FG-GUARD: FINE-GRAINED PERTURBATION DEFENSE WITH FULL PATH AUDIO ATTACK ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-Language Models (ALMs) exhibit a safety inconsistency: a harmful request may be rejected when spoken neutrally yet elicit an unsafe response when the same words are delivered differently. Existing attacks establish this vulnerability empirically, but do not provide a unified account of how an audio manipulation affects the safety of the generated answer. We develop a local, first-order analysis of the complete pathway from acoustic transformation to safety-relevant output. Signal-level transformations first induce structured changes in the log-Mel representation. The ALM and output Jacobians then propagate these changes through the model to a refusal-related opening-token probability. The resulting first-order expression quantifies the signed local change in this continuous proxy. The acoustic transformation determines the time–frequency pattern of the input change, while its effect on the proxy depends on the specific model and input. Motivated by two consequences of the analysis, we propose Fine-Grained Perturbation Defense (FG-Guard). The additive log-Mel formulation motivates a perturbation-based defense in the same representation space, while the joint time–frequency structure of the derived changes motivates a two-dimensional mask over the perturbation. The mask support is estimated empirically from jailbreak and transcription gradients. On Qwen2-Audio-7B-Instruct, FG-Guard reduces AdvWave attack success from 84.61% to 41.03% and AdvBench-Audio from 1.92% to 0.00%.

1 INTRODUCTION

Audio-Language Models (ALMs) such as Qwen2-Audio integrate an audio front end with a language model that generates the response (Chu et al., 2024). This architecture makes safety behaviour dependent not only on lexical content but also on how speech is represented along the audio pathway. On a matched-text evaluation of 95 harmful prompts, Qian & Li (2026) report 4 successful jailbreaks under neutral delivery, compared with 32 under fast speech and 38 under panic delivery. Because the words remain fixed across these conditions, the contrast demonstrates that speech delivery contributes to the attack outcome.

Explaining this discrepancy requires a connection across three levels. Signal-processing studies characterize how transformations such as time-scale modification alter audio features (Laroche & Dolson, 1999; Driedger & Müller, 2016). Language-model safety studies relate internal activations to refusal behaviour (Arditi et al., 2024; Wollschläger et al., 2025). Attack evaluations measure the safety of the generated answer. These lines of work address separate parts of the pathway. A complete analysis must identify the time–frequency changes produced by an acoustic transformation, propagate them to the model state used to begin the answer, and determine their effect on continuous refusal evidence. Without this composition, empirical attack success does not explain which audio features should be modified by a defense or how such an intervention affects safety.

We formulate this complete pathway as a local analytical chain. First, pitch, tempo, and noise are characterized as structured first-order changes in the log-Mel spectrogram (§3.2). Second, the ALM Jacobian propagates these changes through the audio encoder, projector, and decoder to the state preceding the first answer token (§3.3). Third, the output Jacobian maps this state change to a continuous refusal-related opening-token proxy (§3.4). The signed effect on that proxy depends on the evaluated model and input.

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.

The analytical chain motivates two design choices in FG-Guard. The local additive expansion motivates learning a perturbation in log-Mel space. The joint time–frequency structure of pitch and tempo changes motivates a two-dimensional mask instead of a frequency-only mask broadcast over all frames (§3.5). We estimate the active cells empirically using the ratio of jailbreak to transcription gradients, and train the masked perturbation to promote refusal while preserving transcription (§3.6–3.7).

Our primary contribution is the missing analytical link between acoustic delivery and safety-relevant output. Specifically, (1) we derive first-order log-Mel changes for pitch, tempo, and noise under explicit residual assumptions and norm bounds; (2) we compose these changes with the ALM and output Jacobians to obtain a continuous refusal-related opening-token proxy whose direction must be measured for each model and input; and (3) we instantiate the resulting intervention space and time–frequency geometry in FG-Guard, which is evaluated across four attack families and two ALM backbones.

2 Related Work

2.1 Attacks

ALM attacks are either *gradient-based* (optimizing waveform perturbations through the encoder, e.g., AdvWave (Kang et al., 2025), TAGO (Fang et al., 2026), GRM (Wang et al., 2026), and AudioHijack (Chen et al., 2026)) or *transformation-based* (changing delivery via prosody, style, or signal-level perturbations (Qian & Li, 2026; Li et al., 2025; Song et al., 2025; Cheng et al., 2025)). Both groups manipulate the acoustic carrier and provide empirical evidence that model behaviour can change even when the transcript is preserved, consistent with the input-side mechanism analyzed in Section 3.2.

2.2 Defenses

Existing defenses broadly fall into three families (Feng et al., 2026; Luo et al., 2026). *Guard models* reject audio upstream but add latency and may themselves be attacked. *Latent interventions* steer decoder activations toward refusal directions (Lin et al., 2025; Yang et al., 2025) but act downstream of the audio path. *Input-space perturbations* add fixed signals to audio features. The closest predecessor, ALMGuard (Jin et al., 2025), scores Mel bins by a jailbreak-to-transcription gradient ratio and trains a frequency-only perturbation on top bins under a refusal-only loss. A frequency-only perturbation is broadcast across time, whereas pitch and tempo tangents generally vary jointly over time and frequency; this mismatch motivates FG-Guard. Which cells to perturb, whether sparse support suffices, and whether this geometry improves the safety–utility trade-off remain empirical questions.

