAR}$) and give a plug-in bias correction; our π-flip formula is the ranking-level extension.

3 INSTRUMENT

3.1 CORPUS

Five acoustic corruption types are applied to AVE (Tian et al., 2018) clips, each changing exactly one audio property. Severity is defined by physical measurement, not human perceptibility (Table 2): dB RMS attenuation, dBr overlay level, RT60 of the applied room impulse response, or CLAP (Elizalde et al., 2024) cosine distance of the substituted segment (foley donors from ESC-50 (Piczak, 2015)); all values are derived from `manifest_v5.jsonl`.

Table 2: Corpus severity parameters derived from `manifest_v5.jsonl`. Fixed-tier types have the same value for all items at that severity. Variable-tier types show the range and median across all items.

Type	Parameter	Sev 1	Sev 2	Sev 3
muted_source	atten. (dB)	-6 (fixed)	-15 (fixed)	-60 (fixed)
hallucinated	overlay (dBr)	-20 (fixed)	-10 (fixed)	-3 (fixed)
room_mismatch	RT60 (s)	[0.10,0.30] med 0.20	[0.30,1.00] med 0.66	[1.51,4.99] med 3.37
ambience_splice	CLAP dist.	[0.60,0.95] med 0.73	[0.72,0.97] med 0.82	[0.82,1.08] med 0.90
foley_swap	CLAP dist.	[0.64,0.91] med 0.76	[0.76,0.98] med 0.84	[0.85,1.20] med 0.93

Withdrawn types. `speech_content_swap` and `speech_desync` are withdrawn because the C3 gate fired (§3.3): their signal reflects source-dataset identity. `av_desync_lag` and `av_desync_lead` are generated by zero-pad-and-trim; the padding alters mean audio statistics visible to mean-pooled encoder representations and they are withdrawn pending register-only 0 ms controls (C2).

CLIP visual-absence gate. For `hallucinated_sound`, a CLIP (Radford et al., 2021) gate (ViT-B-32-quickgelu, OpenAI weights, threshold 0.196) rejects candidates whose sound source is visually present; items with no passing candidate receive synthesised pink noise (calibration and per-tier fallback counts in Appendix D).

Paired design. Severity probing uses source clips passing gates at both sev1 and sev3: $n \in \{29 \text{ (ambience)}, 37 \text{ (foley)}, 34 \text{ (hallucinated)}, 44 \text{ (muted)}, 46 \text{ (room)}\}$.

3.2 VERIFICATION GATES

Physical verification is run post-generation for every item; items failing their type's gate are excluded. Gate criteria, per-type acceptance rates (68–100 % depending on gate), and item counts are in Appendix D.

3.3 PROCESSING-MATCHED CONTROLS

All corrupted items traverse a PyAV file-rewrite pipeline; clean items do not. Three controls address this:

C1 (ported from VidAudit Cakiroglu et al. (2026)). VidAudit's C1 passes every input through an H.264 pipeline before feature extraction, removing codec-side fingerprints. Our analogue: clean items are passed through the identical PyAV AAC encode/decode pipeline (same parameters), producing clean-re-encoded items. Probes on corrupted-vs-clean-re-encoded isolate content signal from pipeline footprint.

C2 (ours — not from VidAudit). Register-only items: each corruption code path is run but configured to produce no perceptual change (0 ms pad-trim for desync, matched-appropriate RIR for room, 0 dB gain for muted, semantically consistent foley, same-scene ambience); register-only design after Chen et al. (2026a), ported to acoustic AV corruption with physical calibration. A probe that cannot separate corrupted from register-only items detects the code path, not the artifact.

C3 (adapted from VidAudit Cakiroglu et al. (2026)). VidAudit's C3 trains a classifier to discriminate the two real sources, establishing a measurable floor on dataset-bias contribution. Our analogue: a probe trained to discriminate source datasets (VoxCeleb2 vs. AVE) at the same layer and pool as the main analysis. AUC at or above the Bonferroni-corrected threshold (0.54) indicates dataset identity is linearly decodable; any type sourced from the discriminable dataset is disqualified. *C3 applied and acted on:* `speech_content_swap` and `speech_desync` are VoxCeleb2-sourced (Chung et al., 2018); the source probe reached AUC 1.000 at every layer. Both types withdrawn (§3.1).

3.4 SDT APPARATUS

$d' = z(\text{HR}) - z(\text{FAR})$ with criterion c (Green & Swets, 1966; Hautus et al., 2021). Log-linear correction (Snodgrass & Corwin, 1988): add 0.5 to numerator counts and 1 to denominators before z-transformation. Group d' pools counts within a cell before applying z (pooled-counts estimator, not mean of per-item scores). Behavioural bootstrap: resample corrupted and clean items independently, $N = 2000$, seed 42. Severity-probing bootstrap: resample source clips, carrying both severity levels of each resampled clip into every replicate, so transition-difference CIs are computed on a single clip sample rather than by comparing marginal CIs. Within-clip pairing is a correctness requirement, not an optimisation: it raises severity-regression r^2 up to 6.5-fold (Appendix F).

4 PROTOCOL

4.1 PROBE ARCHITECTURE

LogisticRegression ($C = 1$, lbfgs, `max_iter`=400, `class_weight`=balanced), standardised features, StratifiedGroupKFold(5) grouped by source video ID, per-fold AUC averaged; linear decodability shows the information is present, not that the model uses it (Hewitt & Liang, 2019; Belinkov, 2022). Fold AUCs for the primary probe (L9 `last_prompt`, AVE-only): [0.945, 0.846, 0.869, 0.931, 0.904].

Pool disqualification. Pool cosine-similarity analysis and the resulting transition withdrawals are in Appendix G; primary analyses use T1 (within-pool), T3-LP (`hidden_last_prompt`), and T4 (`hidden_decision`).

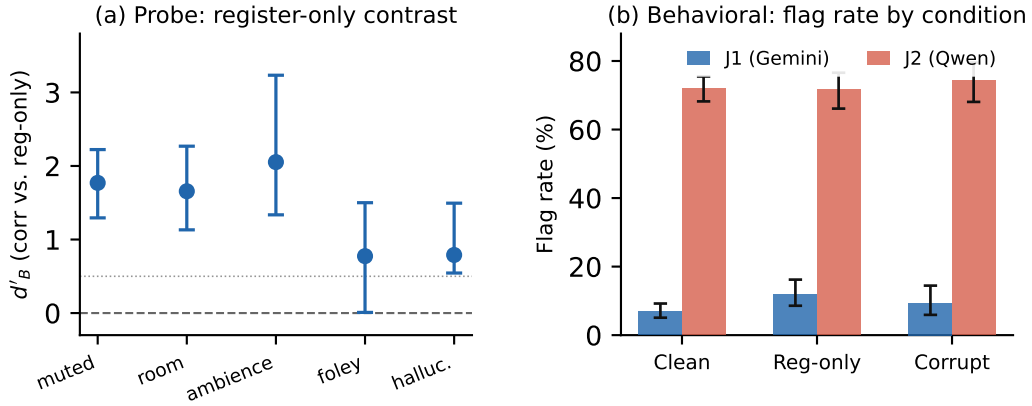


Figure 1: (a) Register-only probe d'_B by type (L35 audio pool, paired clips; error bars = 95 % bootstrap CI). (b) Behavioural flag rates for J1 and J2 across conditions (error bars = Wilson 95 % CI). J1 corrupted: four types (ambience excluded).

