WHERE DOES THE AUDIO JAILBREAK LIVE? A CONTROLLED FREQUENCY-DEPTH AUDIT OF AUDIO-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent attacks and defenses for audio-language models (ALMs) assume that the jailbreak signal in an adversarial waveform concentrates in particular mel bands. We test this assumption with a frequency-depth occlusion protocol: we remove band-limited components of an additive perturbation in the STFT domain, rerun the victim, and score each condition on judge-adjudicated attack success and on the layer-wise divergence of decoder audio-span representations. The protocol is paired with controls that any band-localization claim should pass: bandwidth-matched random bands, equal-Hz and equal-energy partitions, matched-energy scattered removal, and uniform rescaling. For AdvWave-P against Qwen2-Audio (520 AdvBench prompts, attack success 0.77), a standard 8-band mel analysis points to the upper bands, but this locus does not survive the controls. A random contiguous region of the same bandwidth breaks the attack as effectively, every band is about equally load-bearing under an equal-energy partition, and a band's share of perturbation energy alone reproduces the occlusion ranking (Spearman 0.95). The attack does depend on how energy is removed: masking one contiguous band lowers success to 0.05 to 0.29, whereas removing the same energy uniformly leaves it at 0.75 to 0.79, and no single band or pair carries the attack even when it is re-optimized on that support. Applied to this attack, the band selections of ALMGuard and GRM do no better than matched random selections. In depth, band importance and representational divergence are aligned from the decoder entrance onward and remain so after adjusting for band energy, but we find no behavior-specific depth locus. The frequency findings replicate on three further prompt corpora. For this attack, the jailbreak is a broadband property of the jointly optimized perturbation, and band-selective claims should be reported against energy- and bandwidth-matched controls.

1 INTRODUCTION

Audio-language models (ALMs) such as Qwen2-Audio Chu et al. (2024) and LLaMA-Omni Fang et al. (2025) extend LLMs with speech frontends. The new interface widens the safety surface: a harmful instruction refused in text can be accepted if it arrives as adversarial audio. AdvWave Kang et al. (2025) and AdvWave-P Jin et al. (2025) drive high attack success on AdvBench Zou et al. (2023), and similar vulnerabilities have been reported across open and proprietary ALMs Song et al. (2025); Peri et al. (2024); Yang et al. (2025).

Several recent works suggest that the jailbreak-relevant part of an adversarial waveform is concentrated in a subset of mel bands. GRM (Wang et al., 2026) perturbs only bands with a high attack-to-utility gradient ratio, ALMGuard (Jin et al., 2025) learns a sparse mask over sensitive mel bins as a defense, and Gupta et al. (2025) identify a coarse 1 to 8 kHz band for universal audio jailbreaks. If the jailbreak does live in specific bands, defenses can target them. If the apparent concentration is instead produced by how the spectrum is partitioned and how the perturbation's energy is distributed, such defenses target an artifact. None of these works reports controls that separate the two readings, and none examines where in the language decoder the band structure registers.

We ask the question directly for one well-studied attack: which parts of the perturbation are necessary and sufficient for the jailbreak, and at which decoder depth does the representation reflect that structure? Our protocol removes band-limited components of the perturbation in the STFT domain, feeds the reconstructed waveforms back through the victim, and compares jailbreak behavior

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.

and decoder representations across conditions. The central design choice is the control suite: every band-level effect is compared against a bandwidth-matched random band, re-measured under equal-Hz and equal-energy partitions, and contrasted with removing the same energy without spectral selectivity. For AdvWave-P on Qwen2-Audio, the answer is that the jailbreak lives in no particular band and at no particular depth.

In summary, this paper makes four contributions (Table 1 gives the corresponding numbers).

- **A frequency-depth occlusion protocol with a localization control suite** (Section 3). Exactly invertible STFT-domain occlusion of an additive perturbation, scored on judge-adjudicated ASR and layer-wise audio-span divergence, with the matched controls above plus energy-matched scattered removal and an energy-share baseline.
- **The apparent band locus is a partition and energy artifact** (Section 5.1). The upper-band necessity of a standard eight-band analysis is matched by a random contiguous region of equal bandwidth, vanishes under equal-Hz and equal-energy partitions, and is reproduced by band energy share alone (Spearman 0.95).
- **Removal structure matters, band identity does not** (Section 5.2). A contiguous gap breaks the attack while removing the same energy uniformly does not, no band subset suffices even when re-optimized on it, and the band selections of ALMGuard and GRM fall within the range of matched random selections.
- **Alignment without localization in depth** (Sections 5.3 and 5.4). Band necessity and audio-span divergence are aligned from the projector output onward, also after energy and prompt adjustment, but the rise is not behavior-specific, does not clear an autocorrelation-preserving null, and single-layer patching restores the jailbreak only at the decoder entrance.

2 Related Work

Audio jailbreaks on ALMs. ALMs (Qwen2-Audio Chu et al. (2024), LLaMA-Omni Fang et al. (2025)) pair a log-mel frontend with an instruction-tuned LLM, but their text-side safety transfers only weakly to the audio interface. SpeechGuard Peri et al. (2024) exposed brittleness to adversarial waveforms; AdvWave Kang et al. (2025) introduced a dual-phase optimization that handles encoder gradient shattering and exceeds 90% ASR on AdvBench Zou et al. (2023); Jin et al. Jin et al. (2025) proposed the ALMGuard defense and, as the additive attack it defends against, AdvWave-P; AudioJailbreak Song et al. (2025) and Yang et al. Yang et al. (2025) broaden evaluation across attacks and ALMs. All of this work treats the perturbation as a single waveform-level object and leaves its internal structure unexamined.

Classical audio-adversarial and psychoacoustic attacks. Our analysis builds on a longer line of ASR-focused work. Carlini and Wagner Carlini & Wagner (2018) constructed targeted audio adversarial examples; Qin et al. Qin et al. (2019) made them imperceptible and robust with a psychoacoustic masking constraint; Schönherr et al. Sch"onherr et al. (2019) hid attacks below the masking threshold. A standing result is that perturbation energy is pushed into spectral regions below the masking threshold, frequently the mid-to-upper range, which is often read as band-selective jailbreak energy. AdvWave-P itself constrains only $\|\boldsymbol{\delta}\|_\infty$ and has no explicit masking-threshold term, so our findings speak to the analysis of such attacks rather than to the psychoacoustic mechanism as such.

Frequency-domain structure and defenses. GRM Wang et al. (2026) ranks mel bands by attack-vs-utility ratio and perturbs only the top subset; ALMGuard Jin et al. (2025) takes the defense view, learning a sparse mask over sensitive mel bins; Gupta et al. Gupta et al. (2025) probe universal audio jailbreaks with encoder attention and identify a coarse 1–$8\,$kHz band; and the band-stop/low-pass defense family (e.g. Yang et al. Yang et al. (2019)) motivates the single-notch hypothesis. All report band-selective jailbreak energy at the frontend, yet none reports where the signal registers inside the LLM decoder, which is the depth component this paper adds.

Mechanistic interpretability and safety evaluation. Our toolkit is borrowed from text-only interpretability: representation-space probing Alain & Bengio (2017); Belrose et al. (2023), and linear CKA Kornblith et al. (2019) as a geometry check in the appendix. Our causal check borrows the activation-patching and causal-mediation toolkit Vig et al. (2020); Meng et al. (2022); Wang et al. (2023); Conmy et al. (2023).

Removal-based attribution and its pitfalls. Occlusion is a removal-based attribution. In vision, removal has been shown to conflate a feature's information content with the effect of the removal

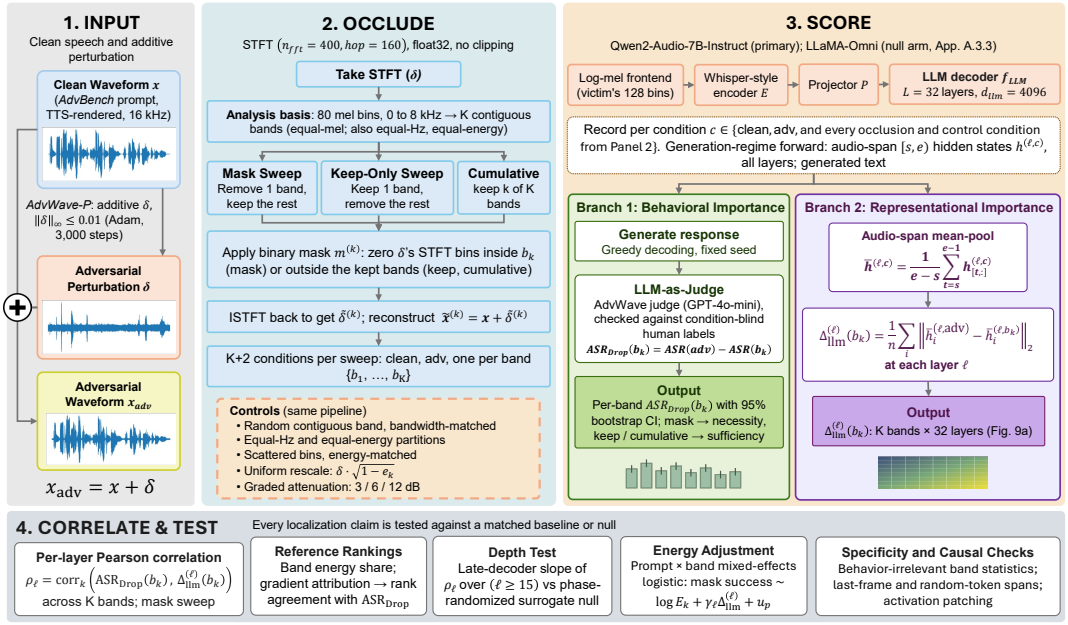


Figure 1: Protocol overview. **Occlude:** the perturbation δ is decomposed in the STFT domain and each analysis band is zeroed (mask) or retained (keep-only, cumulative), then reconstructed; the same pipeline generates the control conditions that any band-localization claim should pass, namely bandwidth-matched random contiguous bands, equal-Hz and equal-energy partitions, energy-matched scattered removal, uniform rescaling by $\sqrt{1-e_k}$, and graded attenuation. **Score:** each condition is played to the victim ALM and scored on judge-adjudicated ASR and on the layer-wise audio-span divergence $\Delta^{(\ell)}_{\text{llm}}$. **Correlate and test:** per-band ASR drop is compared against reference rankings (band energy share, gradient attribution) and correlated with $\Delta^{(\ell)}_{\text{llm}}$ at every decoder layer; the depth trend is assessed against an autocorrelation-preserving surrogate null and an energy-adjusted prompt-by-band model, with specificity and causal checks. The figure shows protocol structure; all numbers are reported in Section 5.

itself: ROAR (Hooker et al., 2019) retrains after removal to separate the two, and sanity checks for saliency methods (Adebayo et al., 2018) show that attributions can be insensitive to what the model has learned. Our uniform-rescale and random-contiguous controls play the same role for spectral occlusion. We also discuss the known limits of single-layer patching in Section 5.4.

3 Methodology

We analyze a single audio jailbreak by removing frequency-localized components of the perturbation and measuring how behavior and decoder representations respond. For each sample the pipeline runs two measurement branches over the same conditions: behavioral importance (each band's contribution to attack success) and representational divergence (the decoder layer at which audio-span geometry reflects which band matters).

3.1 Setup and Mel-Band Occlusion

Let $\mathbf{x} \in \mathbb{R}^T$ be a clean 16 kHz harmful-instruction waveform from AdvBench Zou et al. (2023) and $\mathbf{x}_{\text{adv}} = \mathbf{x} + \boldsymbol{\delta}$ its AdvWave-P adversarial counterpart Jin et al. (2025), optimized with $\|\boldsymbol{\delta}\|_\infty \le 0.01$ (Adam lr 10^{-3}, 3,000 iterations). We compare behavior and hidden states on $\mathbf{x}$, $\mathbf{x}_{\text{adv}}$, and band-occluded variants of $\mathbf{x}_{\text{adv}}$, so any ASR change after occlusion is attributable to the removed band. The victim maps the waveform to a log-mel spectrogram with its own 128-bin filterbank, encodes it into projected audio tokens, and processes them with an L-layer LLM decoder. For the analysis we use a *separate* analysis basis: we partition the spectrum into K contiguous bands whose edges are derived from an 80-bin mel scale over 0–8 kHz (K=8 on Qwen, 10 bins per band; a K=16 refinement halves each block to 5 bins; Omni uses its own 128-bin scale, K=13). This analysis basis is distinct from either victim's frontend filterbank, which is applied downstream to the reconstructed waveform; all band comparisons are stated in Hz, the invariant under the analysis basis. For band b_k

we map its Hz boundaries to STFT bins and reconstruct

$$\tilde{\mathbf{x}}^{(k)}_{\text{mask}} = \mathbf{x} + \text{ISTFT}\left(\text{STFT}(\boldsymbol{\delta}) \odot \mathbf{m}^{(k)}\right), \tag{1}$$

where $\mathbf{m}^{(k)}$ is 0 on the bins covered by b_k and 1 elsewhere, zeroing the perturbation's energy inside b_k while leaving the clean waveform unchanged. The *mask* sweep measures necessity; the *keep-only* sweep measures sufficiency, formalized in Section 3.2.