2.3 Representation-Level Safety and Deformation Stability

Scattering transforms provide Lipschitz stability under small diffeomorphisms including time warping (Mallat, 2012; Andén & Mallat, 2014). Refusal in text LLMs is mediated by directions recoverable from residual-stream activations (Arditi et al., 2024; Wollschläger et al., 2025), and encoder–projector geometry has been linked to refusal circuitry in ALMs (Roh & Houmansadr, 2026). The connection between small input perturbations and large behavioural change follows the local-linearity picture of Goodfellow et al. (2015), quantified by margin-and-Lipschitz arguments (Hein & Andriushchenko, 2017; Tsuzuku et al., 2018). These results respectively describe deformation stability, representation geometry, or decision sensitivity; our focus is their missing composition along the ALM audio path, from a physical transformation to signed safety-relevant output evidence.

3 From Audio Transformations to FG-Guard's Design

Prior work documents audio jailbreaks and proposes defenses largely through empirical outcomes, but lacks a unified analysis linking acoustic changes to safety-relevant output. Consequently, the choice of where and how to perturb an ALM has limited theoretical guidance. We address this gap by tracing a small acoustic change through the full model path: pitch, tempo, and noise change

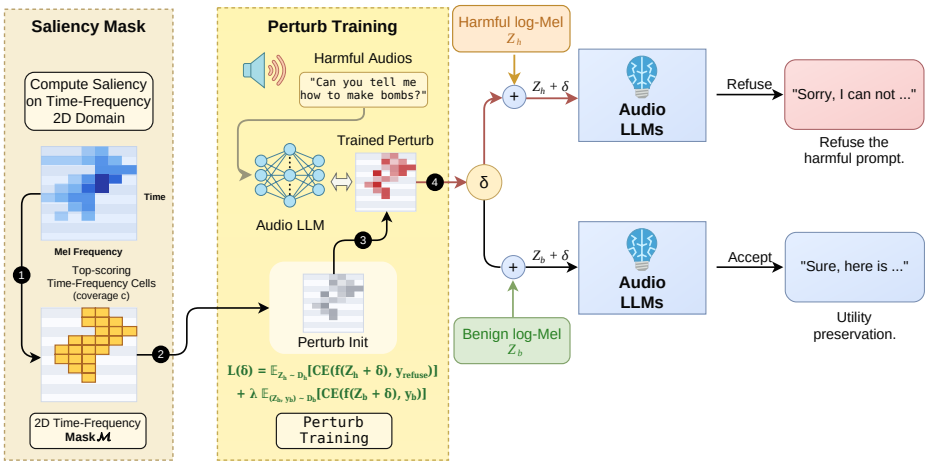


Figure 1: The FG-Guard pipeline. Step 1 computes per-cell gradients of a jailbreak loss and a transcription loss with respect to the log-Mel input. Step 2 ranks their ratio and marks the selected time–frequency cells as the sparse mask $\mathcal{M}$. Step 3 trains a perturbation δ only on those cells, using a refusal loss on harmful audio together with a transcription-preservation loss on benign audio. Step 4 adds the trained δ to each incoming log-Mel representation at inference.

the log-Mel input (Section 3.2); the change propagates to the pre-generation state (Section 3.3) and then to a continuous refusal measure at the output (Section 3.4). This chain identifies an additive intervention space and motivates variation over both time and frequency. We then explain how the analysis motivates FG-Guard's design (Section 3.5) and specify the empirically selected mask and trained perturbation that implement it (Sections 3.6–3.7).

Figure 1 previews the full method: gradient-based mask selection, perturbation training, and application to incoming audio features. Section 3.5 explains how the preceding analysis motivates these components.

3.1 Preliminaries and Threat Model

We first define the representation in which the acoustic changes will be compared, the path through which those changes enter the ALM, and the attacker and defender considered in the experiments.

Log-Mel front end. For waveform x, let $P \in \mathbb{R}_+^{K \times T}$ be its nonnegative STFT power spectrogram over K frequency bins and T time frames. A nonnegative Mel filterbank $B \in \mathbb{R}_+^{M \times K}$ maps this power to M Mel channels. With a small stabilizer $\varepsilon > 0$, the front end Φ produces

$$Z = \Phi(x) = \log(\varepsilon + BP) \in \mathbb{R}^{M \times T}, \qquad z = \mathrm{vec}(Z) \in \mathbb{R}^n, \quad n = MT, \tag{1}$$

where $\mathbb{R}_+$ denotes nonnegative entries and vec stacks the $M \times T$ cells into a vector. The appendices use $\odot$ and $\oslash$ for elementwise multiplication and division, and $\|\cdot\|_F$ and $\|\cdot\|_2$ for the Frobenius and operator norms. For Qwen2-Audio, $M = 128$ and the fixed crop-and-pad grid has $T = 3000$, so $n = 3.84 \times 10^5$ (Chu et al., 2024).