Layer indexing. `model.layers` holds 37 entries; entry 0 has constant activation norm (std = 0.000) and audio probe AUC = 0.500, identifying it as a pre-block module. Layer indices 0–35 denote transformer blocks throughout.

Selection control. Every permutation null performs the same max-over-(layer × pool) selection as the real analysis, so nulls are valid floors for selected sites. Best-layer selection inflates $\Delta d'$(LLM $\rightarrow$ verdict) by construction: anchoring at L35 rather than the argmax layer reduces the apparent readout gap from 4 of 5 types to 1 of 5 (Appendix G).

4.2 Judges

J1 (commercial API judge). J1 is `gemini-3.5-flash-lite` (`models/gemini-3.5-flash-lite`; `google-genai` 2.24.0). Items are submitted as MP4 files via the Gemini file API; verdict extracted by `parse_model_output`; temperature 0.0; `max_output_tokens` 256. Results deduplicated by best-result-wins (`scripts/j1_utils.py`): for each video, the most recent row carrying a valid parse is retained.

J2 (local judge). J2 is Qwen2.5-Omni-3B (Xu et al., 2025), run locally with greedy decoding, producing a binary forced-choice verdict plus an fp32 decision logit. SYSTEM_PROMPT is identical to J1's; the user turn is "Watch the video and answer the question." rather than J1's "Please analyze this video." This prompt difference prevents direct verdict comparison between the two judges and is reported as a confound; each judge's within-condition contrasts stand independently.

5 Results

5.1 Behavioural dissociation: a commercial judge tracks provenance

Table 3 and Figure 1(b) show that J1 flags register-only clips above clean ($p = 0.014$), but corrupted clips (four of five types; ambience excluded below) are not flagged significantly above clean ($p = 0.269$) and are indistinguishable from register-only ($p = 0.394$). The judge's elevated flag rate on manipulated clips is attributable to pipeline history; no acoustic contribution is detectable at $n = 182$, which is not to say none exists. Ambience was collected in an earlier session (`j1_corr_r1`, $n = 37$, 16.2 %); the session gap prevents attributing its difference to type versus session, and it is not pooled (Appendix E). J1 operates at a conservative criterion (clean FAR 6.9 %; $c \approx +1.4$); per-type comparisons are reported descriptively in Appendix E.

Table 3: Behavioral flag rates and pairwise contrasts. Rate: fraction flagged (Wilson 95 % CI). n: items. δ: percentage-point contrast (95 % CI, unpooled SE; two-sided z-test). J1 corrupted: four types (ambience excluded; see §5.1). J2 logit AUC: ROC AUC of fp32 decision logit; three contrasts reported in footnote.

	Clean rate % [CI] (n)	Reg-only (C2) rate % [CI] (n)	Corrupted rate % [CI] (n)
J1 (Gemini)	6.9 [5.1, 9.2] (582)	11.9 [8.6, 16.2] (278)	9.3 [5.9, 14.4] (182)
J2 (Qwen2.5-Omni-3B)	72.0 [68.2, 75.5] (582)	71.6 [66.1, 76.6] (282)	74.2 [68.1, 79.5] (221)
	Reg − Clean δpp [CI] p	Corr − Clean δpp [CI] p	Reg − Corr δpp [CI] p
J1	+5.0 [+0.7, +9.3] 0.014*	+2.5 [−2.2, +7.2] 0.269	+2.5 [−8.2, +3.2] 0.394
J2	−0.4 [−6.8, +6.0] 0.912	+2.2 [−4.6, +9.0] 0.530	−2.6 [−10.4, +5.2] 0.519

J2 logit AUC (DeLong (DeLong et al., 1988), n=221 pairs): corr vs. reg-only 0.54 [0.48, 0.59]; corr vs. clean 0.54 [0.49, 0.58]; reg-only vs. clean 0.50 [0.46, 0.54]. $^*p < 0.05$.

Table 4: Corrupted-sev3 vs. controls at L35 (paired clips, $N_{\text{boot}} = 10{,}000$). [A]: vs. C1 pipeline-matched clean ($n_{\text{C1}} = 498$). [B]: vs. C2 register-only. $\Delta d' = d'_A - d'_B$ (joint bootstrap over shared clips); positive values mean the pipeline inflates the corrupted-vs-clean probe. Bold = CI excludes zero.

Type	n	AUC$_A$	d'_A	AUC$_B$	d'_B [95% CI]	$\Delta d'$ [95% CI]
muted	44	0.885	1.70	0.895	**1.77** [1.29, 2.21]	−0.07 [−0.26, +0.01]
room	46	0.946	2.28	0.879	**1.66** [1.13, 2.25]	**+0.62** [+0.37, +1.11]
ambience	29	0.931	2.10	0.927	**2.05** [1.32, 3.23]	+0.04 [−1.19, +0.39]
foley	37	0.876	1.63	0.708	**0.77** [0.01, 1.50]	**+0.86** [+0.31, +1.36]
hallucinated	34	0.744	0.93	0.712	**0.79** [0.54, 1.49]	+0.14 [−0.10, +0.24]

5.2 WITHIN-MODEL DISSOCIATION (J2)

J2's verdicts do not distinguish the three conditions (Table 3; all three contrasts $p > 0.5$). J2 flags nearly three-quarters of clean clips, operating at a liberal criterion ($c \approx -0.6$) where the binary verdict channel has little room to separate conditions. The continuous decision logit likewise fails to discriminate: corrupted vs. register-only AUC 0.54 [0.48, 0.59] ($n = 221$ pairs, DeLong (DeLong et al., 1988)); corrupted vs. clean 0.54 [0.49, 0.58]; register-only vs. clean 0.50 [0.46, 0.54] (Table 3, footnote). By contrast, the same model's L35 hidden states reach probe AUC 0.71–0.93 on matched items (§5.3). The verdict is not a parser artefact: of 221 corrupted-arm responses, 99 distinct strings occur, and the named corruption type never matches the applied one.

5.3 REGISTER-ONLY PROBE AND THE PIPELINE FOOTPRINT

If an L35 probe classified by processing history only, corrupted and register-only clips would be indistinguishable ($d' \approx 0$). Table 4 shows otherwise: comparison B (corrupted vs. register-only) yields $d'_B = 0.77$–2.05 with 95 % CIs excluding zero for all five types, meeting the pre-registered criterion ($d' > 0.5$, CI excluding zero; `PREREGISTRATION_C2.md`). The acoustic anomaly is linearly decodable under processing-matched controls. Foley's CI only barely excludes zero; its retention rests additionally on positive test d' across a 12-condition regularisation sweep (Appendix E).