Reconstruction and reference evaluation. All reconstructions use a single float32 no-clip STFT-mask-ISTFT path (exact inversion, no 16-bit quantization; Table 4).

Protocol interface. The method takes a clean waveform and an additive, spectrally co-located perturbation, and needs white-box access to decoder hidden states only for the depth branch.

3.2 Behavioral Importance

For each condition $c \in \{\text{clean}, \text{adv}, \text{mask}(b_k), \text{keep}(b_k)\}$ we generate the model response under the same decoding as AdvWave and adjudicate with the LLM-as-a-judge pipeline (Section 4). Attack success is $\text{ASR}(c) = \frac{1}{n}\sum_i \mathbb{1}[\text{judge}(c_i) = \text{jailbroken}]$, and behavioral importance of band b_k is $\text{ASR_Drop}(b_k) = \text{ASR}(\text{adv}) - \text{ASR}(b_k)$. All $K{+}2$ conditions share a single fixed seed (and greedy decoding), and the layer-wise probes (Section 3.3) are computed from the generation-regime forward (a single full-input forward on the exact generation input, before the first sampled token, recording `output_hidden_states` and the audio-token-span activations directly), so the behavioral ASR and representational divergence for a condition are deterministic given the seed, though they come from two different passes. We collapse each band's necessity into positive importance $\mathcal{I}^{+}(b_k) = \text{ASR}(\text{adv}) - \text{ASR}(\text{mask}(b_k))$ (large when removing b_k breaks the attack) and its sufficiency into negative importance $\mathcal{I}^{-}(b_k) = \text{ASR}(\text{adv}) - \text{ASR}(\text{keep}(b_k))$ (small when b_k alone sustains it); we report bootstrap 95% CIs (settings in Section 3.4) and the Spearman correlation between the two rankings.

3.3 Layer-wise Representation Probes

We hypothesize that the geometry of the audio-token span reflects band importance at the layers where the decoder representation is most tightly coupled to the attack. For every condition we save the projector output, all decoder residual streams, and the audio-token span $[s, e)$ from the generation-regime forward. At layer ℓ we mean-pool the span, $\bar{\mathbf{h}}^{(\ell,c)} = \frac{1}{e-s}\sum_{t=s}^{e-1} \mathbf{h}^{(\ell,c)}_{[t,:]}$, and define per-band divergence $\Delta^{(\ell)}_{\text{llm}}(b_k) = \frac{1}{n}\sum_i \|\bar{\mathbf{h}}^{(\ell,\text{adv})}_i - \bar{\mathbf{h}}^{(\ell,b_k)}_i\|_2$, then correlate band ASR drops with band divergences per layer, $\rho_\ell = \text{corr}_k\big(\text{ASR_Drop}(b_k), \Delta^{(\ell)}_{\text{llm}}(b_k)\big)$ (Pearson; Spearman as a sensitivity). We summarize depth by the late-decoder ($\ell \geq 15$) slope of ρ_ℓ, tested as in Section 3.4. This tests whether ρ_ℓ rises, not whether the rise is specific to the safety behavior (Sections 5.3 and 5.4).

3.4 Statistical Inference and Uncertainty

To validate the judge we hand-label a condition-blind set (220 responses: the 120-row adversarial cell, a 40-row clean control, and 60 disagreement-stratified rows; single annotator, no second annotator available, Section 7; composition, fixed seed, and blind key in Appendix A.3.2). On the 120-row adversarial cell the judge shows low false-positive (FPR $\approx$0.10 among human negatives) and slightly low sensitivity; a Rogan-Gladen misclassification correction at the n=520 headline ASR_{adv}=0.767 gives $\text{ASR}^{*}_{\text{adv}}\approx$0.87 and leaves every qualitative contrast unchanged (Section 7).

Depth trend and its nulls. The depth statistic is ρ_ℓ over the K bands at layer ℓ, summarized by its slope over the late decoder ($\ell \in [15, 31]$). Adjacent residual-stream layers are strongly autocorrelated (effective $n \approx 4.3$ over 17 layers), so a layer-index permutation test, which destroys this autocorrelation, is used only as an ordering check. For significance we use phase-randomized surrogates of the ρ_ℓ sequence, which preserve its power spectrum (Theiler et al., 1992; Schreiber & Schmitz, 2000). Because a smooth monotone rise contributes low-frequency power that the surrogates inherit, this test is conservative for trends, and a non-rejection is uninformative. Prompt-level uncertainty on $\Delta\rho = \rho_{30} - \rho_0$ is a percentile bootstrap over prompts (2000 resamples at $n = 520$).

Energy investigation and partition sensitivity (prompt-by-band mixed-effects model). Because $\text{corr}(\Delta^{(\ell)}, E)\approx$0.92–0.98, a per-layer regression of mask success on divergence and band energy is ill-conditioned (VIF $10^{2.7}$–$10^{3.1}$); we therefore move to the prompt-by-band level. Over the 520×8 observations we fit, per layer ℓ, a mixed-effects logistic $\text{logit}\, P(\text{mask success}{=}1) = \alpha + \beta_\ell \log E_k + \gamma_\ell \Delta^{(\ell)}_{\text{llm}}(b_k) + u_p$ (random prompt intercepts u_p), well-conditioned with one continuous covariate beyond band energy, and report the energy-adjusted coefficient γ_ℓ per layer with a prompt-resampling (block) bootstrap CI, normalizing each layer's divergences by their per-sample

Question	Test	Answer
Is there a privileged band? (§5.1)	Mask sweep vs. matched random band; equal-Hz/energy partitions	**No.** Random band equally effective; equal-energy bands equally load-bearing; energy share $\rho_s{=}0.95$
Does removal structure matter? (§5.2)	Band mask vs. uniform rescale; scattered removal	**Yes.** Mask 0.05–0.29 vs. rescale 0.75–0.79
Is any band subset sufficient? (§5.1)	Keep-only; cumulative addition	**No.** Attack returns only with (nearly) all bands
Do prior band selections localize it? (§5.2)	ALMGuard mask and GRM rule vs. matched random	**No.** Within matched-random range
Are behavior and representation aligned? (§5.3)	ρ_ℓ; permutation; energy-adjusted model	**Yes,** from the projector output onward
Is there a behavior-specific depth locus? (§5.3)	Band statistics; surrogate null; patching	**Not established**
Do findings replicate? (§5.5)	JailbreakBench, MaliciousInstruct, StrongREJECT	**Partly.** Frequency findings yes; depth rise not significant in any corpus

Table 1: Summary of findings (Qwen2-Audio, AdvWave-P, $\epsilon{=}0.01$, $n{=}520$ unless noted).

standard deviation for comparability (Section 5.3). A leave-one-band-out analysis reads the band-level statistic as conditional on the partition (LOO spread 0.172, $\approx 2\times$ the bootstrap SE); quantifying the partition-conditional part of the prompt-level interval would require re-running mask inference on fresh partitions (about one GPU-day), so we state this rather than leave the interval unqualified (Section 7).

4 Experimental Setup

Victim models, attack, and corpus. We use Qwen2-Audio-7B-Instruct Chu et al. (2024) (Whisper-large-v3 encoder, $L{=}32$, $d_{\text{llm}}{=}4096$) and LLaMA-Omni Fang et al. (2025) as victims, both with a 128-bin log-mel frontend distinct from our analysis basis (Section 3.1). We attack with AdvWave-P Jin et al. (2025), the additive variant of AdvWave Kang et al. (2025) (co-located full-waveform $\boldsymbol{\delta}$, adversarial budget and optimization in Section 3.1 and Figure 1), on AdvBench Zou et al. (2023) prompts. The $K{=}8$ mask and keep sweeps, the $K{=}16$ sweep, and all depth estimators run on the full $n{=}520$ corpus; the epsilon and resolution-ladder analyses remain at $n{=}50$. The main sweep uses greedy decoding (a single deterministic generation per prompt per condition); a decoding-stability re-measurement under stochastic decoding is reported separately (Appendix A.7) and is not the headline ASR. Each sample yields $K{+}2$ conditions (clean, full adversarial, K occluded variants), adjudicated by the judge below. A representation-space visualization (Appendix A.2), the LLaMA-Omni null arm, and the cross-model transfer matrix use the same conditions; Omni is read strictly as a null (its small margin and judge noise floor make per-band effects unmeasurable in principle), and cross-victim waveforms are fed unchanged through the target with per-band sweeps on the source-side bands so every cell of the $2{\times}2$ ASR matrix carries a full per-band profile.

Judge. We use the AdvWave LLM-as-judge Kang et al. (2025) at its default threshold with GPT-4o-mini as the scorer, so our ASR numbers are directly comparable to prior work; the prompt, model, threshold, and decoding are the shipped defaults of the released implementation. As a sensitivity reference we additionally adjudicated every response with a stricter proprietary judge (GPT-5.6-sol); those labels appear only in the judge-agreement analysis (Table 3) and the Limitations, not as the headline ASR, because the stricter judge applies a higher threshold on borderline completions. We report the per-band table under both judges in the judge-agreement analysis so the headline contrasts are visible under either scorer; the clean-renderer specification is in Appendix A.4.

5 Results and Discussion

We report the full-corpus ($n{=}520$) evaluation on Qwen2-Audio-7B-Instruct, where the unaltered adversarial waveform jailbreaks the model: $\text{ASR}_{\text{clean}}{\approx}0.008$ and $\text{ASR}_{\text{adv}}{=}0.77$ (both $n{=}520$), so the protocol operates on a working, high-margin jailbreak on which band necessity and depth alignment can be disentangled. Every ASR_Drop carries the judge uncertainty of Section 7; all per-band

behavioral intervals are 95% bootstrap (settings in Section 3.4). Table 1 summarizes the findings; each row is developed in the section indicated.

5.1 Per-Band Behavioral Importance

Figure 2 reports both sweeps at full scale (n=520). On Qwen2-Audio the upper mel bands are necessary: removing b_8=[70, 79] drops ASR from 0.77 to 0.05 (+0.71, CI [0.67, 0.75]), with b_5–b_7 within noise of it, so we report a broad mid-to-upper 1.7–8 kHz region (79% of the spectrum), not a peak. This necessary region does *not* separate from a same-bandwidth random contiguous baseline (over 20 draws ASR drives to a mean of 0.015, indistinguishable from the 0.05 left by the necessary region), so removing *any* half of the analyzed spectrum breaks the attack. In the keep-only sweep no single band is sufficient; because keep-only saturates, we measure cumulative band-addition directly; the attack recovers only once the necessary set is retained.

Resolution controls. On K=8 bands of equal Hz width the necessity gradient collapses: every band's residual ASR falls to ≤ 0.30, so every equal-Hz band is necessary. The perturbation's energy density is high at low frequency (0–1000 Hz alone carries 23.0% of delta energy) and widest at high frequency, so every 1000-Hz band holds a critical share. Building contiguous bands at equal-energy quantiles of the corpus-mean $|\mathrm{STFT}(\delta)|^2$ spectrum (K=8; the decisive control), masking any single band drives ASR from 0.77 to 0.07–0.16, so *all* bands are nearly equally load-bearing. Refining the equal-energy partition to K=16, the single most load-bearing region is the narrow top-frequency band b_{15}=[7520, 7960] Hz near Nyquist (residual ASR 0.10, drop +0.67), the one region the coarse K=8 result weighted most heavily, resolving necessity into a non-contiguous set whose top element is this top narrow band (Appendix A.9). Read together, no frequency-specific necessity gradient survives once both bandwidth and energy are fixed; the K=8 "mid-to-upper necessity" is at least partly an artifact of the analysis partition. Retention by importance and by energy give nearly identical sufficiency curves (0.29 at k=7, full adversarial 0.78 only at k=8), so retention order is immaterial.

5.2 Protocol Diagnostics: Occlusion Fidelity and the Energy Confound

The protocol's inferences rest on the occlusion instrument and on what a band's energy means; full numbers are in Appendix A.3.3. Occlusion fidelity holds at the decibel level quantified there (STFT-mask-ISTFT and mel-filter leakage are both small fractions of the removed band's energy), so the contiguous region is not a leakage artifact. Causally, a matched-energy scattered control removes the same number of STFT bins and the same total removed energy as each contiguous mask but scatters them at random across the spectrum; across 20 draws this drives ASR below the scattered-null 95th percentile on all eight bands (largest at low frequency, shrinking toward the top), so the spectral *structure* of the removal carries signal beyond its amount: a contiguous gap is more damaging than the same energy scattered across the spectrum. This is a statement about contiguity, not about which band is removed. Attenuating rather than zeroing a band by $3/6/12$ dB recovers a monotone ASR response, so the instrument does not saturate. The per-layer divergence is a near-linear function of band energy (Appendix A.3.3), so the depth-alignment height is partly energy-driven; Section 5.3 reports the degenerate band-level energy-conditional partial together with the well-conditioned prompt-by-band energy-adjusted model that resolves it.