ALM model. The audio encoder $\mathcal{E}$ maps the log-Mel vector z to audio features $a = \mathcal{E}(z)$, and the projector $\mathcal{P}$ maps a to a language-aligned representation $u = \mathcal{P}(a)$. The merge operator Π combines u with the fixed text embeddings e_{text} into the full input sequence $H^{(0)} = \Pi(u, e_{\text{text}})$. Decoder block $\mathcal{B}_\ell$ updates the full hidden-state sequence as $H^{(\ell+1)} = \mathcal{B}_\ell(H^{(\ell)})$, for $\ell = 0, \ldots, L-1$, where L is the number of decoder blocks. The selection operator $\mathcal{S}$ extracts the last conditioning position before the first answer token, giving $h^{(\ell)} = \mathcal{S}H^{(\ell)} \in \mathbb{R}^d$, where d is the hidden-state width. We call this selected state the *generation anchor*; it is used to predict how the answer begins. Appendix B gives the corresponding layerwise mappings and Jacobians.

Threat model. The adversary submits audio carrying a harmful instruction, using white-box gradients (Kang et al., 2025; Wang et al., 2026; Fang et al., 2026), black-box search (Chao et al., 2023; Song et al., 2025), or acoustic transformation at fixed transcript (Qian & Li, 2026; Li et al., 2025). The defender has white-box access and knows the attack family when constructing its family-specific mask and perturbation, which are reused without per-query recomputation. Appendix H provides the detailed experimental settings and evaluation sets.

3.2 Audio Transformations Are Structured Log-Mel Perturbations

The first step asks what change the ALM actually receives when the waveform is modified. We analyze pitch, tempo, and additive noise because they occur in transformation-based attacks and represent different signal mechanisms (Song et al., 2025; Qian & Li, 2026; Li et al., 2025). Pitch moves spectral energy across frequency, tempo moves it across time, and noise interacts with the speech spectrum. Although these mechanisms differ, small changes can be expressed in a common form at the log-Mel input.

Let $\theta \in \mathbb{R}^{n_\theta}$ collect the transformation amounts, with $\theta = 0$ denoting the original audio. Under the local smoothness and implementation-residual assumptions stated in Appendices A–D, each transformation has a first-order log-Mel direction. Collecting these directions as the columns of $G(z) = [g_\rho, g_\tau, g_{\text{noise}}, \ldots] \in \mathbb{R}^{n \times n_\theta}$ gives

$$\boxed{z_\theta \;=\; z + G(z)\,\theta + O(\|\theta\|^2).} \tag{2}$$

Eq. 2 is the common conclusion of the three derivations. For tempo, g_τ depends on the frame and the local temporal variation of each Mel channel; Appendix A gives its formal expansion and norm bound on a fixed crop-and-pad grid. For pitch, g_ρ depends on the Mel channel and on how the spectrum changes over time; Appendix C derives it under a local spectral-warp model with a second-order implementation residual. Both are structured over the two-dimensional time–frequency plane. Noise is different: Appendix D shows that its first-order direction is a phase-sensitive speech–noise interference pattern, but it is still additive in the same log-Mel coordinates.

The tempo result models an ideal feature-domain warp; Appendix A specifies how implementation residuals enter for time-scale-modification methods such as a phase vocoder (Laroche & Dolson, 1999) or Waveform Similarity Overlap-Add (WSOLA) (Verhelst & Roelands, 1993).

Figure 2 compares the derived tempo direction with finite differences through Qwen2-Audio's processor. The finite differences estimate the local change produced by a small waveform tempo transformation, while the analytical profile predicts that change from the feature-domain time-warp model. Both profiles vary across the utterance and show broadly similar shifts between more and less sensitive regions. This agreement between the predicted and processor-measured temporal trends supports the first-order model's account of time-varying log-Mel changes.

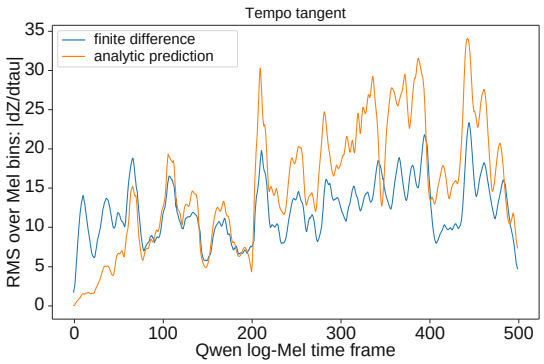


Figure 2: Tempo input-stage check on one utterance. Blue: per-frame RMS of log-Mel finite differences after small waveform tempo changes pass through Qwen2-Audio's processor. Orange: per-frame RMS of the first-order log-Mel direction predicted by the local feature-domain time-warp model. Peaks indicate frames more sensitive to the tempo change; the comparison tests whether the predicted and measured sensitive regions occur at similar times.

In summary, pitch, tempo, and noise have distinct signal-level mechanisms. Under the stated local assumptions, all three yield additive log-Mel changes as in Eq. 2; pitch and tempo also vary jointly across time and frequency. This identifies the log-Mel input as an intervention point and motivates a 2D perturbation. The next subsection propagates the input change through the ALM to determine its model-dependent effect.

3.3 LOG-MEL CHANGES REACH THE PRE-GENERATION STATE

Eq. 2 describes the change presented to the ALM. We now propagate it to the generation anchor defined in Section 3.1. Assume the encoder, projector, merge operation, and decoder blocks have locally Lipschitz Jacobians near the current input while the sequence structure remains fixed. Let $J_{\ell \leftarrow z}$ denote their composite Jacobian from the log-Mel vector z to the layer-ℓ anchor state $h^{(\ell)}$. The chain rule gives

$$\boxed{\delta h^{(\ell)} \;=\; J_{\ell \leftarrow z}\, G(z)\, \theta \;+\; O(\|\theta\|^2).} \tag{3}$$

The two factors have separate roles: $G(z)\theta$ specifies the log-Mel change caused by the acoustic transformation, and $J_{\ell \leftarrow z}$ maps that change through the ALM to the state used before generation. Appendix B gives the layer-by-layer product and a bound on the resulting local state displacement.