The footprint is type-dependent. For room and foley, the C1-clean comparison overestimates discrimination relative to the register-only baseline ($\Delta d' = +0.62$ and $+0.86$, both CIs excluding zero): the encode–decode–re-encode pipeline leaves a detectable signature independent of the corruption. For muted, ambience, and hallucinated, $\Delta d'$ spans zero and d'_B accounts for essentially all of d'_A. Because the footprint is not predictable from d'_A, a design running C1 but not C2 would have reported room and foley at inflated values without knowing it. The ambience interval is wide because $n = 29$ pairs let A and B swing in opposite directions within a resample, not because of AUC clipping (Appendix E).

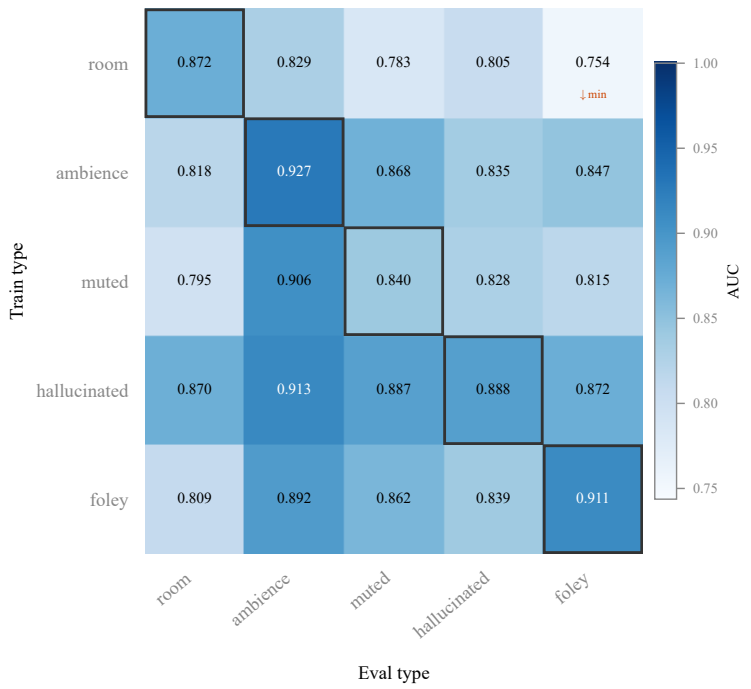


Figure 2: Corrected cross-artifact transfer matrix at L9 `last_prompt` (grouped CV; clean items held out consistently for every cell). Diagonal (outlined): within-type CV AUC. Off-diagonal: probe trained on row type, evaluated on column type. Minimum off-diagonal 0.754 (room → foley).

Probe footprint and behaviour are not the same quantity. Per-type probe footprints do not predict J1's per-type register-only elevation (Spearman $\rho = 0.30$, $n = 5$, uninformative). Muted has no representational footprint ($\Delta d' = -0.07$) but the second-largest behavioural elevation ($+7.4$ pp); room has a significant footprint ($+0.62$) but ranks third ($+4.0$ pp). A single latent "footprint" variable does not link the two measures.

5.4 A TYPE-AGNOSTIC ACOUSTIC-ANOMALY DIRECTION

A generic corrupted-vs-clean probe (L9 `last_prompt`, AVE-only, $N = 1457$) reaches AUC 0.899, versus 0.826 [0.766, 0.885] at the pre-LLM audio encoder (shuffled-label null max 0.527, below the Bonferroni threshold 0.54). Figure 2 shows the corrected cross-artifact transfer matrix: off-diagonal AUC lies in $[0.754, 0.91]$ with mean 0.841 against a diagonal mean of 0.888 (gap 0.046 [0.037, 0.055], fold-level t-CI, $df = 4$). In a descriptive two-factor decomposition of the 25 cells (not independent; no p-values), the diagonal indicator — the only type-specific term — explains the *least* variance ($R^2 = 0.181$; training type 0.362, evaluation type 0.256; full $R^2 = 0.799$, $R^2_{\text{adj}} = 0.679$). Column increments (Table 12) exclude zero only for foley ($+0.089$) and hallucinated ($+0.062$); for room, ambience, and muted, training on the target type gives no measurable advantage over training on an unrelated type, and muted's increment is negative.

See Appendix G for column increments, the converging discrete-vs.-global split, the four-transition decomposition, and full layer-curve analysis.

5.5 SEVERITY IS ENCODED AND NOT REPORTED

Three types show strong severity encoding at the pre-LLM audio encoder: room ($d'_{\text{AEM}} = 2.12$ [1.52, 2.72]), ambience (1.92 [0.97, 3.03]), and muted (1.81 [1.39, 2.57]). J2's (Qwen2.5-Omni-3B) verdict logit d' is indistinguishable from zero for all five types (Table 8 in Appendix E), and the logit AUC for corrupted vs. register-only is 0.52–0.55 per type, with every DeLong CI spanning 0.5 (Appendix G). For muted, the severity signal is lost at the readout transition T4 ($\Delta d' = +0.47$ [0.10, 0.82]): the decision token at L35 still carries signal ($d' = 0.45$) that the verdict logit does not.

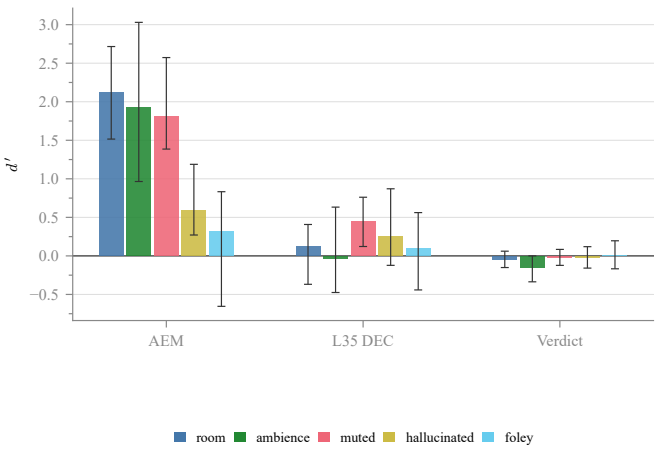


Figure 3: d' chain (sev1 vs. sev3, paired clips, joint bootstrap $N = 500$) at the AEM encoder, L35 DEC token, and J2 (Qwen2.5-Omni-3B) verdict logit; types ordered by AEM d'. Verdict CIs all span zero (Table 8).

For room and ambience, the readout gap is unresolved (T3-LP and T4 both span zero; T2 and T3a withdrawn; see Appendix G). Figure 3 shows the full d' chain; the four-transition decomposition and GLMM robustness are in Appendix G.