Global-rescale control: selective versus non-selective removal. The equal-energy and energy-share baselines above separate *how much* energy is removed from *where* it is removed. To isolate location at fixed removed-energy without any spectral selectivity, for each K=8 band we scale the whole delta by $\sqrt{1-e_k}$ (removing that band's energy fraction e_k spread uniformly across the spectrum) and re-run the n=520 greedy evaluation under the identical judge. Every global-rescale ASR lands at the adversarial baseline (0.75–0.79; uniform $\frac{1}{8}$ removal 0.76), a +0.49 to +0.70 gap above the matched band-mask ASR (0.05–0.29) on every band. A non-selective removal of the same energy thus *cannot* reproduce the band-localized breakage: rescaling cannot single out any band, so the variable that matters is the spectral *structure* of the removal, not which band is removed.

Named prior band selections vs. matched random. We re-expose the band selections of two prior methods to the protocol under a matched-random control (Appendix A.10). ALMGuard's sensitive-bin mask, applied as a band-stop to the existing deltas at n=520, drops ASR to 0.15, and GRM's key-band rule, reproduced by re-optimizing a delta on its projected support at n=20, reaches 0.75; both sit within the range of bandwidth- and energy-matched random selections (means 0.17 and 0.86), so neither named prior localizes the necessary band beyond matched random. We test only whether these selections localize the energy AdvWave-P needs, not the methods themselves.

5.3 Per-Layer Representational Importance and the Alignment Trend

The depth statistic is read at $K{=}8$ on the $n{=}520$ pooled divergences (Figure 8, appendix; the $K{=}16$ depth row is only at $n{=}50$, Table 2). The pooled $n{=}520$ alignment rises from $\rho_0{=}{+}0.568$ at the projector output to $\rho_{30}{=}{+}0.877$ at layer 30 (increment $+0.309$, clearing the ρ_0 CI $[0.27, 0.60]$); a paired prompt bootstrap on $\Delta\rho$ gives $+0.307$ (CI $[0.271, 0.342]$, excluding zero), and band-label permutation nulls on ρ_{30} are significant ($p{=}0.001$ at $K{=}8$, $p{<}5{\times}10^{-5}$ at $K{=}16$), so the top-layer alignment is significantly nonzero; the significance of the *trend* itself is assessed against the autocorrelation-preserving surrogate null below. Table 2 consolidates every estimator so each number traces to a row. The rise is confined to the pooled span; the last-audio-frame curve is flat and lower ($\rho^{\text{last}}{\approx}0.45$–$0.49$ across layers, and a length- and position-matched random-token span reproduces the pooled level), so the finding is specific to pooled-span representation.

No depth locus. The rise is smooth, but three observations argue against reading it as localization. *First*, the rise is not specific to the jailbreak: behavior-irrelevant band statistics (center frequency, bandwidth, clean spectral flatness) rise with depth nearly as strongly as the safety ranking, so at band resolution any quantity aligned with band structure rises mechanically. *Second*, the rise is not established statistically: a phase-randomized surrogate that preserves the layer autocorrelation gives $p{=}0.183$ for the late-decoder slope (Section 3.4 explains why this null is conservative for smooth trends), so we read it as inconclusive. In the energy-adjusted mixed model the association is positive at every layer but not monotone in depth: the top-layer coefficient exceeds the projector coefficient (contrast $+0.29$, CI $[0.11, 0.51]$), while the late-decoder mean ($+0.52$, CI $[0.31, 0.76]$) is not above the projector value. *Third*, single-layer activation patching restores the jailbreak only near the decoder entrance (Section 5.4): because downstream positions have already read the occluded audio at lower layers, this design cannot detect a representation that forms deeper, and shows only that full-span substitution suffices at the entrance.

After energy and prompt adjustment the divergence still predicts band necessity. Because the band-level analysis is ill-conditioned (both sides of ρ_ℓ track band energy), we move to the $520{\times}8$ prompt-by-band level, restrict to the 399 prompts that the full attack leaves jailbroken, and fit the mixed-effects model of Section 3.4, with all divergences recomputed from the generation-regime forward. The divergence coefficient is *positive at every layer* ($+0.63$ at $\ell{=}0$, $+0.93$ at $\ell{=}30$; bootstrap 95% CI $[0.70, 1.19]$), so once band energy and prompt identity are conditioned out, a band's representational divergence positively predicts its necessity within a prompt. This is an energy-adjusted correlate, not a localization claim: the depth trend itself is not established (no behavior-specific rise, surrogate null inconclusive, patching a null), so we report alignment without localization. Consistent with this, the raw depth increment is a weak, resolution-dependent effect (Appendix A.9).

5.4 Activation Patching Is a Null for Localization

To probe whether the depth correlation is causal, we inject the adversarial audio-span representation back into a band-occluded stream at a single swept layer, in a raw-overwrite and a norm-matched mode Zhang & Nanda (2024). A hard self-patch gate preceding every cross-patch is a true identity at every swept layer ($312/312$) once the injection reference is re-collected from the generation regime. Under the pre-specified criterion (restoration rising with layer, high late and low early, control flat), the data are a null for localization: restoration of the necessary band is 0 at layers 18–30, 0.08 at 16, 0.33 at 12, 0.67–0.75 at 8, and 0.83 at the projector output, so restoration is *monotone-decreasing* with depth and behaves similarly on the control band, so the near-projector restoration is not specific to the necessary band. Because the injection is the full adversarial audio-span representation, this tests whether the jailbreak survives *full-span* substitution at each depth; it cannot separate one band from another, so it is not a band-localization null. Patching the full-adversarial span into the *clean* stream (the ceiling reference) reconstitutes the jailbreak only at the decoder entrance and nowhere deeper; the early-strong/late-null profile is the inverse of the correlational ρ_ℓ rise. At $n{=}12$ samples per layer, a true upper-decoder restoration below ≈ 0.2 cannot be excluded (95% Clopper–Pearson upper bound 0.26), while one of 0.2 or more would be seen with ≥ 0.9 power.

5.5 Additional Findings: Geometry, Transfer, Detector, and Robustness

Geometry, robustness, and replication. We claim no residual-stream advantage for detection: a supervised log-mel input-space classifier matches the residual-stream probe on the same negatives (AUROC 0.999 vs. 1.00; Appendix A.3.5). Robustness cross-checks hold in the appendix: gradient attribution reproduces the occlusion ranking (Spearman $+0.97$, $p{<}0.001$), as does raw delta-energy share alone at $n{=}520$ (Spearman $+0.95$); a $b_1{=}[0, 9]$ band-stop notch reduces ASR by about two

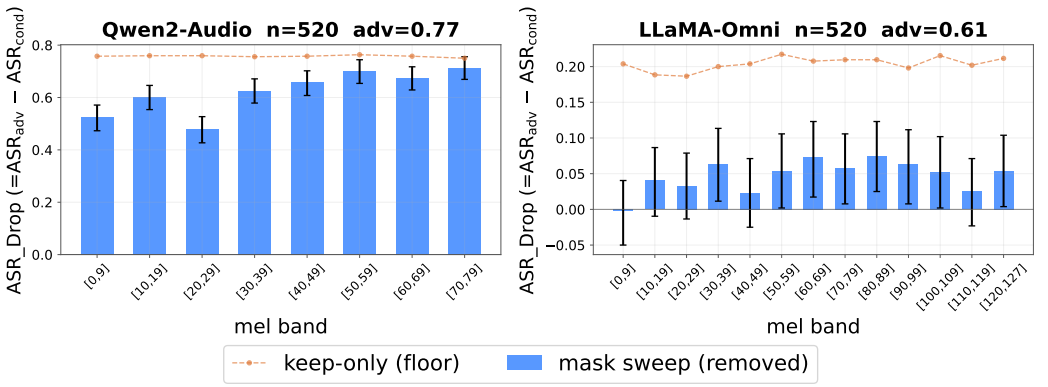


Figure 2: Per-band behavioral importance at the full-corpus scale (n=520, mask sweep in blue with prompt-level 95% bootstrap CIs; keep-only drop in orange). On Qwen2-Audio (left, ASR_{adv}=0.77) masking any upper band breaks the attack, but a random contiguous region of equal bandwidth has the same effect (Section 5.1), so the upper-band appearance is not a privileged band: the ordering inside the block is not resolved at this 10-bin resolution, and an equal-energy partition makes every band equally load-bearing. No single band is sufficient: retaining any one band collapses ASR to the floor. On LLaMA-Omni (right, ASR_{adv}=0.61) the mask drops are all small and each CI overlaps zero.

thirds at n=520 (0.77→0.24), with decoding/optimization-seed stability and clean-speech scope in the appendix (Appendix A.7, A.8).

Cross-corpus replication at full scale. The frequency and removal-structure findings replicate on three further malicious-prompt corpora at full scale, with no necessity filter (Table 8, Appendix A.11): JailbreakBench (n=70), MaliciousInstruct (n=93 of 100), and StrongREJECT (n=190 of 276) all place the highest mel band 70–79 as the single most-necessary band (drop +0.81 to +0.96 of adversarial ASR when masked) and 20–29 as the least-damaging; random contiguous masking of equal bandwidth matches in every corpus, and no single band subset suffices. The depth rise is not significant in any of the three corpora under the same phase-randomized null (p=0.195, 0.230, 0.499; cf. primary p=0.183), a qualitative confirmation of the frequency results rather than a new headline number.

LLaMA-Omni null arm and cross-model transfer. On the n=520 corpus, Omni's $\text{ASR}_{\text{clean}}$=0.41 and ASR_{adv}=0.61, a 0.20 white-box margin versus 0.76 on Qwen, so per-band effects and the depth alignment are null (Appendix A.3.4); direct cross-model transfer is a two-directional null with no transferable band or depth signature (Appendix A.3.6).

6 Conclusion

We asked where an audio jailbreak lives, in frequency and in decoder depth, and answered it for AdvWave-P on Qwen2-Audio with an occlusion protocol and energy- and bandwidth-matched controls. The answer is negative in both directions. The upper-band locus of a standard mel partition disappears once bandwidth and energy are matched, a band's energy share alone reproduces its measured importance, and the band selections of two prior methods do no better than matched random ones. The attack depends on the structure of the removal: it survives a uniform loss of energy but not a contiguous spectral gap, and no band subset carries it. In depth, behavior and representation are aligned from the decoder entrance onward, but not in a way that localizes the jailbreak to a layer. The practical lesson is methodological: band-level importance should be reported against bandwidth-matched random bands, equal-energy partitions, and an energy-share baseline; these controls cost little and would have changed the conclusion here. Whether the same holds for other attacks and victims, in particular attacks optimized in the mel domain, is the main open question.

7 Limitations

Scope. All positive results come from one attack (AdvWave-P, $\epsilon = 0.01$) on one victim (Qwen2-Audio). LLaMA-Omni judges 40% of clean harmful prompts as jailbroken, leaving too little margin to measure per-band effects, so it provides no test of generality. The protocol applies to additive, spectrally co-located perturbations, not unchanged to suffix-appended, semantic or mel-domain attacks.

Budget. Attacks optimized at $\epsilon = 0.005$ and 0.02 did not survive cold-start re-inference (0.84 and 0.76 at training time fall to 0.32 and 0.40 on re-inference at n=50, for reasons we have not isolated), so every behavioral result is at $\epsilon = 0.01$. The untrained-ϵ probe result therefore concerns detection of the perturbation family, not of successful jailbreaks.

Occlusion target. We occlude the perturbation while holding the clean waveform fixed. This isolates the structure of the perturbation, but it is not a deployable filter, which would have to act on the mixture. The notch result in Appendix A.8 should be read accordingly.

Depth. The depth analysis is correlational. The band-level correlation is computed over K=8 points whose two sides both track band energy; the prompt-level model addresses this but is not causal. Our single-layer activation-patching design cannot detect representations that form above the decoder entrance; cumulative or KV-cache patching is left to future work.

Judge. Labels come from the AdvWave judge (GPT-4o-mini), validated against one author-annotator on a condition-blind set; a second annotator was not available. A stricter judge lowers the adversarial ASR (0.76 to 0.64 at n=50) but preserves the ordering of every band-level condition (agreement ≥ 0.92 on masked conditions).

Replication. The replication corpora use a different TTS renderer, and each is run on its full jailbroken pool (n=70, n=93, n=190 for JailbreakBench, MaliciousInstruct, and StrongREJECT, no necessity filter), so the cross-corpus claims replicate the qualitative frequency and partition-dependence findings qualitatively; they do not clear the depth significance null in any corpus. The re-optimized sufficiency result is a pilot (n=20).

USE OF AI AND AI-ASSISTED TOOLS IN THE WRITING PROCESS

In accordance with the ICLR policy, we disclose the use of generative AI tools in this work and take full responsibility for its content. This section does not count toward the page limit.

Generative AI used for required-disclosure tasks.