Eq. 3 links the acoustic change to the pre-generation state. The next step maps that state change to an output quantity whose direction can be evaluated for a given model and input.

3.4 FROM PRE-GENERATION STATE TO REFUSAL EVIDENCE

In autoregressive generation, the first few tokens condition the subsequent response and can influence whether it follows a refusal. Token-level contrastive decoding likewise shows that modifying next-token distributions can affect generation safety (Zhao et al., 2024). To measure the output-side effect of Eq. 3, let $\mathcal{R}$ be a chosen, model-specific set of tokens associated with refusal-style openings. With next-token logits $o = \mathcal{O}(h^{(L)})$ and $p = \mathrm{softmax}(o)$, define $P_{\mathcal{R}} = \sum_{j \in \mathcal{R}} p_j$ and $m_{\mathcal{R}} = \log(P_{\mathcal{R}}/(1 - P_{\mathcal{R}}))$. The log-odds $m_{\mathcal{R}}$ provide a continuous measure of the model's preference for these openings before generation.

Proposition 1 (Local safety-relevant output shift)**.** *Suppose the composed map from z to next-token logits has a locally Lipschitz Jacobian near z, the sequence structure is fixed, and $0 < P_{\mathcal{R}} < 1$. Together with the second-order input expansion in Eq. 2, for a sufficiently small audio-transformation parameter θ, the opening-token log-odds obey*

$$\boxed{\Delta m_{\mathcal{R}} = (\nabla_o m_{\mathcal{R}})^\top J_{\mathcal{O}} J_{L \leftarrow z} G(z)\theta + O(\|\theta\|^2).} \tag{4}$$

Eq. 4 completes the path. Reading it from right to left, $G(z)\theta$ describes the acoustic change at the input, $J_{\mathcal{O}} J_{L \leftarrow z}$ propagates that change to the output logits, and $\nabla_o m_{\mathcal{R}}$ measures the resulting change in opening-token log-odds. A positive first-order term increases the local probability mass of the selected token set, while a negative term decreases it. Its sign depends jointly on the transformation, model, and input; audio physics alone does not determine it. Appendix E gives the output derivative and optional hidden-state interpretation.

Thus, Eq. 4 provides a signed, model- and input-specific local prediction for refusal-related opening-token log-odds. Appendix F gives a complementary local-margin interpretation.

3.5 HOW THE ANALYSIS MOTIVATES FG-GUARD'S DESIGN

The analytical chain leads to four design decisions. Figure 1 summarizes the resulting pipeline: gradients select a 2D mask, a shared perturbation is trained within it, and that perturbation is added at inference. The analysis motivates the intervention space and its two-dimensional form; data determine the active cells and perturbation values.

Intervention space. Eq. 2 describes small audio transformations as additive changes in log-Mel coordinates, and Eq. 3 propagates those coordinates through the ALM. This motivates adding δ at the same input, $\tilde{Z} = Z + \delta$, so the defense acts directly in the representation where these changes arise.

Two-dimensional values. The derived tempo direction varies across frames; the pitch direction varies across Mel channels and, as the local spectrum changes, across frames. A frequency-only value broadcast over every frame of a band cannot generally represent either pattern. FG-Guard therefore learns a separate value at each selected (m, t) cell. Table 1 compares this design with

frequency-only ALMGuard (Jin et al., 2025); the controlled no-mask comparison in Table 2 separately tests the value of sparse support.

Sparse support. To select among the $n = 3.84 \times 10^5$ log-Mel cells, we rank them by an empirical ratio of safety-related and transcription gradients and retain the highest-ranked fraction c (Section 3.6). In the Qwen2-Audio ablation, masking lowers attack success on both reported attacks and improves AIR-Bench relative to otherwise identical unmasked training (Table 2); the masked run also reaches the stopping rule sooner.

Trained, reusable perturbation. Eq. 4 shows that the sign of a local output shift depends on the model and input. We therefore fit perturbation values across harmful and benign examples while the ALM remains frozen (Section 3.7). The resulting family-specific δ is optimized across training inputs and reused without per-query optimization.

The next two subsections specify how the mask is selected and how the masked perturbation is trained.

3.6 Empirical Mask Construction

We now implement the sparse-support choice by estimating which cells may improve safety with less effect on transcription. For a sample $(x, y_{\mathrm{jb}}, y_{\mathrm{asr}})$, where y_{jb} is the target used by the jailbreak loss and y_{asr} is the reference transcript, we measure the sensitivity of both losses to each log-Mel cell,

$$\nabla^{\mathrm{jb}}_{m,t} = \left|\partial \mathcal{L}_{\mathrm{jb}} / \partial Z_{m,t}\right|, \qquad \nabla^{\mathrm{asr}}_{m,t} = \left|\partial \mathcal{L}_{\mathrm{asr}} / \partial Z_{m,t}\right|, \tag{5}$$

and score each cell by their ratio, following Jin et al. (2025) but without collapsing the time axis:

$$S_{m,t} \;=\; \frac{\nabla^{\mathrm{jb}}_{m,t}}{\nabla^{\mathrm{asr}}_{m,t} + \varepsilon_S}. \tag{6}$$

The numerator estimates how strongly a cell affects the jailbreak-related loss; the denominator estimates its effect on transcription. Their ratio favors cells with a large safety-related gradient relative to the transcription gradient. We average both gradient maps over the mask-construction set before taking the ratio, which reduces sample-level variation. The top-scoring cells at coverage c (the retained fraction) form the binary mask $\mathcal{M} \in \{0,1\}^{M \times T}$.