6 Discussion

What the judges respond to. Both judges fail to separate corrupted from register-only clips, in opposite ways: J1 at a conservative criterion flags anything that has been through the pipeline slightly more often; J2 at a liberal criterion flags nearly everything. In neither case does the verdict track the acoustic manipulation, while J2's own hidden states do. J2's continuous decision logit corroborates the verdict: AUC 0.54 $[0.48, 0.59]$ for corrupted vs. register-only (DeLong), compared to probe AUC 0.71–0.93 on matched items. For a benchmark whose negatives are pipeline-edited and whose controls are not, a judge's accuracy therefore mixes three components that the standard design cannot separate: sensitivity to the manipulation, sensitivity to provenance, and criterion.

The anomaly direction is generic in type, not in content. High cross-type transfer first suggested a provenance reading; the register-only probe falsifies it (pre-registered, `PREREGISTRATION_C2.md`). The control requirement survives in a different form: because the direction is type-agnostic, a corrupted-vs-clean probe cannot attribute its signal to a type, and because the pipeline footprint is type-dependent (§5.3), the size of the correction is not knowable without running it.

Mechanism. The d' chain (AEM $\rightarrow$ L35 $\rightarrow$ verdict; §5.5) shows severity is encoded before the LLM but not expressed in the verdict. T1 is near zero for all five types, so severity signal is not lost at the encoder–LM interface. For muted, the loss is at readout (T4), consistent with F3 underuse (Peng & Nuchged, 2026). For room and ambience, both interpretable downstream transitions span zero and the stage of loss is unresolved. Activation patching of the L35 last-prompt state into the decision position (`PREREGISTRATION_CAUSAL.md` §2c) would test whether a routing gap is causal; it is pre-registered but not yet run. SWS's Probe-Guided Logit Adjustment (Quang et al., 2026) intervenes at the output layer, not at intermediate positions, and so does not address this question.

Recommendations for benchmark design. For the three surveyed benchmarks that construct negatives via a processing pipeline (AV-Phys-Bench uses human-annotated negatives and is outside this critique), any probe or judge result without same-pipeline controls (C1, C2) is an upper bound on acoustic sensitivity. Two additions are cheap: a clean false-alarm rate (one inference pass over held-out clean items) and a register-only arm (re-encoding the source clips through the pipeline).

Without the first, a judge that over-flags clean material is indistinguishable from a sensitive one; without the second, elevated flag rates on corrupted items cannot be attributed to content rather than provenance. More generally: report HR and FAR separately, d' and c alongside accuracy, and corpus prevalence (π-flip when $\pi > 0.5$); release item-level outputs, which OmniHalluc-L (Dong et al., 2026) explicitly withholds.

Limitations. The behavioural dissociation rests on two judges with different user-turn prompts and on $n = 182$ corrupted J1 items (four types); the corrupted-vs-clean interval $[-2.2, +7.2]$ pp admits an acoustic contribution we lack power to detect. All representational claims — the register-only probe, the transfer matrix, the transition decomposition, and the representation–behaviour gap — are single-model (Qwen2.5-Omni-3B); hidden states are unavailable for API judges. Foley and hallucinated are inconclusive at current n. The audio-mean pool's contamination ($\cos > 0.95$ at all layers) withdraws T2 and T3a, leaving the room and ambience mechanism unresolved; activation patching and the rating condition are pre-registered but not run. Desync types are withdrawn pending 0 ms controls, and there is no human psychophysical baseline.

USE OF AI ASSISTANCE

We used Claude (Anthropic) in this work. The authors set the research goals, directed the work, and manually inspected the generated audio–video stimuli. LLM coding agents wrote and executed the data-generation, inference, probing, and statistical-analysis scripts on the authors' hardware; the authors reviewed the code and its outputs, identified errors, and requested revisions, which the agents implemented. The authors critiqued the experimental design and statistics, with further critique from an LLM assistant, which also proposed analyses and interpretations and helped draft and revise the manuscript; the authors decided which suggestions to adopt. Every reported number was produced by a script and traced to a result file, and the authors verified all claims and take full responsibility for the paper.

REFERENCES

Tony Alex, Wish Suharitdamrong, Sara Atito, Armin Mustafa, Philip J. B. Jackson, Imran Razzak, and Muhammad Awais. PAL: Probing audio encoders via LLMs — audio information transfer into LLMs. 2025. arXiv:2506.10423.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Mert Onur Cakiroglu, Zhihe Lu, Mehmet Dalkilic, and Hasan Kurban. Auditing generalization in AI-generated video detection: A six-control protocol and the VidAudit toolkit. 2026. arXiv:2606.31004.

Jianlin Chen, Wenhui Chen, Ziyao Lin, and Chi Man Vong. A judge should know what changed: Construct validity for LLM-as-a-judge evaluation. 2026a. arXiv:2608.24419.

Zixuan Chen, Depeng Wang, Hao Lin, Li Luo, Ke Xu, Ya Guo, Huijia Zhu, Tanfeng Sun, and Xinghao Jiang. AVID: A benchmark for omni-modal audio-visual inconsistency understanding via agent-driven construction. 2026b. arXiv:2604.13593.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Interspeech*, 2018.

Zijun Cui, Xiulong Liu, Hao Fang, Mingwei Xu, Jiageng Liu, Zexin Xu, Weiguo Pian, Shijian Deng, Feiyu Du, Chenming Ge, and Yapeng Tian. Do joint audio-video generation models understand physics? 2026. arXiv:2605.07061.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

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

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. arXiv:2309.05767.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 2, pp. 107–112, 2018.

Michael J. Hautus, Neil A. Macmillan, and C. Douglas Creelman. *Detection Theory: A User's Guide*. Routledge, 3rd edition, 2021.

John Hewitt and Percy Liang. 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)*, pp. 2733–2743, 2019.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report LLM-as-a-judge evaluations. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2511.21140.

Linkai Peng and Baorian Nuchged. Represented but ignored: A causal account of prosodic underuse in audio-language models. 2026. arXiv:2608.19211.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia (ACM MM)*, pp. 1015–1018, 2015.

Trung Nguyen Quang, Yiming Gao, Fanyi Pu, Kaichen Zhang, Shuo Sun, and Ziwei Liu. Senses wide shut: A representation-action gap in omnimodal LLMs. 2026. arXiv:2605.13737.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Joan G. Snodgrass and June Corwin. Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117(1):34–50, 1988.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision (ECCV)*, pp. 252–268, 2018.

Hiroyasu Usami, Keisuke Hara, Ayato Tsuboi, and Naohiko Matsuda. LLM judges have dark current: A psychometric datasheet for LLM-as-a-judge evaluation. 2026. arXiv:2606.15610.

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. 2025. arXiv:2503.20215.

Tianhong Zhou, Mingyang Han, Boyu Li, Yuxuan Jiang, Jiaxin Ye, Dongxiao Wang, Haoxiang Shi, Kunpeng Wang, Jun Song, Cheng Yu, and Bo Zheng. AV-SyncBench: Decoupled benchmarking of temporal and semantic audio-visual synchronization. 2026. arXiv:2607.00726.