- **Automated judge.** The LLM-as-a-judge used to adjudicate every ASR label is itself a generative model (a `gpt-4o-mini` backend), disclosed in Section 4 and validated against a condition-blind human label in Section 3.4.
- **Clean-speech rendering.** The clean audio stimuli were rendered with a text-to-speech model (a closed API in the primary run and an open-weights edge TTS for the replication corpora), disclosed in Appendix A.4; this is a documented construction step of the pipeline, not a source of the reported measurements.
- **Code, methodology, and analysis assistance.** LLM-assisted tooling was used to debug and refine the occlusion, reconstruction, and statistical-analysis code, and to support the data analysis. All experimental results reported here (every ASR, confidence interval, ρ value, and p-value) were produced by executing that code against real model queries on the documented panel; no generative model emitted any reported number.

Generative AI used for recommended-disclosure tasks. LLMs were used to draft and edit manuscript prose, to format references, and to suggest repository structure. LLM-generated suggestions for code and statistics were reviewed, executed, and verified against the recorded experimental data before they were accepted. We reviewed all AI-assisted work; we do not rely on any AI-generated claim that was not independently checked. The reported figures were produced programmatically from the recorded outputs, and the raw harmful model outputs are not reproduced in the paper (see the Ethical Considerations statement). We take full responsibility for the final content of this work, including the text, claims, code, and artifacts produced with the aid of generative AI.

ETHICAL CONSIDERATIONS

Reused resources; no new artifacts. All attacks, prompts, and victim models are public; we introduce no new attack waveforms, harmful-behavior corpus, or victim weights. The coarse band ranking and the white-box requirement of the depth analysis make the work more useful to defense than to attack. There are no user studies and no fine-tuned models; compute is under 100 A100-hours.

REPRODUCIBILITY

Code, scripts, the occlusion and hidden-state pipeline, and per-sample responses and judge scores will be released under a research-use license, and the submitted artifacts (Appendix A.5) include sample waveforms with their victim outputs and judge scores so the band-by-depth claims can be spot-checked. The judge, bootstrap procedure, depth-slope statistic, sampling configuration, and the

float32 no-clip reconstruction path are specified in the paper and its appendix. We do not release perturbation waveforms (they require the public AdvWave and AdvBench resources to reproduce).

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Mlyniec, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. https://arxiv.org/abs/1810.03292.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, 2018.

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, 2024.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. https://openreview.net/forum?id=89ia77nZ8u.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless speech interaction with large language models. In *International Conference on Learning Representations (ICLR)*, 2025. https://openreview.net/forum?id=PYmrUQmMEw.

Isha Gupta, David Khachaturov, and Robert Mullins. "i am bad": Interpreting stealthy, universal and robust audio jailbreaks in audio-language models, 2025.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1806.10758.

Weifei Jin, Yuxin Cao, Junjie Su, Jason Xue, Jie Hao, Ke Xu, Jin Song Dong, and Derui Wang. ALMGuard: Safety shortcuts and where to find them as guardrails for audio–language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. https://openreview.net/forum?id=pCRm6g0RnA.

Mintong Kang, Chejian Xu, and Bo Li. AdvWave: Stealthy adversarial jailbreak attack against large audio-language models. In *International Conference on Learning Representations (ICLR)*, 2025. https://openreview.net/forum?id=0BujOfTqab.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. https://arxiv.org/abs/2202.05262.

Raghuveer Peri, Sai Muralidhar Jayanthi, Srikanth Ronanki, Anshu Bhatia, Karel Mundnich, Saket Dingliwal, Nilaksh Das, Zejiang Hou, Goeric Huybrechts, Srikanth Vishnubhotla, Daniel Garcia-Romero, Sundararajan Srinivasan, Kyu Han, and Katrin Kirchhoff. SpeechGuard: Exploring the adversarial robustness of multi-modal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*,

pp. 10018–10035, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.596. URL https://aclanthology.org/2024.findings-acl.596/.

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5231–5240, 2019.

Lea Sch"onherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. In *Network and Distributed System Security Symposium (NDSS)*, 2019.

Thomas Schreiber and Andreas Schmitz. Surrogate time series. *Physica D: Nonlinear Phenomena*, 142(3–4):346–382, 2000.

Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Tang, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, Zhenhao Chen, and Xiuying Chen. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models, 2025.

James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J. Doyne Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1–4):77–94, 1992.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. https://openreview.net/forum?id=NpsVSN6o4ul.

Yunqiang Wang, Hengyuan Na, Di Wu, Miao Hu, and Guocong Quan. Grm: Utility-aware jailbreak attacks on audio llms via gradient-ratio masking, 2026.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. Audio is the achilles' heel: Red teaming audio large multimodal models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9292–9306, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.470. URL https://aclanthology.org/2025.naacl-long.470/.

Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song. Characterizing audio adversarial examples using temporal dependency. In *International Conference on Learning Representations (ICLR)*, 2019.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. https://arxiv.org/abs/2309.16042.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

A Appendix

A.1 Methodology and Protocol

We operate under the occlude–score–correlate protocol of Figure 1: decompose the perturbation $\boldsymbol{\delta}$ in the STFT domain, zero (mask) or retain (keep) each band in turn, re-synthesize, play each condition to the victim ALM, score with the judge, and correlate per-band behavioral importance across decoder layers with the layer-wise audio-span divergence $\Delta^{(\ell)}_{\mathrm{llm}}$.

A.2 Representation-Space Visualizations

We supplement §5.5 with linear CKA heatmaps across all $K{+}2$ conditions at layer 30 and two-dimensional PCA and t-SNE projections of the audio-span-pooled hidden states at layer 30.

A.3 Supplemental Analyses and Diagnostics

Activation-patching gate and scale detail. A hard self-patch gate precedes every cross-patch: patching a layer with its own hidden must reproduce the unpatched generation exactly. The sweeps we first reported collected the injection reference in a separate cache-free forward, which is not numerically equal to the hidden the cache-enabled generating forward produces (relative L2 over the audio span 0.08–2.26, i.e. not floating-point noise). Under that reference the exact-text self-patch identity held only near the top of the stack ($12/12$ at layers 30–31, failing at essentially all layers below, in both injection modes); this is the scale-mismatch signature that, unreconciled, produced an internal contradiction between the gate and the per-layer stability report. Re-collecting the reference from the generation regime per (Zhang & Nanda (2024)) removes the mismatch, and the self-patch identity then holds at *every* swept layer ($312/312$: 12 samples $\times$ 13 layers $\times$ 2 modes). The remaining norm divergence between the injected adversarial span and the incumbent occluded- or clean-stream span (mean $|\mathrm{ratio}{-}1|{=}0.10/0.20/0.37/0.71$ for $\ell{=}31/30/28/20$, up to $\sim 2.6\times$ at $\ell{=}20$) is the RMSNorm-scale property the norm-matched mode is designed to absorb; the corrected runs report both modes, which agree (Section 5.4).

A.3.1 Per-Band Behavioral Detail

The depth-alignment curve across all decoder layers is shown in Figure 7, which contrasts the pooled-audio-span estimator with the last-audio-frame estimator; the pooled-span representation carries the depth trend, while the last-frame curve is flat. The full-depth curve used for the slope analysis in §5.3 is reproduced in Figure 8, and the per-layer band-by-layer heatmap in Figure 9.

A.3.2 Judge-Misclassification Correction (Rogan-Gladen)

To correct judge-error we validate the primary judge on the condition-blind set of Section 3.4. On the 120 blind adversarial responses (the adversarial cell, labeled by the author; single-annotator, Section 7) the judge marks $P_j{=}0.733$ jailbroken ($88/120$), the human-positive rate is 0.825 ($99/120$), and the judge has conditional false-positive rate $FPR{=}0.095$ ($2/21$ human negatives) and false-negative rate $FNR{=}0.131$ ($13/99$ human positives). Corrected for misclassification, $\mathrm{ASR}^{*}_{\mathrm{adv}} = (\mathrm{ASR}_{\mathrm{adv}} - FPR)/(1 - FPR - FNR)$: at the $n{=}520$ headline $\mathrm{ASR}_{\mathrm{adv}}{=}0.767$ this gives $0.868 \approx 0.87$, and the correction leaves every qualitative contrast unchanged (Section 7).

Table 4 reports the $K{=}8$ contiguous per-band jailbreak ASR under the float32 no-clip reconstruction, the single reconstruction path used for every behavioral number (STFT/ISTFT with $n_{\mathrm{fft}}{=}400$ and hop 160 at 16 kHz on the perturbation delta, exact inversion, float32, no int16 clip). The earlier clipped-16-bit masks are not reported as a second column because that earlier set was produced against a phase2_gen_v2 audio set whose string-sorted stem names left 86% of condition wavs absent at inference time; its collapsed ASR was a missing-wav artifact, not a measurement of the 16-bit path. The 16-bit-versus-float32 question is instead settled directly on the waveform (relative RMSE $1.76{\times}10^{-5}$, signal-to-quantization noise ≈ 70.5 dB), so the reconstruction choice does not rescale which bands drive the jailbreak; the float32 numbers are the reference reconstruction.

A.3.3 Energy Diagnostics

Table 5 reports each band's share of total $\boldsymbol{\delta}$ energy, its Hz edges, and the round-trip occlusion residual (the fraction of in-band energy that re-enters the supposedly removed band), the data underlying the energy-confounded reading in §5.2. The per-band ASR-drop versus per-band delta-energy scatter is shown in Figure 11; the residual scatter below the energy trend indicates a frequency-specific component beyond the energy confound, but the reading is correlational and we do not lean on it as independent evidence.

A.3.4 LLaMA-Omni Behavior

The Omni ASR and depth-alignment sweeps are reported in Figure 12, and the cross-architecture summary (mask/keep band drops and the depth-slope row) in Table 6.

A.3.5 Detector Depth Curves

Figure 13 reports the linear-probe and Mahalanobis detector AUROC by decoder depth on both ALMs. The input-versus-residual-stream comparison (adversarial-versus-benign-perturbed, plus the untrained-epsilon column) is in Table 7, discussed in §5.5.

A.3.6 Cross-Model Transfer

Figure 14 reports the 2×2 ASR matrix and the per-band sweep across each cross-victim. The transfer result is a two-directional null, summarized in §5.5.

A.4 Clean-Speech Renderer

The clean AdvBench waveforms are rendered by OpenAI `tts-1-hd` ("tts-1-hd"), voice `alloy`, speed 1.0, at the engine's native 24 kHz, stored as an MP3 container under a .wav name, and resampled to 16 kHz mono when fed to either victim (the analysis and generation protocol is otherwise agnostic to the clean renderer). Both Qwen2-Audio and LLaMA-Omni consume the identical clean files (byte-identical), which is the basis of the clean-condition comparison in §5.5. If the corpus were re-rendered under a different voice or engine this renderer specification would be the record of the original.

A.5 Submitted Artifacts and Supplementary Audio Materials

Alongside this submission we upload a curated set of concrete inputs and outputs for each of the three second-corpus benchmarks (`jailbreakbench`, `malicious_instruct`, `strongreject`) so that the band-by-depth claims can be spot-checked without re-running the pipeline. For each corpus we ship the first 20 samples (indices 0–19) in the same numeric ordering used throughout the paper, so rows line up with the per-sample analysis tables. Each corpus folder contains:

- `input/`: the clean rendered prompt audio (16 kHz mono float32).
- `adversarial/`: the AdvWave-P adversarial waveform (ε=0.01, PGD).
- `noise/`: the raw perturbation $\boldsymbol{\delta} =$ `adversarial` $-$ `input`, stored at its native float32 amplitude.
- `noise_playable/`: the same noise scaled to a fixed peak of 0.5 so its structure is audible; `noise/` is the faithful version and should be used for any analysis.
- `outputs/`: the victim model's verbatim responses to the clean and adversarial clips, tagged `[refused]` or `[jailbroken]`.
- `manifest.csv`: one row per sample (0–19) with the prompt text and the clean/adversarial judge labels.

The clean audio was rendered with `edge-tts` (voice `en-US-AriaNeural`), `Qwen2-Audio-7B-Instruct` (bf16) is the victim, and the `[refused]`/`[jailbroken]` labels were produced by a GPT-4o-mini safety judge at temperature 0. Across the shipped samples the clean clips are overwhelmingly refused while the adversarial clips are mostly jailbroken: `jailbreakbench` 20/20 refused clean and 18/20 jailbroken adversarial, `malicious_instruct` 17/20 and 18/20, `strongreject` 19/20 and 15/20. The raw perturbation (mean RMS $\approx$ 0.0065) is perceptually inaudible yet reverses the model's refusal on most samples, and the same native sample format (16 kHz mono float32) is used everywhere. The corresponding per-condition judge scores and code live in the submission repository (Section 7).