3.7 Masked Perturbation Training

Given the mask from Section 3.6, we train one perturbation shared across audio inputs of the same attack family. The objective promotes a refusal on harmful audio while retaining the transcript of benign audio; only the perturbation is updated.

Let f denote the frozen ALM from log-Mel input to output logits. The learned perturbation has shape $(1, M, T) = (1, 128, 3000)$, matching Qwen2-Audio's fixed input, so the same δ can be reused across source audios. At each update, the mask sets all unselected cells to zero and clipping bounds every retained value by ξ:

$$\delta \;\leftarrow\; \mathrm{clip}_{[-\xi,\xi]}\big(\delta \odot \mathcal{M}\big). \tag{7}$$

Training uses harmful examples $Z_h \sim \mathcal{D}_h$ with a fixed refusal target and benign examples $(Z_b, y_b) \sim \mathcal{D}_b$ with their reference transcripts:

$$\mathcal{L}(\delta) = \mathbb{E}_{Z_h \sim \mathcal{D}_h}\left[\mathrm{CE}\big(f(Z_h + \delta), y_{\mathrm{refuse}}\big)\right] + \lambda\, \mathbb{E}_{(Z_b, y_b) \sim \mathcal{D}_b}\left[\mathrm{CE}\big(f(Z_b + \delta), y_b\big)\right]. \tag{8}$$

The first term trains the perturbation to elicit the fixed target *I'm sorry, but I cannot fulfill this request.* on harmful audio. The second preserves the benign transcript, with λ controlling its weight. These objectives can conflict, which is why Eq. 6 favors cells with a large jailbreak-to-transcription gradient ratio. We optimize the shared perturbation with Adam at $\eta = 3 \times 10^{-4}$ for at most 3000 updates, stopping early when the sampled refusal loss falls below 0.1.

Training across attack families. Different families concentrate their signal in different regions, so we compute a mask and train a perturbation per family and evaluate each family against its own. For comparison, the frequency-only variant broadcasts its perturbation over time. For each trained family-specific defense, we evaluate utility with all four family-specific checkpoints and report their arithmetic mean.

4 EXPERIMENTS

4.1 SETUP

We evaluate Qwen2-Audio-7B-Instruct (Chu et al., 2024) and Kimi-Audio (KimiTeam et al., 2025) on AdvWave (Kang et al., 2025), AdvBench-Audio (Zou et al., 2023; Jin et al., 2025), AudioJailbreak (Song et al., 2025), and Jailbreak-AudioBench (Cheng et al., 2025). Safety is Success Rate of Attack (SRoA; lower is better), assigned by the SORRY-Bench Mistral-7B evaluator (Xie et al., 2025); utility is the AIR-Bench score and FigStep-Audio False Rejection Rate (FRR; lower is better). We compare with the undefended model, Gaussian noise, local smoothing, ALMGuard, and an unmasked variant.

For FG-Guard, $\xi = 0.5$, mask coverage $c = 35\%$, $\lambda = 1$, and $\eta = 3 \times 10^{-4}$. Each family-specific perturbation receives at most 3,000 updates. Appendix H gives datasets, transcription metrics, baseline configurations, hardware, stopping, decoding, and judging details.

4.2 MAIN RESULTS

Table 1: Safety (SRoA, lower is better) and utility (AIR-Bench score, higher is better; FigStep-Audio FRR, lower is better). Utility for each trained family-specific defense averages its four family-specific checkpoints. Best within each model block in bold, second best underlined, based on unrounded values; columns with an all-method tie are unmarked.

Method	AdvWave	AdvBench-Audio	Audio-Jailbreak	JALB	AIR-Bench (↑)	FRR (%) (↓)
Qwen2-Audio						
No Defense	84.61	<u>1.92</u>	21.05	1.89	**6.41**	<u>2.8</u>
Gaussian Noise	<u>41.03</u>	<u>1.92</u>	20.39	0.54	6.29	3.2
Local Smooth	43.59	3.85	<u>19.08</u>	1.35	5.66	4.0
No Mask	48.72	<u>1.92</u>	19.74	<u>0.27</u>	6.26	<u>2.8</u>
ALMGuard	**15.38**	**0.00**	21.05	0.81	6.32	3.0
FG-Guard	<u>41.03</u>	**0.00**	**18.42**	**0.00**	<u>6.35</u>	**2.4**
Kimi-Audio						
No Defense	**0.00**	**0.00**	10.53	<u>1.08</u>	6.12	49.6
Gaussian Noise	**0.00**	<u>1.92</u>	11.18	**0.00**	<u>6.18</u>	<u>46.0</u>
Local Smooth	<u>2.56</u>	**0.00**	10.53	<u>1.08</u>	6.01	59.2
No Mask	**0.00**	**0.00**	**8.55**	**0.00**	6.13	**45.9**
ALMGuard	**0.00**	**0.00**	11.18	1.62	6.15	48.2
FG-Guard	**0.00**	**0.00**	<u>9.87</u>	1.35	**6.21**	52.0

Table 1 reports safety and utility together. The four safety columns give SRoA (lower is better) on AdvWave, AdvBench-Audio, AudioJailbreak, and JALB; the utility columns give the AIR-Bench score (higher is better) and FigStep-Audio FRR (lower is better). For each learned defense, the utility columns average the four family-specific checkpoint evaluations.