A Data Sources

Numeric claims in the main text are sourced from the result files below.

1. Generic probe AUC 0.899: `results/probe_pass3_results.json#ave_only_artifact_probe.best_auc`
2. Cross-artifact transfer $[0.754, 0.91]$: `results/probe_transfer_matrix_cv.json#leakage_check` (corrected grouped CV; replaces the leaked estimate in `probe_transfer_matrix.json`)
3. Register-only probe d'_B and $\Delta d'$: `results/probe_c2_contrast.json` (`PAIRED_ONLY=True`, $N_{\text{boot}} = 10{,}000$)
4. J1 contrasts: `results/headline_fraction.json#contrasts`; GLMM and pooled OR: `results/pooled_behavioral_test.json`
5. J2 contrasts: `results/j2_contrasts.json`
6. Room AEM $d' = 2.12$ [1.52, 2.72]: `results/probe_dprime_results.json#room_mismatch.stage1_aem`
7. Muted T4 $\Delta d' = +0.47$ [0.10, 0.82]: `results/probe_four_transitions.json#muted_source.transitions.t4`
8. T3-LP CIs (room $[-0.09, 0.88]$; ambience $[-0.42, 1.02]$): `results/probe_four_transitions.json#room_mismatch.transitions.t3lp` and `#ambience_splice.transitions.t3lp`

B Reproducibility

Judge models. J1: `gemini-3.5-flash-lite` (`models/gemini-3.5-flash-lite`; `google-genai` 2.24.0), temperature 0.0, `max_output_tokens` 256, items submitted as MP4 via the Gemini file API. J2 / probing: Qwen2.5-Omni-3B downloaded from HuggingFace, revision `f75b40e3da2003cdd6e1829b1f420ca70797c34e`.

Compute. Instrumented inference: `scripts/run_qwen_instrumented.py`, run on a single consumer GPU.

Analysis scripts. Probing: `scripts/probe_pass3.py`. Tables: `scripts/make_tables.py`. Figures: `scripts/make_figures.py`.

C Pre-registration log

Pre-registered hypotheses and their outcomes:

- **C2 provenance-shortcut hypothesis** (`PREREGISTRATION_C2.md`): Falsified. Register-only probe yields $d' = 0.77$–2.05 all types; the probe encodes acoustic content, not only pipeline history. Revised interpretation: type-agnostic acoustic-anomaly direction.
- **Severity ladder** (`PREREGISTRATION_LADDER.md`): Confirmed for four of five types. Encoder d' separates sev1 from sev3 with CIs excluding zero for room, ambience, muted, and hallucinated; foley is inconclusive (CI spans zero; §G). Verdict does not scale with severity for any type. The prereg requires $\geq$4 types for full confirmation; the current result meets that threshold (4/5 with CI excluding zero).
- **Activation patching LP→DEC** (`PREREGISTRATION_CAUSAL.md` §2c): Deferred. T4 (representation-to-output readout) is the established finding for muted_source; activation patching to test causality is not yet run.
- **Rating condition** (`PREREGISTRATION_RATING_CONDITION.md`): Not yet run. Pre-registered; deferred pending main results freeze.

D CORPUS VERIFICATION STATISTICS

Per-type item counts, derived from `manifest_v5.jsonl` (`scripts/compute_acceptance_rates.py`). Pool size is 308 source clips; *ambience_splice* has fewer items at sev-2/3 because some clips produced no candidate segment meeting the CLAP-distance threshold.

Table 5: Per-type corpus item counts (all splits combined). Manifest counts reflect source clips with ≥ 1 accepted item per severity tier, not gate acceptance rates (see Table 6).

Type	sev 1	sev 2	sev 3
room_mismatch	308	308	308
ambience_splice	303	284	282
muted_source	308	308	308
hallucinated_sound	308	308	308
foley_swap	308	308	308

Gate acceptance rates. Sourced from `build4/build_stats.json` test-split cell statistics (`scripts/compute_gate_acceptance.py`).

Table 6: Per-type gate acceptance rates (test split, n_accepted / n_attempts). *hallucinated_sound* rates reflect the fill rate including resampling; synthetic-noise fallback counts are in the gate notes below.

Type	Gate	Sev 1	Sev 2	Sev 3	Notes
room_mismatch	none	100%	100%	100%	
muted_source	precond.	100%	98%	100%	
ambience_splice	CLAP dist.	80%	68%	84%	47/59, 47/69, 47/56
foley_swap	CLAP dist.	81%	84%	86%	47/58, 47/56, 47/55
hallucinated_sound	CLIP visual	69%	63%	65%	47/68, 47/75, 47/72

Gate notes. *ambience_splice*: substituted segments must meet the per-tier CLAP-distance lower bound; clips producing no qualifying segment are absent from the manifest. The missing sev-3 items (26 clips below 308) are distinct source clips, not re-failures of the same clip. CLAP distances use Microsoft CLAP (`msclap`, model version 2023; Elizalde et al., 2024), identified from the corpus-build code and the 1024-dimensional cached embeddings; the checkpoint hash was not logged.

hallucinated_sound: a CLIP gate (ViT-B-32-quickgelu, OpenAI weights, threshold 0.196) rejects candidates whose sound source is visually present in the frame. `data/calibration/thresholds.json` records `visual_absence_max_sim_p90` $\approx$ 0.192–0.194 for calibration-set items; the threshold is set just above this calibrated maximum. Clips exhausting the attempt budget (250 candidates) receive synthesised pink noise; per-tier synthetic-noise fallback counts were 3–4 items per tier in the test split.

E SUPPLEMENTARY RESULTS

J1 per-type register-only rates. Descriptive only: all five CIs include zero, per-type power ranges from 13 % to 54 %, and the five comparisons share one clean denominator, so no type-level test is reported (the pooled test and GLMM in Appendix G are the correct frame).

Ambience (separate session). An earlier run (`j1_corr_r1`) collected $n = 37$ ambience-splice corrupted items with a flag rate of 16.2 %, 6.9 pp above the four-type corrupted rate (9.3 %). Because the session differs, the difference cannot be attributed to type versus session, and ambience is excluded from the primary corrupted arm.

Table 7: J1 register-only flag rate by type vs. the shared clean baseline (6.9 %, $n = 582$).

Type	C2 %	n_{C2}	Δ pp	95 % CI
muted_source	14.3	49	+7.4	[−2.6, +17.4]
room_mismatch	10.9	46	+4.0	[−5.2, +13.2]
ambience_splice	10.6	66	+3.7	[−4.0, +11.4]
foley_swap	14.5	55	+7.7	[−1.9, +17.2]
hallucinated_sound	9.7	62	+2.8	[−4.8, +10.4]

Pairing verification (register-only probe). For every type, the number of corrupted sev3 clips equals the number of shared clips ($n_{\text{corr}} = n_{\text{shared}}$): every corrupted clip has a register-only counterpart. Register-only items without a corrupted pair (ambience 38, hallucinated 29, foley 20, muted 5, room 0) are excluded under `PAIRED_ONLY`; for all types except room this changes the training distribution relative to an unpaired run.