A.6 Gradient Band-Attribution Baseline

Because occlusion costs $K+2$ forward passes per sample, a cheaper gradient-based band attribution is a useful cross-check. On the saved ε=0.01 adversarial waveforms we backpropagate the attack objective (the same next-token cross-entropy used by AdvWave-P, evaluated on the contaminated waveform) to $\boldsymbol{\delta}$ and aggregate the squared gradient over each analysis-basis band in the STFT domain (n=50, no re-optimization, no generation). The gradient band ranking is strongly correlated with the occlusion ranking (Spearman $\rho = +0.97$, $p<0.001$; Pearson $+0.84$), so the occlusion protocol's primary value is validation of a cheaper gradient attribution rather than an independent source of the ranking. Raw energy share, which needs no gradient, already recovers the ranking at Spearman 0.95 at n=520; the occlusion protocol's distinct value lies in the controls (random contiguous, uniform rescale, scattered) that neither baseline provides. Recomputing the depth alignment with the gradient band saliency as the behavioral target in place of the occlusion ASR drop reproduces

the trend: ρ_ℓ rises with depth from $\rho_0{=}0.70$ to $\rho_{30}{=}0.94$ (late-decoder slope $+0.0102$ per layer; a layer-index permutation test is significant as an ordering check only, consistent with the primary trend and not as slope significance under an autocorrelation-preserving null), so the depth rise is not a property of the occlusion audio specifically; it tracks band geometry, which the gradient and occlusion rankings agree on. The occlusion protocol therefore retains its behavioral-ordering value, and its claim of uniqueness is only that it obtains the ranking without a differentiable attack objective.

A.7 Decoding and Optimization Seed Stability

Decoding seeds. The main sweep uses greedy decoding, giving one deterministic generation per prompt per condition; because the attack family is fragile to cold-start re-inference at neighboring budgets (Section 7), we re-generate the adversarial condition and the highest– and lowest-importance occluded bands under stochastic decoding ($\mathrm{do_sample}{=}\mathrm{true}$, $t{=}1.0$, 3 independent seeds, $n{=}520$, cold start). The adversarial ASR is stable and far above the cold-start failure levels seen at $0.005/0.02$: per-seed $0.70/0.75/0.56$, mean 0.67 (spread 0.19) versus greedy 0.77. The band necessity ranking is sampling-robust: the top band $[70, 79]$ stays collapsed in every seed ($0.05/0.11/0.04$, mean 0.07), and the low band $[0, 9]$ stays far weaker in every seed ($0.22/0.31/0.18$, mean 0.24, versus adversarial 0.67), so mid-to-upper necessity holds at each seed even though the low band's absolute level carries decoding variance.

Optimization seeds. We additionally verify stability across the *optimization* seed, which varies the attack init rather than decoding. We re-optimize AdvWave-P deltas on a 20-prompt subset ($\varepsilon{=}0.01$, 3000 Adam iterations) under three independent random initializations ($\boldsymbol{\delta}$ drawn uniformly in $[-\varepsilon, \varepsilon]$). The band structure is seed-robust: every prompt's top-energy band is *identical* across all three seeds (band $b_8{=}[70, 79]$ for $20/20$ prompts), the pairwise Spearman correlation of per-band energy fractions across seeds is 0.97 on average (minimum 0.88; $19/20$ prompts have all three seed-pair correlations ≥ 0.9), and the top-band energy fraction is tightly concentrated (mean 0.21, std 0.015). All re-optimized deltas respect the bound ($\|\boldsymbol{\delta}\|_\infty{=}0.01$). This confirms the top-energy band is stable across optimization seeds; note that this reads which band holds the most delta energy under an L_∞ budget, which is partly accounted for by the band's Hz width (the top mel band covers the largest fraction of the spectrum), and it does not by itself establish frequency-specific necessity (Section 5.1).

A.8 An Arbitrary Single Notch as a Defense Hypothesis

A single band-stop notch at an arbitrary location already removes most of the attack: removing a low band ($b_1{=}[0, 9]$, about 10% of delta energy, the second-least-important band at $n{=}520$) reduces ASR by about two thirds at $n{=}520$ ($0.77{\to}0.24$; the $n{=}50$ subset gave $0.76{\to}0.34$). We report this as a property of the perturbation, not a validated defense: a defender must notch the mixture $x{+}\boldsymbol{\delta}$, not the known perturbation, and a broad low notch would suppress clean speech's fundamental-frequency band. The clean-speech utility of such a notch is not quantified, so we do not recommend it.

A.9 Resolution Ladder Across K

To check whether the raw depth increment is stable across the occlusion resolution, we rerun the same sweep on the existing $\varepsilon{=}0.01$ deltas at two further partition widths, $K{=}4$ (20-bin bands) and $K{=}20$ (4-bin bands), in addition to the $K{=}8$ and $K{=}16$ already reported. At $K{=}4$ the per-band ASR drops saturate: every 20-bin band removes enough of the concentrated upper-spectrum energy that $\mathrm{ASR_Drop}$ is 0.72–0.76 for all four bands (adversarial 0.76), so the across-band variance collapses and the raw depth increment is not a reliable statistic (point estimate -0.575, driven by near-constant ASR-drop input rather than a real sign reversal). Read at the resolutions where the bands separate, the raw increment is $+0.36$ at $K{=}8$, $+0.13$ at $K{=}16$, and $+0.094$ at $K{=}20$ (all $n{=}50$; the $n{=}520$ $K{=}8$ estimate is $+0.309$) (ASR drops rise from 0.00 on band $[20, 23]$ to 0.74 on band $[76, 79]$, so the fine partition still concentrates necessity in the upper region): it is positive at all three resolutions but decreases monotonically toward zero as K grows, so the depth alignment is not carried by any one coarse band and does not require the coarsest partition to appear. Together with $K{=}4$'s saturation, the ladder shows the depth result is a fine-resolution phenomenon: once bands separate ASRs, the increment is consistently positive, and it decays roughly monotonically from $K{=}8$ to $K{=}20$ ($+0.36 \to +0.13 \to +0.094$). At $K{=}8$ the energy-conditioned increments are degenerate at this resolution: the linear-, log- and rank-energy partial correlations all exceed the partial-correlation bound $|\rho|{\leq}1$ ($+1.40$, $+1.17$, $+0.51$ at $n{=}520$) because energy and the representational divergence are near-collinear, so we do not report an energy-conditioned increment at any n (Section 5.3); the earlier linear-'suppressor' reading was a numerical artifact of that collinearity.

Read with Section 5.3, the raw depth increment is a weak, resolution-dependent effect rather than a load-bearing depth statistic: it tracks the rising ρ_0 base (Table 2): as more, narrower bands separate finer from coarser necessity, the depth alignment moves closer to the projector-output floor, so the increment read at $K{=}16$ is the one most sensitive to the energy confound; we caution against reading it as a resolution-independent floor. We report the point estimates here; the $K{=}16$ energy-conditioned increment is the one most studied, but as Section 5.3 shows it does not survive re-parameterizing energy, so the ladder is read as a caution (a weak, resolution-dependent effect) rather than a load-bearing depth statistic.

A.10 Audit of Named Prior Band-Selectivity Claims

Two prior methods claim to localize audio-jailbreak energy to a specific band set within the frontend: GRM Wang et al. (2026) ranks mel bands by an attack-vs-utility ratio and perturbs only the top subset, and ALMGuard Jin et al. (2025) learns a sparse mask over sensitive mel bins as a defense signal. Our central finding (Section 5.1) is that the perceived band locus of the AdvWave-P perturbation is an artifact of bandwidth and energy allocation in the analysis basis once location is separated from quantity. To test whether either named prior's specific selection identifies something the bandwidth- and energy-matched random selection does not, we re-expose each prior to the protocol under a matched-random control. ALMGuard and GRM were designed for their own attacks and objectives; we test only whether their band selections localize the energy that AdvWave-P needs, not the methods themselves.

ALMGuard (band-stop, no re-optimization, $n{=}520$). We take ALMGuard's released sensitive-bin mask (a 128-mel saliency map for Qwen2-Audio), map it to our 201-bin STFT analysis basis, and apply it as a band-stop to the existing AdvWave-P deltas. The ALMGuard-restricted removal drops ASR to 0.15; twenty energy- and bandwidth-matched random masks (same bin count, same Hz coverage) give mean 0.17, empirical range $[0.16, 0.17]$. The ALMGuard mask falls inside the matched-random interval, so on the pre-specified control it is *not* significantly better than random at breaking the attack; the named prior does not localize the necessary band beyond what a bandwidth- and energy-matched random selection achieves.

GRM (band-restricted re-optimization, $n{=}20$). GRM's selection is mel-domain and falls outside our occlusion protocol's interface, so we reproduce its rule rather than its perturbations: we map GRM's key-band mask (109 of 201 STFT bins after the mel-to-STFT mapping) and re-optimize a fresh AdvWave-P delta whose STFT support is projected onto those bands only (the same machinery as the re-optimized-sufficiency experiment, Appendix A.11), then re-infer cold-start on the $n{=}20$ jailbroken JailbreakBench prompts. The GRM-restricted delta reaches ASR 0.75 $(15/20)$; twenty bandwidth-matched random masks (the same 109-bin count drawn uniformly) reach mean 0.86, empirical range $[0.75, 0.95]$, bootstrap 95% CI of the mean $[0.84, 0.885]$. The GRM-restricted attack sits at the lower edge of the matched-random range: it is *not* more load-bearing than a matched random selection and, directionally, marginally *underperforms* the random mean. On the same pre-specified criterion, the GRM selection does not localize the attack beyond matched random.

Both named priors therefore fail to identify a band set that carries the attack more than a bandwidth- and energy-matched random selection does. This is the partition-aware reading we adopt for prior band-selective claims (Section 2): the recovery of a mid-to-upper "locus" in these methods is, at least in part, an artifact of where the analysis basis and the L_∞ budget allocate energy, not evidence that the jailbreak lives specifically in those bands.

A.11 Second Corpus: Cross-Corpus Replication

To test whether the AdvBench-based conclusions are specific to that corpus rather than properties of the attack, we rerun the mask, keep-only-style sufficiency, cumulative, equal-Hz, and scattered conditions, plus the depth probe, on three independent malicious-prompt corpora: JailbreakBench (JBB-Behaviors, 89 prompts after deduplication), MaliciousInstruct (100), and StrongREJECT (276 non-AdvBench). For each corpus we render clean audio with a single fixed voice (open-weights edge TTS, because the original renderer is a retired server-versioned API; the acoustic difference is a documented risk), run AdvWave-P ($\varepsilon{=}0.01$) with per-sample targets, and measure cold-start ASR. On JailbreakBench we run the occlusion analysis on the full jailbroken pool ($n{=}70$, the $70/89$ prompts the attack leaves jailbroken, no necessity filter) so the replication is not conditioned on selection on necessity; MaliciousInstruct ($93/100$ jailbroken) and StrongREJECT ($190/276$) are now run at the same full-corpus scale, so all three corpora use their full jailbroken pool with no necessity filter. Every condition and the depth probe reuse the exact protocol of Sections 3.1–5.3,

with the same Qwen2-Audio victim, GPT-4o-mini judge at temperature 0, and the float32 no-clip reconstruction path.

Table 8 reports the $K{=}8$ mask sweep. In all three corpora the highest mel band (70–79) is the single most-necessary band (drop $+0.81$ to $+0.96$ of adversarial ASR when masked), and the lowest-importance band (20–29) is the least-damaging in all three, reproducing the AdvBench ordering. A $K{=}16$ refinement again resolves necessity into a non-contiguous set whose top element is the top sub-band 75–79 in all three corpora.

The remaining findings replicate as well. No single band of the jointly optimized delta is sufficient on its own in any corpus: the cumulative-by-importance curve stays near zero until roughly the last two bands, then jumps to $\sim$0.9–1.0. A random-k-band baseline is identical in shape, staying flat near zero through $k{=}6$ and reaching the same $\sim$0.9–1.0 only when nearly all bands are kept (it differs from the importance-ordered curve by only a small constant at intermediate k), so sufficiency is a global property of the band set rather than of any particular importance-ordered selection. We also run the constrained-band re-optimized sufficiency experiment, which asks whether the attack objective itself can be carried by a band subset: on $n{=}20$ jailbroken JailbreakBench prompts we re-optimize a fresh AdvWave-P delta whose STFT support is projected onto a single band (or band union) each step, and re-infer cold-start. The full-delta attack reaches ASR 0.95 $(19/20)$, while the band-only re-optimized delta never restores it: the least-important band $[20, 29]$ yields ASR 0.00, the most load-bearing band $[70, 79]$ yields 0.25, and the two-band union yields 0.40. No band subset carries the attack objective on its own, so re-optimized sufficiency fails as well: sufficiency is a property of the full band set, not of any single band or pair, even when the attack is re-optimized with the band as the only support. An equal-Hz partition drives every band to a low residual ASR in all three corpora, reproducing the partition-dependence claim. The matched-energy scattered control shows the same two-part structure as AdvBench: scattering the removed bins in the low bands (0–39) leaves ASR-drop near zero while the adjacent contiguous mask damages the attack substantially (a contiguity effect beyond energy, as in Section 5.2), whereas in the high bands both contiguous and scattered removal push ASR to the floor (energy dominated). Finally, on depth, the per-layer correlation between a band's behavioral importance and its representational divergence is positive throughout and rises toward the decoder in all three corpora ($\rho_0 = +0.69, +0.87, +0.72$ to $\rho_{30} = +0.88, +0.98, +0.93$); a phase-randomized surrogate null on the late-decoder slope, the same autocorrelation-preserving test as the primary run, is not significant in any of the three corpora ($p{=}0.195$, 0.230, 0.499; cf. primary $p{=}0.183$), so the cross-corpus study replicates the frequency findings, while the depth rise is not significant in any corpus.