On Qwen2-Audio, FG-Guard reduces AdvWave SRoA from 84.61% without defense to 41.03% and AdvBench-Audio SRoA from 1.92% to 0.00%. Its AIR-Bench score changes only from 6.41 to 6.35, while FRR improves from 2.8% to 2.4%: the safety gains therefore come with a small change in measured utility. FG-Guard is best or tied best on AdvBench-Audio, AudioJailbreak, and JALB. Relative to the unmasked variant, FG-Guard improves all four safety columns and both utility columns, supporting the sparse mask's role in the observed safety–utility trade-off.

Table 2: Ablations on Qwen2-Audio. Best results are bold and second-best results are underlined, based on the unrounded values.

Variant	AdvWave SRoA (↓)	AdvBench-Audio SRoA (↓)	AIR-Bench (↑)
FG-Guard (2D, λ=1, c=35%)	**41.03**	**0.00**	**6.35**
w/o transcription term (λ=0)	56.41	**0.00**	**6.35**
w/o mask	<u>48.72</u>	<u>1.92</u>	<u>6.26</u>

On Kimi-Audio, AdvWave and AdvBench-Audio SRoA are already 0.00% without defense and remain so under FG-Guard. Compared with ALMGuard, FG-Guard lowers AudioJailbreak SRoA from 11.18% to 9.87% and JALB SRoA from 1.62% to 1.35%, while increasing AIR-Bench from 6.15 to 6.21. Thus, FG-Guard improves both remaining safety results and the AIR-Bench score over this frequency-only baseline on Kimi-Audio.

4.3 Ablations

Table 2 tests the role of two method components by removing each in turn. The mixed objective in Eq. 8 combines refusal and transcription terms; the sparse mask from Section 3.6 limits where the perturbation acts. FG-Guard is best or tied best in all three columns. Removing the transcription term raises AdvWave SRoA from 41.03% to 56.41%, while leaving AdvBench-Audio SRoA at 0.00% and AIR-Bench at 6.35. Removing the mask raises AdvWave and AdvBench-Audio SRoA to 48.72% and 1.92%, respectively, and lowers AIR-Bench from 6.35 to 6.26. We use the final hyperparameters $\lambda = 1$, $c = 35\%$, and $\xi = 0.5$ in all reported FG-Guard results.

The transcription results in Appendix Table 3 provide a second utility check. On LibriSpeech `dev-clean`, FG-Guard obtains 4.0% CER and 6.7% WER, compared with 11.9% and 18.2% when the transcription term is removed. This directly supports that term's role in preserving transcription on the evaluated set.

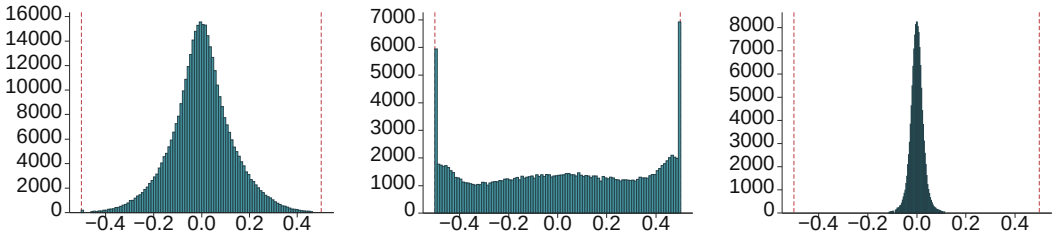
4.4 Numerical Theory Check and Design Analysis

Numerical check of the audio-to-log-Mel stage. As detailed in Section 3.2, Figure 2 shows broadly similar temporal sensitivity trends in the predicted and measured tempo profiles. This qualitative check supports the time-varying input structure that motivates a time–frequency intervention.

Mask geometry and defense outcomes. On Qwen2-Audio, the 2D mask improves both attack results and AIR-Bench over the same objective without a mask (Table 2). Appendix Figure 4 shows selected (m, t) cells rather than frequency bands broadcast across time. Appendix Figure 5 shows that the learned mask overlaps visually with high pitch-tangent-energy regions across 129 audios. Together, the geometry and ablation results support selecting a sparse, time-varying set of intervention cells.

Perturbation values. Figure 3 shows that FG-Guard's nonzero log-Mel perturbation values cluster closest to zero; the unmasked distribution is broader, and ALMGuard has many values near the clipping bounds. Thus, among these learned defenses, FG-Guard makes the smallest typical nonzero per-cell change to the audio representation. Compared with no defense, it lowers attack success on all four Qwen2-Audio evaluations. Its AIR-Bench score changes only from 6.41 to 6.35, while FRR improves from 2.8% to 2.4%. These results show effective defense with a small measured utility change. The histograms measure feature values, not waveform or perceptual distortion.