Ambience $\Delta d'$ interval width. The ambience $\Delta d'$ CI $[-1.188, +0.388]$ around $+0.044$ is not caused by AUC clipping: the upper bootstrap AUC for comparison B is 0.989, far below the clip at $1-10^{-6}$, and the comparison-B AUC interval $[0.828, 0.989]$ is only $1.65\times$ asymmetric. The additional asymmetry in $\Delta d'$ arises at $n = 29$ pairs when the A and B out-of-fold AUCs swing in opposite directions within a resample; it reflects small-n uncertainty.

Overfitting diagnostic (foley, hallucinated). Both types memorised training data in the severity probe, where foley test d' collapsed to chance under PCA50. The register-only contrast is a different problem (presence vs. absence of the inserted event, not severity gradation), so overfitting was re-evaluated on it. A sweep over $C \in \{0.01, 0.1, 1, 10\}$ and dimensionality $\{\text{raw}, \text{PCA50}, \text{PCA200}\}$ gives positive test d' in all 12 conditions: 0.71–0.85 (foley) and 0.72–0.86 (hallucinated). Training d' reaches ceiling ($+6.72$, AUC clipped at $1-10^{-6}$) in 11/12 foley and 8/12 hallucinated conditions, but test d' does not collapse under PCA50 for either type (foley 0.76–0.85; hallucinated 0.72–0.83), the opposite of the severity-probe behaviour. Foley's primary bootstrap CI only barely excludes zero ($P{>}0 = 0.977$); its retention rests additionally on this 12-condition consistency. Hallucinated's CI fully excludes zero ($P{>}0 = 1.000$).

Verdict and transition tables. Table 8 gives J2's verdict-logit d' for sev1 vs. sev3 (all CIs span zero; negative point estimates are sampling noise and no inversion is claimed). Table 9 gives all transition estimates and CIs; T2 and T3a are reported for completeness only (disqualified pool, §4.1).

Table 8: Qwen2.5-Omni-3B verdict-logit d' for sev1 vs. sev3 (paired clips). All CIs span zero; negative point estimates are sampling noise, not signal inversion.

Type	n	Verdict d' [95%CI]
room	46	−0.05 [-0.15, 0.06]
ambience	29	−0.15 [-0.34, 0.00]
muted	44	−0.02 [-0.12, 0.08]
hallucinated	34	−0.02 [-0.16, 0.12]
foley	37	0.00 [-0.17, 0.20]

F PAIRING GAIN

Within-clip pairing increases severity-regression r^2 by $6.5\times$ for room (r: $0.265 \to 0.676$) and $2.1\times$ for ambience (r: $0.436 \to 0.630$), both from the same run; hallucinated shows the largest gain (unpaired r essentially zero). Severity-probing designs that pool items across clips without within-clip pairing are underpowered for types with high within-clip variance, and will systematically underestimate the representation–action gap.

Table 9: Four-transition decomposition (L35-anchored, sev1 vs. sev3, joint bootstrap $N = 500$). Bold = 95% CI excludes zero. T1 uses corrected within-pool audio-token positions; values are from `probe_four_transitions_corrected.json`. †T2 and T3a compare cross-position representations that diverge under TMRoPE interleaving; both are withdrawn. T3-LP and T4 use clean pools. Consequence: room and ambience have no valid transition with CI excluding zero.

Type	n	T1 interface	T2† depth	T3a† route	T3-LP route	T4 readout
room	46	$-0.08\,[-0.25, 0.16]$	$0.15\,[-0.11, 0.69]^{\dagger}$	$0.82\,[0.38, 1.37]^{\dagger}$	$0.20\,[-0.09, 0.88]$	$0.17\,[-0.33, 0.50]$
ambience	29	$0.00\,[-0.44, 0.22]$	$-0.08\,[-0.89, 0.71]^{\dagger}$	$1.60\,[0.73, 2.35]^{\dagger}$	$0.45\,[-0.42, 1.02]$	$0.11\,[-0.36, 0.85]$
muted	44	$-0.01\,[-0.23, 0.14]$	$0.62\,[0.08, 1.02]^{\dagger}$	$0.48\,[0.07, 0.97]^{\dagger}$	$0.21\,[-0.30, 0.51]$	$0.47\,[0.10, 0.82]$
hallucinated	34	$-0.08\,[-0.20, 0.29]$	$0.13\,[-0.20, 0.69]^{\dagger}$	$0.02\,[-0.76, 0.53]^{\dagger}$	$-0.03\,[-0.85, 0.39]$	$0.28\,[-0.14, 0.94]$
foley	37	$-0.03\,[-0.20, 0.13]$	$-0.14\,[-0.72, 0.45]^{\dagger}$	$0.17\,[-0.53, 0.89]^{\dagger}$	$-0.43\,[-0.78, 0.52]$	$0.09\,[-0.51, 0.60]$

Table 10: Pairing gain: LLM regression under unpaired vs. clip-demeaned (paired) analysis. Both figures from the same run (`probe_dprime_results.json`). r^2 ratio = paired r^2 / unpaired r^2; do not report as percentage of r.

Type	n	r_{unpaired}	r_{paired}	r^2_{unpaired}	r^2_{paired}	r^2 ratio
room	46	0.265	0.676	0.070	0.457	6.5×
ambience	29	0.436	0.630	0.190	0.397	2.1×
muted	44	0.565	0.886	0.319	0.785	2.5×
hallucinated	34	0.032	0.585	0.001	0.342	$\gg$100×
foley	37	0.150	0.324	0.023	0.105	4.6×

G WHERE THE SEVERITY SIGNAL GOES

Pool disqualification. `hidden_audio_mean` (pool 0) and `hidden_video_mean` (pool 1) have $\cos > 0.95$ at all 36 transformer layers (range 0.960–0.997, pooled over 1733 items; computed from item-mean pooled representations); neither pool can support a cross-pool audio-specific claim. T1 (AEM → L0_audio) is a within-pool transition: both endpoints track the same audio-token positions (token id 151646), so the encoder-to-LM interface is measurable within that scope; the TMRoPE interleaving objection does not apply to T1. T2 (L0_audio → L35_audio) and T3a (L35_audio → L35_dec) compare positions whose LLM-layer representations diverge under TM-RoPE interleaving; both are withdrawn.

FOUR-TRANSITION DECOMPOSITION

Table 11: Transition definitions. T1 uses exact audio-token positions (within-pool; corrected) and is interpretable. †T2 and T3a use `hidden_audio_mean`, which §4.1 disqualifies; both are withdrawn (TMRoPE interleaving makes cross-position comparisons non-separable). T3-LP and T4 use clean pools. Consequence: room and ambience have no valid transition with a CI excluding zero.