Qualifications on the second-corpus study. The second-corpus results carry three qualifications that the primary study does not. First, the renderer differs (edge TTS vs. the original closed API), so absolute ASR values are not directly comparable across corpora; we report per-corpus drops against each corpus' own baseline for this reason. Second, all three second-corpus columns use each corpus's full jailbroken pool (no necessity filter, $n{=}70$, $n{=}93$, $n{=}190$), so the replication is not conditioned on selection on necessity, but each pool is smaller than the $n{=}520$ primary run and the pools differ in size across corpora, so precision varies across columns. Third, the depth probe here is the pooled-L2, Pearson variant read against the same autocorrelation-preserving surrogate null as the primary run; the monotone rise is clear in all three corpora but does not clear that null ($p{=}0.195$, 0.230, 0.499), matching the primary $p{=}0.183$, and the StrongREJECT rise saturates early (reaching a high value by the mid layers rather than continuing to climb). We therefore read the second corpus as replication of the qualitative findings, not as new headline numbers.

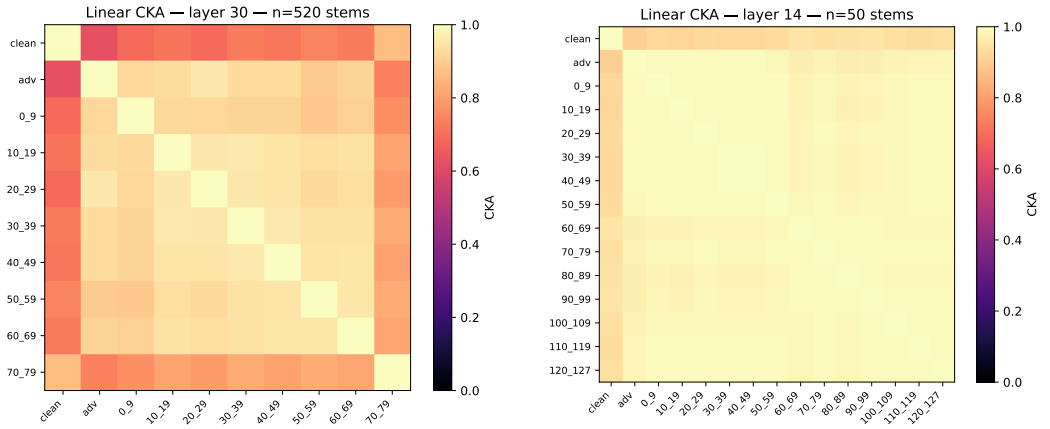


Figure 3: Linear CKA (Qwen2-Audio at ℓ=30, n=520; LLaMA-Omni at ℓ=14, n=50 pairing, which lacks the n=520 sampling point at ℓ=14). **Left:** Qwen2-Audio (10×10). **Right:** LLaMA-Omni (15×15). Cells are linear CKA between audio-span-pooled residual states (Kornblith et al., 2019). Band-removed conditions form a broad high-similarity block, so the behaviorally relevant differences in Figure 9 live in the high-dimensional distance rather than in CKA-detectable geometry.

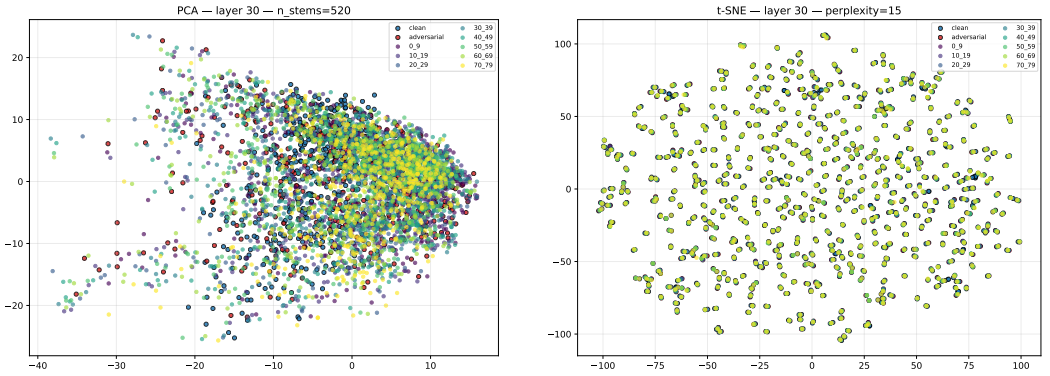


Figure 4: Qwen2-Audio at layer 30, n=520. **Left:** PCA. **Right:** t-SNE (perplexity 15). Each point is one (sample, condition) pair across clean, full adversarial, and K=8 band-removed variants. Conditions overlap in both projections, confirming the jailbreak direction is not on the first two unsupervised axes.

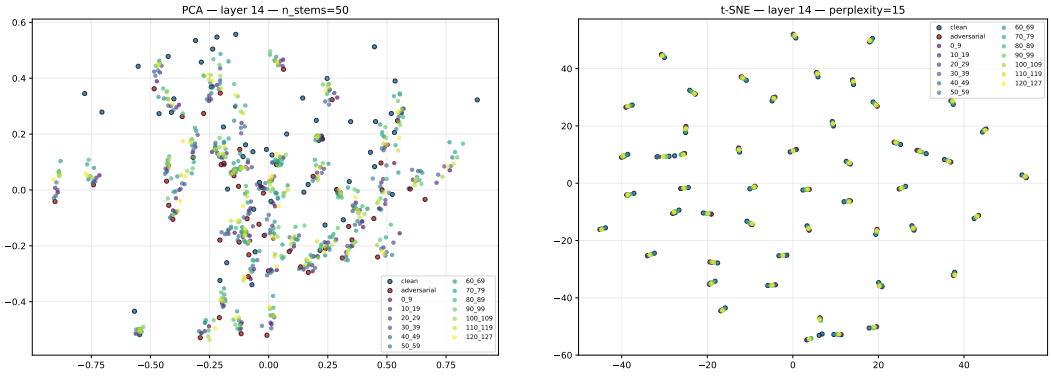


Figure 5: LLaMA-Omni at layer 14, n=50. **Left:** PCA. **Right:** t-SNE (perplexity 15). As on Qwen2-Audio (Figure 4), heavy condition overlap confirms a structured high-dimensional displacement rather than a 2D cluster.

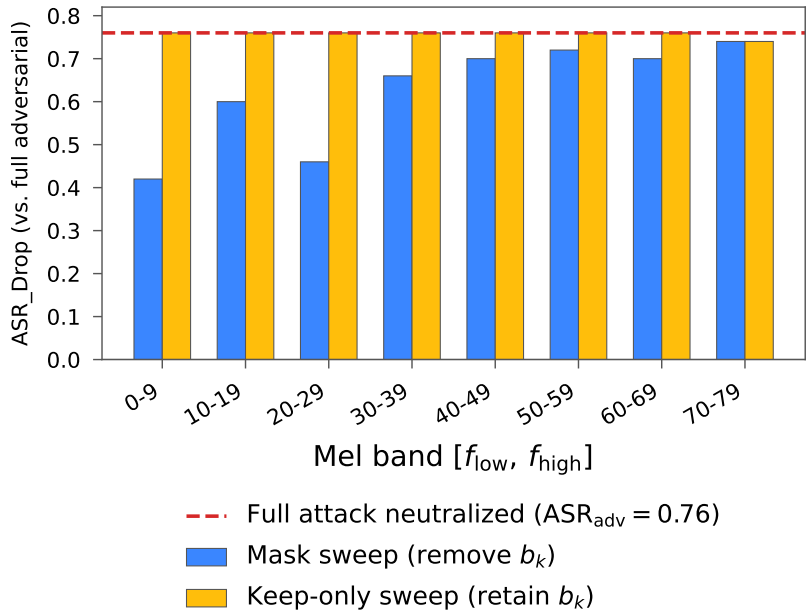


Figure 6: n=50 version of the per-band behavioral importance (mask sweep in blue, keep-only in orange) on Qwen2-Audio. Its mask ranking is recovered at n=520 with Spearman $\rho = +0.97$ against the full-corpus ranking (§5.1); it is shown here for continuity with the n=50 analysis in prior sections.

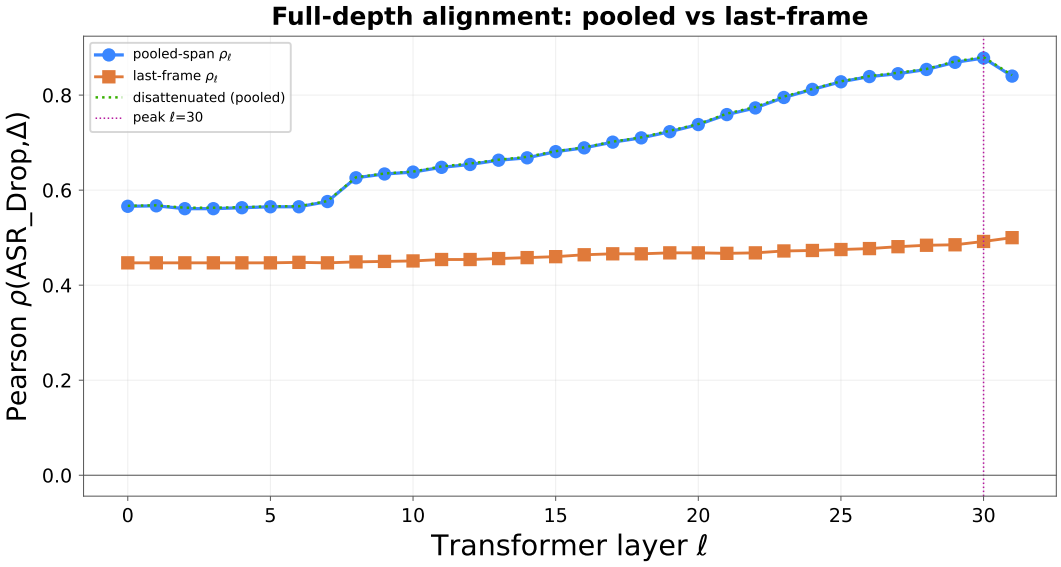


Figure 7: Full-decoder-depth alignment from per-prompt hidden states (n=50 recollection subset; the n=520 full-depth trend appears in Figure 8). Pooled audio-span ρ_ℓ rises with depth while the last-audio-frame ρ_ℓ stays flat and lower ($\rho^{\text{last}} \approx 0.45$–$0.49$ at all layers), so the depth alignment is specific to the pooled-span representation, not the final audio-frame token.

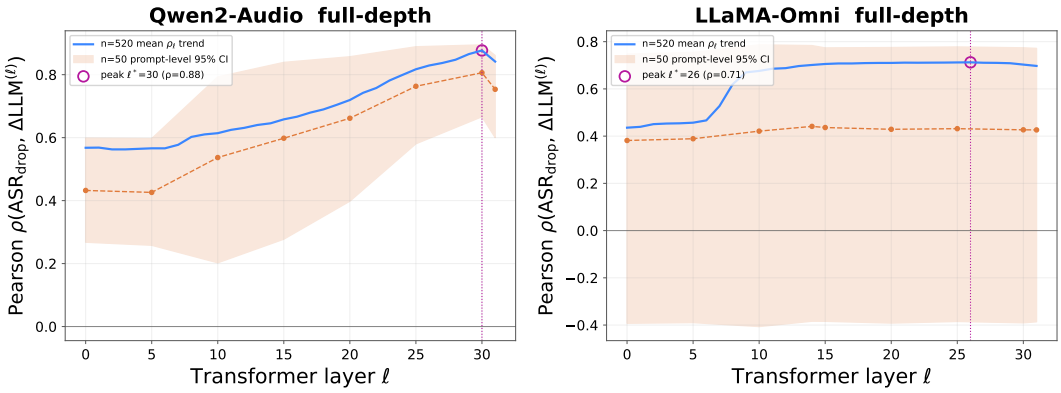


Figure 8: Full-decoder-depth alignment curve. The solid blue line is the per-band Pearson correlation ρ_ℓ=corr(ASR_Drop, $\Delta^{(\ell)}_{\text{llm}}$) computed over the n=520-mean divergences at all 32 layers; the orange band is the prompt-level 95% bootstrap CI from the n=520 subset at the sampled layers $\{0, 5, 10, 15, 20, 25, 30, 31\}$. On Qwen2-Audio the 520-mean trend rises from $\rho_0 \approx 0.568$ at the projector output to ρ_{30}=0.877; on LLaMA-Omni the same 520-mean estimator is flat ($\rho \approx 0.71$ at its upper extent, indistinguishable from zero over layers). The n=520 curve is a 520-mean point estimate, not a bootstrapped band: it reports the full-depth trend, and the n=520 CI band remains the prompt-level uncertainty statement for the sampled layers (§5.3). The argmax layer (ℓ=30 on Qwen, ℓ=26 on Omni) is marked for reference only and is not a reported statistic, and in particular is not used as a "critical layer". The depth statistic reported in the text is the late-decoder monotone slope of ρ_ℓ assessed against the autocorrelation-preserving surrogate null (p=0.183), not a peak layer.