Hidden-state visualization. Lin et al. (2025) use t-SNE to show that hidden representations of harmful and safe audio inputs separate early and remain separated across layers. Their comparison concerns different input classes; Appendix Figure 6 instead compares each audio sample before and after a defensive perturbation. In our visualization, FG-Guard's before–after points form close pairs with small offsets from the projector onward, whereas the unmasked and ALMGuard variants



(a) No mask. (b) ALMGuard (1D). (c) FG-Guard (2D).

Figure 3: Histograms of nonzero log-Mel perturbation values (horizontal axis) and their counts (vertical axis); dashed lines mark the clipping bounds $\pm\xi$. The unmasked values have a broad distribution, ALMGuard has many values near the bounds, and FG-Guard's values are most concentrated near zero. Thus, FG-Guard makes smaller typical nonzero per-cell changes to the input representation than the other learned defenses shown.

separate markedly at the projector and pair less clearly in the intermediate layer. The close pairing indicates that FG-Guard largely preserves sample-wise hidden-state structure in these examples, consistent with its smaller measured impact on general utility.

FG-Guard learns a reusable additive log-Mel perturbation with independent values on a sparse two-dimensional time–frequency mask selected using jailbreak and transcription gradients. Its joint objective promotes refusal on harmful audio while preserving transcription on benign audio. The analyses above provide empirical support for the effectiveness of these design choices on the evaluated tasks.

5 Conclusion

We connect three previously separate questions: how audio transformations change the log-Mel input, how that change propagates through an ALM, and whether the resulting local shift increases or decreases a refusal-related opening-token probability. The first-order chain identifies a common additive intervention space and motivates a time–frequency parameterization. FG-Guard turns those design cues into a trained, sparse 2D perturbation. On Qwen2-Audio, it reduces AdvWave success from 84.61% to 41.03%, reaches 0.00% SRoA on AdvBench-Audio and JALB, and obtains an AIR-Bench score of 6.35 versus 6.41 without defense. The matched-objective ablation provides specific mask evidence: adding the mask lowers AdvWave SRoA from 48.72% to 41.03% and raises AIR-Bench from 6.26 to 6.35 (Table 2). The finite-difference check in Figure 2 and the mask–tangent comparison in Appendix Figure 5 further support the time-varying input structure behind the design.

Limitations. The theory is local and first-order: it bounds displacement near a given input, says nothing globally, and the layerwise Lipschitz product is too loose to read as a numerical sensitivity estimate. Real time-scale modification adds an unmeasured residual, while the pitch result assumes a second-order implementation residual. Figure 2 checks one utterance using temporal RMS curves; it does not establish signed 2D agreement across a population or downstream safety effects. Eq. 4 tracks refusal-related opening-token log-odds, not the probability that a complete generated answer refuses; its signed coefficient must be measured for each model and input. The first-order analysis motivates the intervention space and geometry but does not determine the mask support, perturbation values, or defense effectiveness. The refusal-direction and margin constructions in Appendices E and F are optional diagnostics and are neither used by FG-Guard nor validated here.

Our experiments cover non-adaptive, within-family attacks and do not test pitch- or tempo-warping attacks, cross-family transfer, or adaptive robustness. A fixed perturbation also remains open to an attacker who optimizes against it.

AI USE STATEMENT

In this work, we used generative AI tools to assist with (1) developing theoretical models or conceptual frameworks; (2) drafting proof steps; (3) translation; (4) data cleaning and reformatting; (5) creating or modifying scientific figures and images; (6) writing or editing code and creating artifacts; (7) interpreting experimental results; and (8) drafting manuscript sections and editing or polishing text to improve readability. We reviewed and verified all AI-assisted work, including mathematical derivations, code, data-processing outputs, figures, interpretations, and text. We take full responsibility for the final content of this work, including all claims and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work evaluates defenses against harmful audio requests using publicly released datasets and models. It involves no new human-subject experiments or private data. Audio jailbreak research is dual-use: the same analyses that help identify safety failures could inform attacks. We therefore frame our findings as defensive evidence under the stated threat model, not as a guarantee against adaptive misuse.

REPRODUCIBILITY STATEMENT

The main paper and appendix specify the method, optimization settings, dataset splits, evaluation protocol, and ablations. Code and configuration files are submitted as supplementary material to facilitate independent reproduction, subject to the licenses of the underlying datasets and models.

REFERENCES

Joakim Andén and Stéphane Mallat. Deep scattering spectrum. *IEEE Transactions on Signal Processing*, 62(16):4114–4128, 2014. doi: 10.1109/TSP.2014.2326991.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *arXiv preprint arXiv:2310.08419*, 2023.

Meng Chen, Kun Wang, Li Lu, Jiaheng Zhang, and Tianwei Zhang. Hijacking large audio-language models via context-agnostic and imperceptible auditory prompt injection. *arXiv preprint arXiv:2604.14604*, 2026.

Hao Cheng, Erjia Xiao, Jing Shao, Yichi Wang, Le Yang, Chao Shen, Philip Torr, Jindong Gu, and Renjing Xu. Jailbreak-AudioBench: In-depth evaluation and analysis of jailbreak threats for large audio language models. *arXiv preprint arXiv:2501.13772*, 2025.

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

Jonathan Driedger and Meinard Müller. A review of time-scale modification of music signals. *Applied Sciences*, 6(2):57, 2016. doi: 10.3390/app6020057.