Label	Varies	From	To	Pool
T1 interface	processing stage	AEM	L0 audio-mean	audio-token pos. (corrected)
T2 depth†	depth	L0 audio-mean	L35 audio-mean	audio-mean†
T3a route†	position	L35 audio-mean	L35 DEC	audio-mean†
T3-LP route	position	L35 LP	L35 DEC	last-prompt
T4 readout	repr. to output	L35 DEC	verdict logit	decision

The T1 (AEM→L0) and T2 (L0→L35) transitions characterise encoder–LLM interface and depth; PAL Alex et al. (2025) is the closest prior architectural analysis of this interface.

Severity is encoded before the LLM for room ($n = 46$, AEM $d' = 2.12$ [1.52, 2.72]), ambience ($n = 29$, 1.92 [0.97, 3.03]; room is the lead exemplar due to narrower CI), and muted ($n = 44$, 1.81 [1.39, 2.57]); hallucinated (0.59 [0.27, 1.19]) and foley (0.32 [−0.65, 0.83]) are underpowered, not null. Verdict d' is indistinguishable from zero for all five types (Table 8). The d' chain figure is in §5.5.

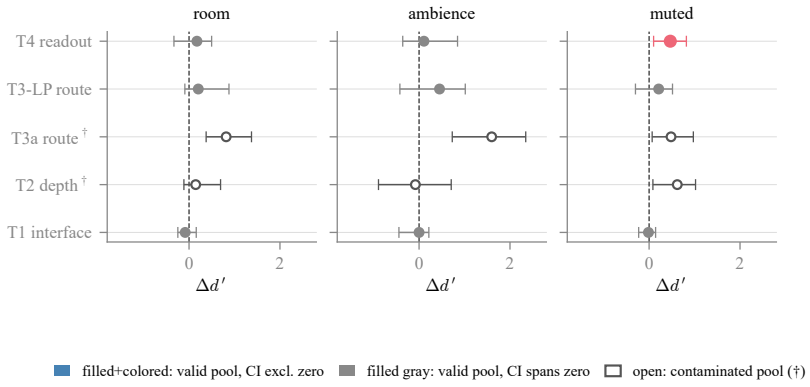


Figure 4: Four-transition decomposition for room, ambience, muted. Open circles: withdrawn audio-mean pool (T2, T3a). Filled gray: valid pool, CI spans zero. Filled colour: valid pool and CI excludes zero (muted T4 only). Full CIs in Table 9.

Transitions. T1 (encoder→L0, within-pool) is near zero for all five types (all CIs spanning zero): no measurable loss at the encoder–LM interface. For muted, T4 (readout) excludes zero ($+0.47$ [0.10, 0.82]): signal present at the decision token (L35 DEC $d' = 0.45$) is absent from the verdict logit. This matches the F3 underuse pattern of Peng & Nuchged (2026), extended from speech prosody to non-speech acoustic corruption. For room and ambience, T3-LP (room $+0.20$ [-0.09, 0.88]; ambience $+0.45$ [-0.42, 1.02]) and T4 both span zero; T3a is disqualified (audio-mean pool, §4.1). The stage at which room and ambience lose severity signal is therefore unresolved. Hallucinated and foley: all transitions span zero (inconclusive, not null).

GLMM robustness (J1 register-only elevation). All five per-type register-only deltas are positive ($+2.8$ to $+7.7$ pp; sign test $p = 0.031$, exact one-sided). A Bayesian mixed-effects logistic regression (`flagged` ∼ `condition` + (1 | type), Laplace MAP) gives OR = 1.78 (pooled OR 1.83, 95 % CI [1.12, 2.96]); the type random-intercept SD is at the boundary (0.002, [0.000, 0.018]), which with $k = 5$ groups is expected regardless of true heterogeneity. Per-type comparisons share one clean denominator and are individually underpowered (13–54 % power); they are reported descriptively in Appendix E.

Table 12: Column increments: within-type AUC minus mean off-diagonal AUC in that column (fold-level t-CI, $df = 4$). A negative increment means other probes detect this type better than the specialist. **Bold** = CI excludes zero or point estimate is negative (muted). Sorted by increment descending.

Eval type	Diagonal AUC	Off-diag col mean	Column increment [95% CI]
foley	0.911	0.822	0.089 **[0.068, 0.109]**
hallucinated	0.888	0.827	0.062 **[0.023, 0.101]**
room	0.872	0.823	0.049 [-0.004, 0.102]
ambience	0.927	0.885	0.042 [-0.001, 0.086]
muted	0.840	0.850	−0.011 **[-0.062, 0.042]**

Column increments and converging split. Column increments exclude zero only for foley ($+0.089$) and hallucinated ($+0.062$); for room, ambience, and muted, training on the target type gives no measurable advantage over training on an unrelated type, and muted's increment is negative. The two types with a type-specific probe component (foley, hallucinated) are exactly the two with the weakest encoder magnitude encoding (AEM $d' = 0.32$ [-0.65, 0.83] and 0.59 [0.27, 1.19]); the three without (room, ambience, muted) have strong encoder magnitude encoding ($d' = 1.81$–2.12). The split maps onto discrete event insertion (foley, hallucinated: binary presence) versus global acoustic transform (room, ambience, muted: continuous, gradable magnitude).

H FULL LAYER CURVES

Figure 5 shows d' at every transformer layer (0–35) for two clean pools (LP = last-prompt token, DEC = decision token) per corruption type. The pre-LLM audio encoder (AEM) d' is the horizontal dotted reference. Source: `results/probe_severity_clean_pools_results.json` (`scripts/probe_severity_clean_pools.py`). Figure generated by `scripts/fig_layer_curves.py`.