Estimator	n	ρ_0	ρ_{30}	$\Delta\rho$ (CI)
Pooled, n=520 mean trend	520	0.568	0.877	+0.309
Prompt-level bootstrap	520	0.564	0.872	+0.307 [0.271, 0.342]
Disattenuated (split-half)	520	0.568	0.878	+0.310
Max-pooled span	520	0.665	0.732	+0.067 (peak 0.84)
Cosine (norm-normalized Δ)	520	0.501	0.835	+0.334
Continuous judge-score target	520	0.58	0.88	+0.30
K=16 bands, pooled	50	0.79	0.92	+0.13
Mixed-effects (γ_ℓ, energy-adj.)	520	+0.63	+0.93	late mean +0.52
Partial increment, leave-one-band-out (K=8)	520	—	—	n/a (spread > CI on LOO)

Table 2: Every estimate of the depth-alignment correlation ρ_ℓ=corr(ASR_Drop, $\Delta^{(\ell)}_{\text{llm}}$) quoted in the text, by estimator, sample size, and the increment $\Delta\rho=\rho_{30}-\rho_0$ over the projector output. All rows except the K=16 row are recomputed at n=520 from per-stem representational divergences and per-stem ASR labels. The prompt-level bootstrap (row 2) has a paired-bootstrap CI on $\Delta\rho$ of $[0.271, 0.342]$ excluding zero; the band-label permutation null on ρ_{30} is significant (p=0.001 at K=8, $p<5\times10^{-5}$ at K=16; Section 5.3). At n=520 the energy-conditioned partial estimators are degenerate: the increment under linear, log, and rank energy is +1.40, +1.17, and +0.51 respectively, all outside $[-1, 1]$ because energy and the representational divergence are near-collinear (corr $\approx$ 0.92–0.98). We therefore do not report an energy-conditioned increment at any n; the linear-'suppressor' reading was a numerical artifact of that collinearity (Section 5.3). The K=16 pooled row is the n=50 estimate: the K=16 n=520 behavioral sweep is complete (Section 5.1), but per-stem K=16 hidden states for the depth trend were not collected.

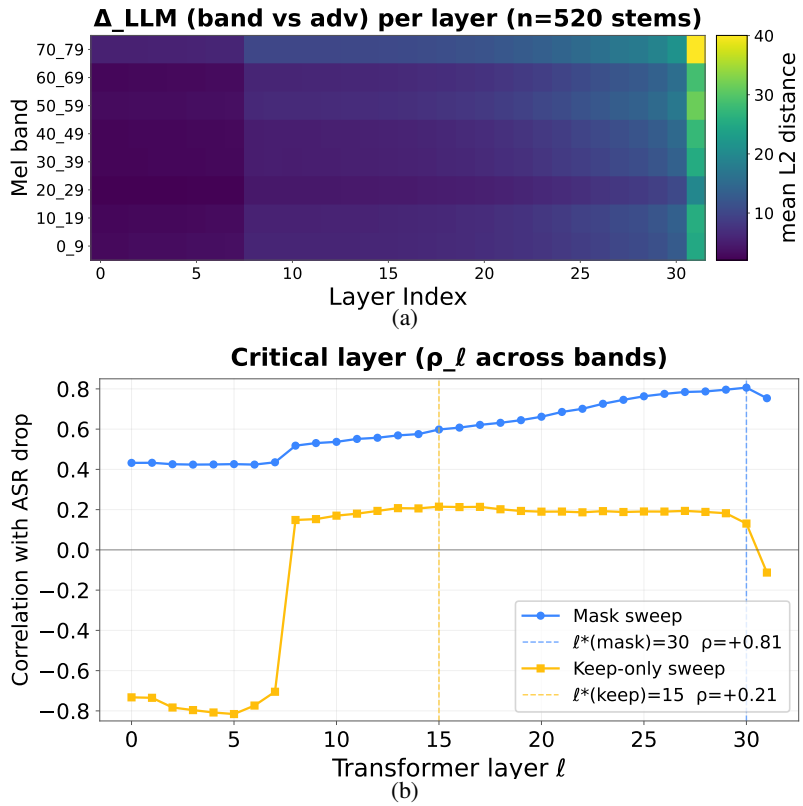


Figure 9: Frequency-depth localization on Qwen2-Audio. **(a):** $\Delta^{(\ell)}_{\text{llm}}(b_k)$ heatmap, 32 decoder layers $\times$ $K{=}8$ analysis-basis mel bands, $n{=}520$. **(b):** layer-wise ρ_ℓ between band-wise ASR_Drop and $\Delta^{(\ell)}_{\text{llm}}$; mask alignment rises with depth and saturates over the upper layers (the late-decoder monotone slope of ρ_ℓ is the depth statistic; the level at layer 30 is $\rho_{30}{=}+0.877$), with neighboring layers in the range $\ell{\in}[25, 31]$ statistically indistinguishable, so we read the depth axis as a trend rather than a single committed layer (the argmax is marked in the plot for reference only and is not a reported statistic); keep-only alignment remains weak ($\ell{=}15$, $\rho{=}+0.21$).

condition	$\text{ASR}_{\text{primary}}$	ASR_{ref}	agreement(κ)
adversarial (no mask)	0.76	0.64	0.80 (0.53)
$b_1 = [0, 9]$	0.32	0.28	0.92 (0.81)
$b_2 = [10, 19]$	0.18	0.14	0.96 (0.85)
$b_3 = [20, 29]$	0.28	0.24	0.92 (0.79)
$b_4 = [30, 39]$	0.18	0.14	0.96 (0.85)
$b_5 = [40, 49]$	0.04	0.04	1.00 (1.00)
$b_6 = [50, 59]$	0.04	0.04	1.00 (1.00)
$b_7 = [60, 69]$	0.08	0.08	1.00 (1.00)
$b_8 = [70, 79]$	0.02	0.02	1.00 (1.00)

Table 3: Adjudication agreement between the primary judge (GPT-4o-mini, the AdvWave default) and a stricter reference judge (GPT-5.6-sol) on the $K{=}8$ mask sweep and the adversarial cell ($n{=}50$ each, Qwen2-Audio, $\varepsilon{=}0.01$). Each entry is the per-sample re-adjudication of one $n{=}50$ sweep (source run `r26asr_infer`); its ASR magnitudes track that source and are not the headline float32 numbers of Table 4. The two judges agree on the lower bands but diverge on the adversarial cell: the reference judge withholds the jailbreak label more often on borderline completions, so the judges differ in threshold rather than in which conditions are affected (Section 7); every band-level ordering is preserved. The primary judge is the reference evaluation; the stricter judge is a sensitivity check.

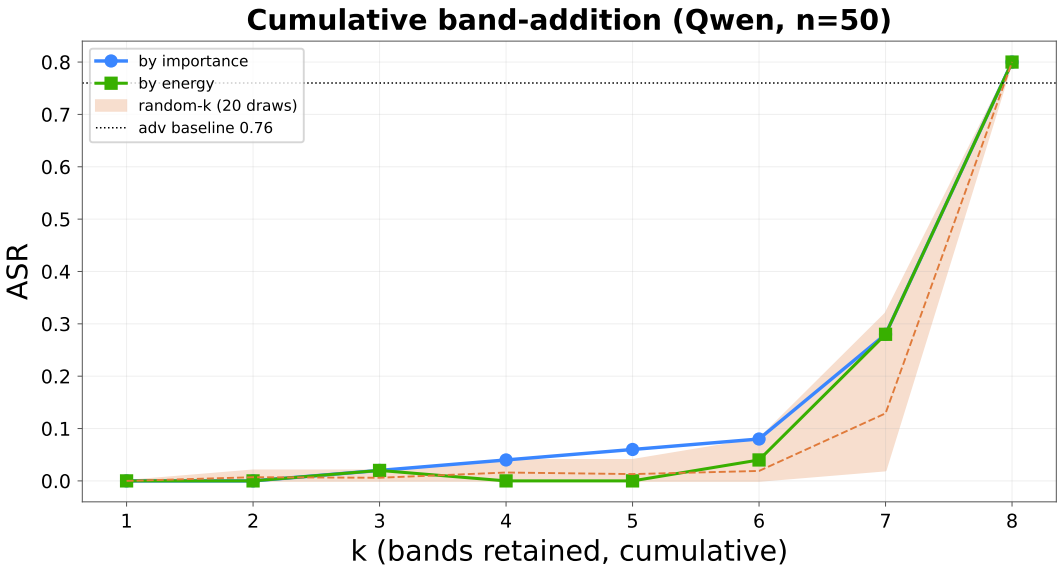


Figure 10: Cumulative band-addition on Qwen2-Audio (n=50). ASR as a function of k retained bands, ordered by mask-sweep importance and by retained energy, against an energy-matched random-k baseline (orange band, 20 draws). Retaining the least-important bands first leaves ASR at the floor through k=6 (importance path $0.00, 0.00, 0.02, 0.04, 0.06, 0.08$), rises to 0.28 at k=7, and reaches the full adversarial ASR only when all bands are retained (k=8, 0.80). The importance- and energy-ordered paths are closely matched, so *retention order is immaterial*; the energy-matched *random* baseline stays near zero until k=7–8, which is the control's purpose: it shows the sufficiency signal is carried by bandwidth+energy coverage, not by which specific bands are retained.

Mel-band (δ bins)	Float32 ASR (new path)
0–9	0.24
10–19	0.17
20–29	0.29
30–39	0.14
40–49	0.11
50–59	0.07
60–69	0.09
70–79	0.05

Table 4: K=8 contiguous mask per-band jailbreak ASR (n=520) under the float32 no-clip reconstruction, the single reconstruction path used for every behavioral number. An earlier clipped 16-bit run collapsed to a missing-wav artifact and is not reported; the 16-bit-vs-float32 question is settled directly on the waveform (relative RMSE 1.76×10^{-5}, signal-to-quantization-noise $\approx$ 70.5 dB), so the float32 numbers above are the reference reconstruction.

		Qwen2-Audio (80-bin analysis basis)			LLaMA-Omni (128-bin analysis basis)	
Band	Hz $[f_{\text{low}}, f_{\text{high}}]$	En.	Res.	Hz	En.	Res.
$[0, 9]$	0–344	0.10	0.014	0–214	0.056	0.019
$[10, 19]$	382–725	0.089	0.030	238–451	0.047	0.049
$[20, 29]$	764–1117	0.072	0.035	475–689	0.051	0.047
$[30, 39]$	1162–1656	0.101	0.022	713–926	0.046	0.047
$[40, 49]$	1722–2455	0.105	0.017	950–1184	0.046	0.046
$[50, 59]$	2554–3640	0.146	0.010	1213–1513	0.056	0.041
$[60, 69]$	3786–5396	0.147	0.007	1550–1932	0.073	0.028
$[70, 79]$	5613–8000	0.242	0.003	1980–2469	0.071	0.024
$[80, 89]$	–	–	–	2530–3154	0.078	0.018
$[90, 99]$	–	–	–	3232–4029	0.087	0.015
$[100, 109]$	–	–	–	4129–5148	0.084	0.012
$[110, 119]$	–	–	–	5275–6576	0.084	0.008
$[120, 127]$	–	–	–	6739–8000	0.221	–

Table 5: Per-band share of total perturbation ($\boldsymbol{\delta}$) energy, Hz edges per analysis-basis band (computed at $f_{\max}$=8 kHz), and the round-trip occlusion residual (fraction of original in-band energy that re-enters the supposedly removed band after STFT-mask-ISTFT-restft). The bands are analysis-basis edges; all comparisons are stated in Hz because that is the invariant under the analysis basis (Section 3.1). Residuals are below 5% everywhere and smallest on the high bands; mel-filter attenuation in neighboring bands is under 1.5% of the in-band attenuation (not shown). The round-trip residual is computed only for the full 10-bin band groups; Omni's final 8-bin band $[120, 127]$ is left "–" because the diagnostic bands are formed as whole 10-bin groups and that tail band is not part of the check.

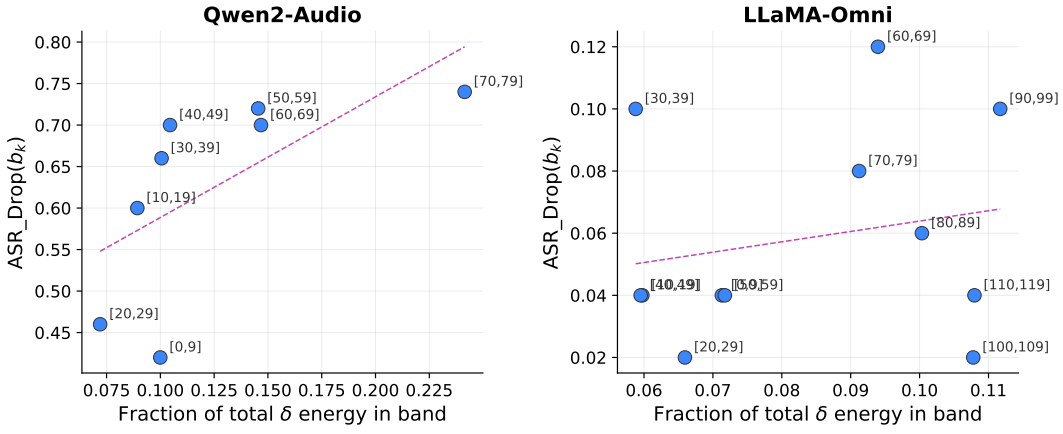
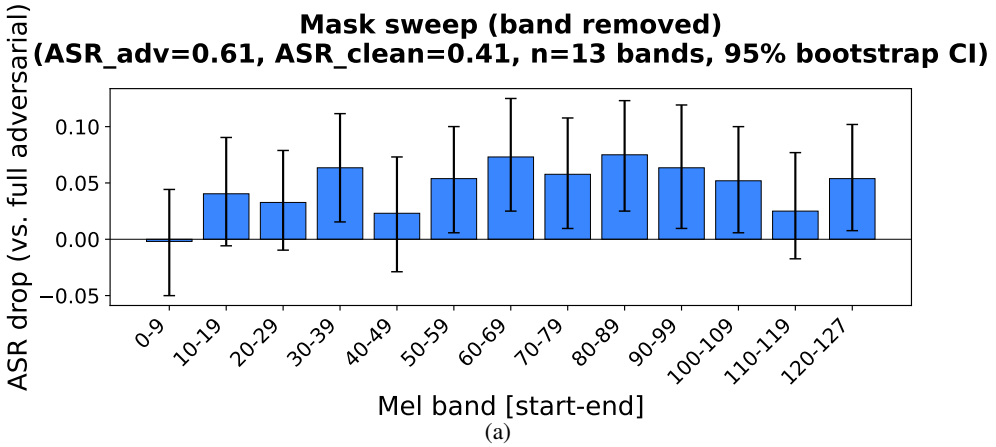
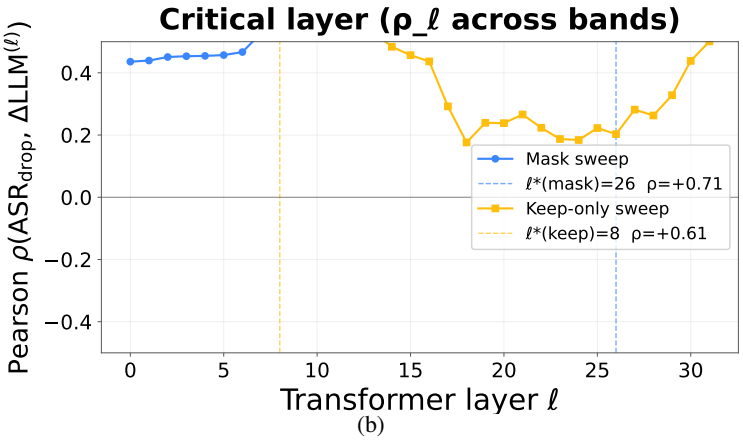


Figure 11: ASR_Drop against the fraction of total $\boldsymbol{\delta}$ energy carried by each band, band identity annotated. On Qwen2-Audio a band's energy share alone reproduces the occlusion ranking (Spearman 0.95, Section 5.1); several mid-to-upper bands sit modestly above the least-squares line, but the reading is correlational and, as Section 5.3 shows, small residual scatter is not evidence of a frequency-specific component independent of energy. On LLaMA-Omni the relationship is weak and per-band effects are within noise.



(a)



(b)

Figure 12: LLaMA-Omni null arm (128-bin analysis basis, $K{=}13$; read strictly as a null, not a power study, per Section 5.5). **(a):** mask-sweep ASR_Drop(b_k) vs. full adversarial baseline ($\text{ASR}_{\text{adv}}{=}0.61$, $n{=}520$), with 95% bootstrap CIs; per-band drops are modest and each interval crosses zero, so no single band is distinguishable from the others here. **(b):** ρ_ℓ; at its upper extent the level is $\rho{=}+0.71$ (a level, not a slope), modest and not distinguishable from a non-significant rise on this marginal-margin arm, so this architecture does not support a claimed depth localization. The mask–keep disagreement (a keep-only ρ at layer 8) is $+0.61$. The per-layer band sweep is measured on the Omni 128-bin analysis basis, so the mel-bin index denotes different frequencies than the Qwen 80-bin basis; see the Hz edges in Table 5.

Metric	Qwen2-Audio	LLaMA-Omni
Evaluated prompts n	520	520
Analysis bands K	8	13
ASR, clean	0.008	0.41
ASR, adversarial	0.77	0.61
Top mask band (basis index)	$[70, 79]$	$[80, 89]$
Top mask band (frequency)	5.6–8.0 kHz	2.7–3.4 kHz
ASR_Drop at top band	0.71	0.08
Late-decoder ρ_ℓ slope	rising	$\approx$ flat

Table 6: Cross-architecture comparison. The LLaMA-Omni arm is an n=520 corpus read strictly as a null arm (Section 5.5). ASR_Drop is measured against each victim's own adversarial ASR. The slope row summarizes the late-decoder trend of ρ_ℓ (Section 5.3); we report the slope rather than an argmax depth because the latter is not stable as a functional. The keep-only sweep is omitted: it is floor-saturated on Qwen (every retained band collapses ASR to $\leq$0.02) and returns toward the clean rate on LLaMA-Omni, so it carries no rank evidence (§5.1); the sufficiency instrument is the cumulative band-addition curve instead.

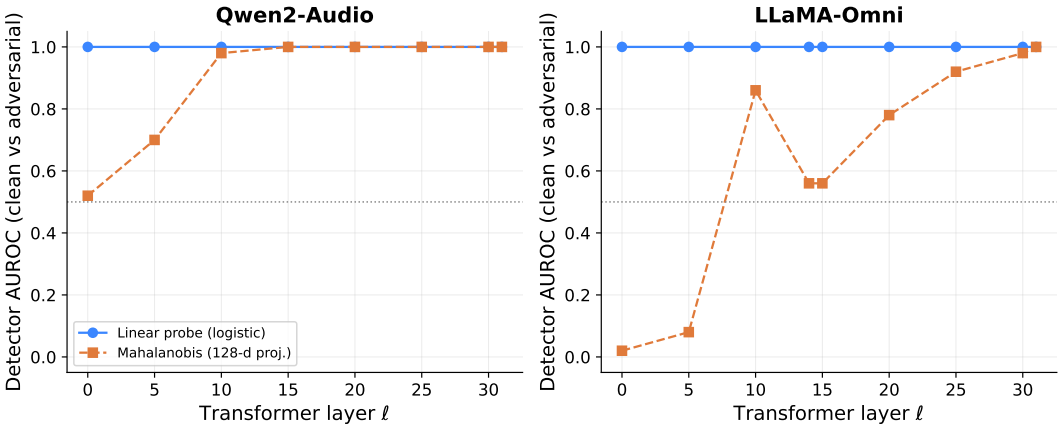


Figure 13: Clean-versus-adversarial detector AUROC as a function of decoder depth, on the n=50 pooled audio-span hidden states. A linear probe (logistic regression, held out by prompt) separates clean from adversarial representations with AUROC 1.0 at every sampled layer on both ALMs; a Mahalanobis detector on a 128-dim random projection rises from chance at the input layers to near-perfect at depth. The jailbreak displacement is therefore large enough to be linearly detectable in the high-dimensional residual stream even though the first two PCA/t-SNE axes show no cluster

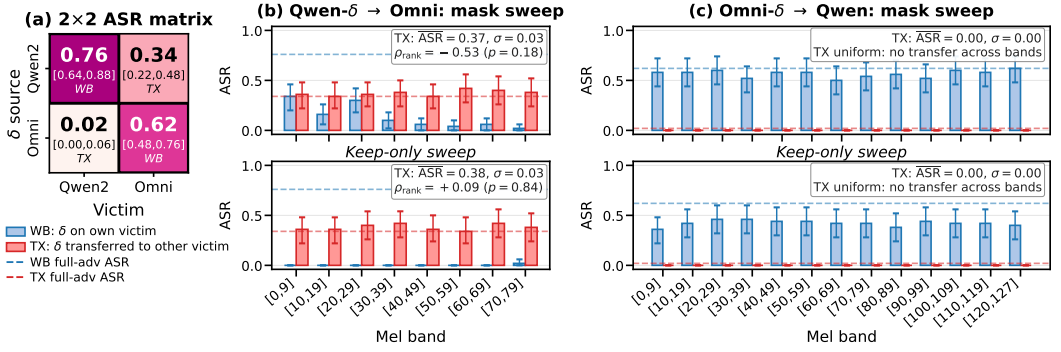


Figure 14: Cross-model AdvWave-P transferability on the n=50 matched subset. **(a)** 2×2 ASR matrix; diagonal cells (*WB*) are white-box, off-diagonal (*TX*) are direct black-box transfer. **(b, c)** per-band mask (top) and keep-only (bottom) sweeps in both directions: blue bars are the source-victim white-box ASR, red bars are the cross-victim transfer ASR; dashed lines mark the full-adversarial WB and TX ASRs. Inline boxes report transfer mean$\pm\sigma$ across bands and Spearman ρ_{rank} between the WB and TX ASR-drop rankings. Brackets in (a) are bootstrap 95% CIs over 1000 resamples.

Detector	AUROC		Untrained-ϵ AUROC
	clean vs. adv	adv vs. benign-perturbed	(0.005 / 0.02)
Input space, high-freq. energy ratio	0.92	0.64	—
Input space, spectral flatness	1.00	0.57	—
Depth linear probe (layer 30)	1.00	1.00	1.00 / 1.00
Input space, log-mel logistic (d=80)	1.00	0.999	—

Table 7: Input-space versus residual-stream detectors on the Qwen2 subset (n=50). The first column is the easy clean-versus-adversarial separation; the second is the deployed adversarial-versus-benign-perturbed separation, where "benign-perturbed" means clean speech corrupted by additive noise (SNR $\in \{15, 25\}$ dB) or a codec round-trip (MP3, Opus at 64 kbps). Fixed unsupervised input scalars nearly match the depth probe on the easy problem but drop sharply once benign corruption is introduced, while the depth probe stays at 1.0 across both settings; a supervised input-space classifier (logistic regression on the victim log-mel spectrogram, held out by prompt) reaches 1.0 and 0.999, essentially matching the depth probe, so the depth probe offers no measurable AUROC advantage over a comparable supervised input baseline in this n=50 setup. The final column reports the untrained-epsilon probe (Section 5.5): a depth linear probe fit only on the ϵ=0.01 adversarial samples, evaluated on clean-versus-adversarial for attacks optimized at the untrained bounds ϵ=0.005 and ϵ=0.02. Separation stays at AUROC 1.00 on both, so the probe detects the attack family rather than the specific optimization budget.

band	JBB (n=70)	Malicious (n=93)	StrongRE (n=190)	AdvBench (n=520)
0–9	+0.70	+0.82	+0.60	+0.68
10–19	+0.69	+0.76	+0.62	+0.78
20–29	+0.45	+0.60	+0.45	+0.63
30–39	+0.75	+0.74	+0.54	+0.81
40–49	+0.73	+0.80	+0.69	+0.86
50–59	+0.78	+0.87	+0.75	+0.91
60–69	+0.77	+0.88	+0.73	+0.88
70–79	+0.89	+0.96	+0.81	+0.93

Table 8: Mask sweep (K=8) across three corpora: adversarial-ASR drop when each band is masked. Each entry is $1 - \mathrm{ASR}(\mathrm{mask})/\mathrm{ASR}_{\mathrm{adv}}$ relative to that corpus' adversarial ASR; all three second-corpus columns use the full-corpus occlusion run on each corpus's full jailbroken pool (JailbreakBench n=70, $\mathrm{ASR}_{\mathrm{adv}}$=0.914; MaliciousInstruct n=93 of 100, $\mathrm{ASR}_{\mathrm{adv}}$=0.978; StrongREJECT n=190 of 276, $\mathrm{ASR}_{\mathrm{adv}}$=0.921), so no necessity-filtered selection conditions the replication. The AdvBench (n=520) column is normalized the same way (drop divided by $\mathrm{ASR}_{\mathrm{adv}}$=0.77) so that all four columns are directly comparable.