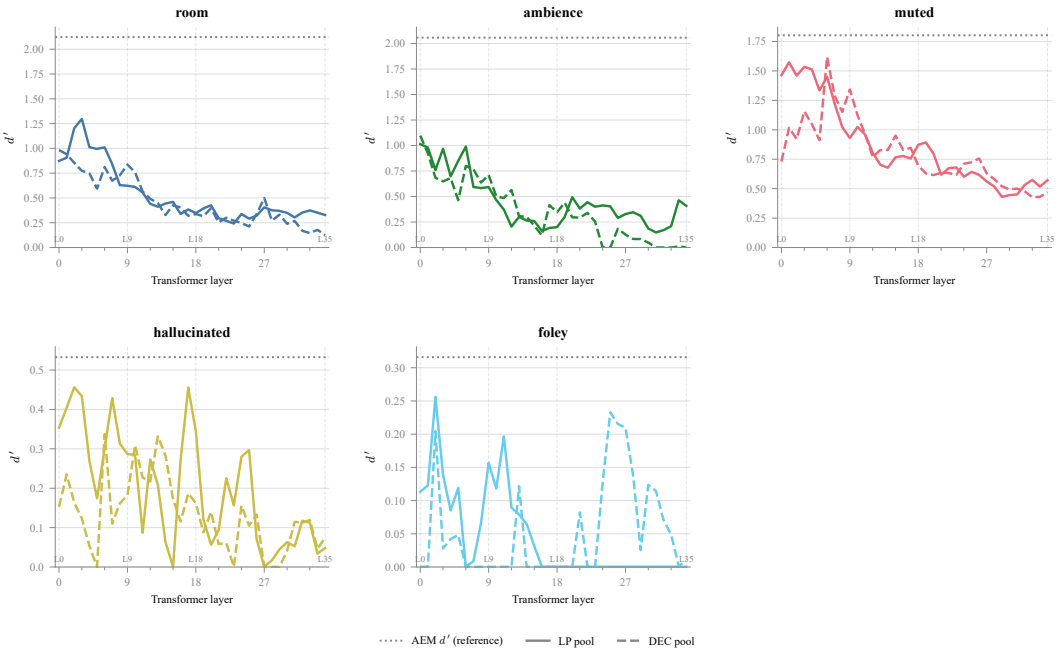


Figure 5: Per-type d' layer curves for LP (solid) and DEC (dashed) pools across all 36 transformer layers. Horizontal dotted line: AEM d' (pre-LLM audio encoder reference). Vertical dashed lines mark L0, L9, L18, L35. Types ordered as in the main text (room, ambience, muted, hallucinated, foley). Depth-decay is visible for room and ambience (peak near L0–L6, decay thereafter); muted maintains signal through L35 DEC, consistent with the T4 readout result.

Reading guide. Both LP and DEC curves being below AEM at deep layers indicates that the signal available at the pre-LLM encoder is not retained proportionally into the LLM representation. Room and ambience peak near L0–L6 then decay; the DEC pool at L35 is near the LP baseline, consistent with T3-LP spanning zero (§G). Muted source retains a non-trivial DEC signal at L35, consistent with T4 excluding zero. The early-layer peaks are also why argmax-layer selection inflates $\Delta d'$(LLM $\rightarrow$ verdict) (§G).

I DETECTION AND MAGNITUDE ARE SEPARABLE CONSTRUCTS

Two probe analyses address different research questions; using either as a proxy for the other mischaracterises the results.

Artifact presence (AUC). AUC for corrupted-vs-clean detection (`probe_pass3.py`) measures whether the encoder separates pipeline-processed audio from originals regardless of severity. *All values are upper bounds*: they are corrupted-vs-clean without register-only controls, and the probe direction is largely type-agnostic ($[0.754, 0.91]$ cross-artifact transfer; §5.4). The register-only probe (§5.3) shows this direction is a generic acoustic-anomaly direction plus, for room and foley, a pipeline footprint.

Severity magnitude (d'). d' for sev-1-vs-sev-3 paired probing (`probe_dprime_analysis.py`) measures whether the encoder represents physical corruption *magnitude* as a graded signal. A type can be strongly detectable while its magnitude representation is weak or absent.

The divergence is evidence, not artefact. The two rankings diverge along the discrete-insertion vs. global-transform axis (Table 13).

Table 13: Artifact presence (AUC) vs. severity discrimination (d') at the pre-LLM audio encoder. Rankings are inverted for foley and room.

Type	Presence AUC†	Rank	Severity d'	Rank
room_mismatch	0.934	3	2.12	1
ambience_splice	0.970	2	1.92	2
muted_source	0.825	4	1.81	3
hallucinated_sound	0.805	5	0.59	4
foley_swap	0.974	1	0.32	5

†Upper bound (corrupted-vs-clean; register-only controls in §5.3).

Global transforms (room, ambience, muted) score well on *both* questions: the event is detectable and the physical magnitude is graded. Discrete insertions (foley, hallucinated) score high on presence AUC but low on severity d': the event is detectable but its CLAP-distance magnitude carries no graded signal at the encoder, consistent with the working hypothesis in §5.4.

Prevalence sweep: no rank inversions in our judge set. Per-type HR and FAR were computed from J2 verdicts in `qwen_full_results_backfilled.jsonl`. Accuracy, balanced accuracy, F1, MCC, and Youden's J were evaluated at every $\pi \in [0.01, 0.99]$ for all five corruption types, and rank crossings were checked for every type pair. No metric produces a rank inversion for any type pair at any prevalence (Figure 6). The cause is structural: all types share one clean-item pool, so FAR is identical across types (FAR $= 0.763$; 380 of the 498 clean originals in this severity run received verdict `true` (`type = none`, `is_corrupted = false` in `qwen_full_results_backfilled.jsonl`; not C1 re-encoded); this pool is distinct from the $n = 582$ J2 clean arm of §5.2, whose rate is $72.0\,\%$). With constant FAR, every prevalence-dependent metric reduces to ranking types by HR, and the HR ordering (foley > room > muted > hallucinated > ambience) is invariant to π. Ranking divergence requires heterogeneous FAR — different clean corpora or different judge models.

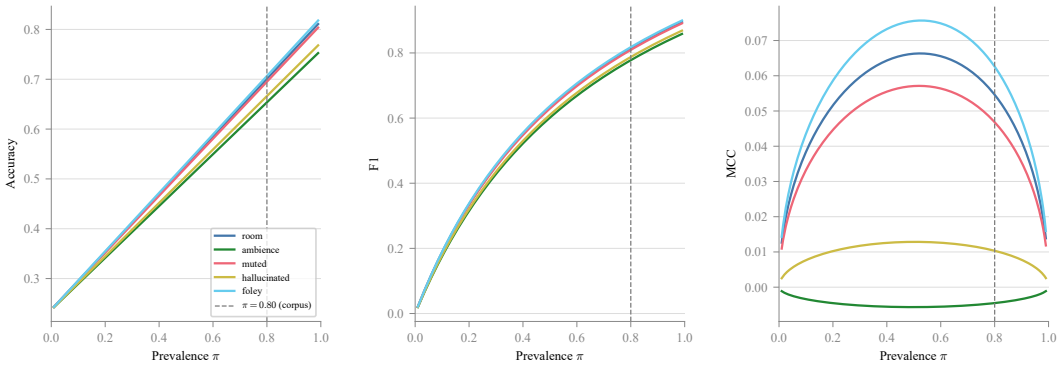


Figure 6: Accuracy, F1, and MCC as a function of prevalence π for each corruption type, using J2 HR and FAR from `qwen_full_results_backfilled.jsonl`. Lines are parallel: no rank inversions across $[0.01, 0.99]$. Vertical dashed line: $\pi = 0.80$ (our corpus).

J MEASUREMENT-NOISE SOURCES

Three sources of measurement noise: (1) clip-level variance in corruption severity (non-fixed-tier types); (2) probe fold variance (fold AUCs [0.945, 0.846, 0.869, 0.931, 0.904]); (3) within-clip severity variance, accounted for by the paired bootstrap design.