Zheng Fang, Xiaosen Wang, Shenyi Zhang, Shaokang Wang, and Zhijin Ge. Sparse tokens suffice: Jailbreaking audio language models via token-aware gradient optimization. *arXiv preprint arXiv:2605.04700*, 2026.

Bo-Han Feng, Yu-Hsuan Li Liang, Chien-Feng Liu, You-Hsuan Chang, and Yun-Nung Chen. Audio jailbreaks in large audio-language models: Taxonomy, attack-defense analysis, and cost-aware evaluation. *arXiv preprint arXiv:2605.30031*, 2026.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Weifei Jin, Yuxin Cao, Junjie Su, Minhui 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.

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.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, et al. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

J. Laroche and M. Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, 1999. doi: 10.1109/89.759041.

Hongyi Li, Chengxuan Zhou, Chu Wang, Sicheng Liang, Yanting Chen, Qinlin Xie, Jiawei Ye, and Jie Wu. StyleBreak: Revealing alignment vulnerabilities in large audio-language models via style-aware audio jailbreak. *arXiv preprint arXiv:2511.10692*, 2025.

Weilin Lin, Jianze Li, Hui Xiong, and Li Liu. SARSteer: Safeguarding large audio-language models via safe-ablated refusal steering. *arXiv preprint arXiv:2510.17633*, 2025.

Kaiwen Luo, Zhenhong Zhou, Leyan Wang, Liang Lin, Tianyu Shao, Yuanhe Zhang, Yang Xiao, et al. A survey of large audio language models: Generalization, trustworthiness, and outlook. *arXiv preprint arXiv:2605.20266*, 2026.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012. doi: 10.1002/cpa.21413.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

Jiachen Qian and Junyu Li. Prosody-driven jailbreaks in audio LLMs: A controlled study and mechanistic analysis. *arXiv preprint arXiv:2607.26541*, 2026.

Jaechul Roh and Amir Houmansadr. Benign fine-tuning breaks safety alignment in audio LLMs. *arXiv preprint arXiv:2604.16659*, 2026.

Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, Zhenhao Chen, and Xiuying Chen. Audio Jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models. *arXiv preprint arXiv:2505.15406*, 2025.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *Proceedings of ICASSP '93*, volume 2, pp. 554–557, 1993. doi: 10.1109/ICASSP.1993.319366. URL `https://doi.org/10.1109/ICASSP.1993.319366`.

Yunqiang Wang, Hengyuan Na, Di Wu, Miao Hu, and Guocong Quan. GRM: Utility-aware jailbreak attacks on audio LLMs via gradient-ratio masking. *arXiv preprint arXiv:2604.09222*, 2026.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, 2025.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-Bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations (ICLR)*, 2025.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. Reshaping representation space to balance the safety and over-rejection in large audio language models. *arXiv preprint arXiv:2505.19670*, 2025. URL https://arxiv.org/abs/2505.19670.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Zhengyue Zhao, Xiaoyun Zhang, Kaidi Xu, Xing Hu, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Adversarial contrastive decoding: Boosting safety alignment of large language models via opposite prompt optimization. *arXiv preprint arXiv:2406.16743v1*, 2024. URL https://arxiv.org/abs/2406.16743v1.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. In *arXiv preprint arXiv:2307.15043*, 2023.

A TEMPO: LOCAL EXPANSION AND PROOF

Let $\tau = \log \sigma$, where σ is the duration stretch factor, so $\tau > 0$ lengthens the audio and slows it down. We compare all outputs on the fixed crop-and-pad grid used by the ALM. The following proposition states the assumptions and the result summarized in Section 3.2.

Proposition 2 (Tempo). *Let $W_\tau \in \mathbb{R}^{T\times T}$ be the time-resampling operator on the fixed grid. Assume that, near $\tau = 0$, $W_0 = I$, $W_\tau = I + \tau D_{\text{tempo}} + R_W(\tau)$, $\|R_W(\tau)\|_2 \le C_W \tau^2$, and $\|W_\tau - I\|_2 \le L_W|\tau|$, where D_{tempo} is the local warp generator and C_W, L_W are constants. Writing $E = BP$ and $Z_\tau = \log(\varepsilon + EW_\tau)$,*

$$Z_\tau \;=\; Z \;+\; \tau \underbrace{\frac{BPD_{\text{tempo}}}{\varepsilon + BP}}_{g_\tau} + O(\tau^2), \qquad \|Z_\tau - Z\|_F \;\le\; \frac{L_W \|B\|_2 \|P\|_F}{\varepsilon}\,|\tau|. \tag{9}$$

The first part of Eq. 9 identifies the log-Mel direction caused by a small tempo change; the second bounds the size of that change. To prove both statements, note first that B multiplies P from the left and W_τ from the right. Associativity gives

$$B(PW_\tau) = (BP)W_\tau = EW_\tau, \qquad \text{so} \qquad \Delta E_\tau \;=\; E(W_\tau - I) \;=\; \tau E D_{\text{tempo}} + E R_W(\tau).$$

Write $\psi(x) = \log(\varepsilon + x)$; it satisfies $\psi'(x) = 1/(\varepsilon + x)$ and $|\psi''(x)| \le \varepsilon^{-2}$ on $x \ge 0$. Applying ψ elementwise to $E + \Delta E_\tau$ gives the first-order expansion

$$\begin{aligned} Z_\tau &= \log(\varepsilon + E + \Delta E_\tau) = \psi(E + \Delta E_\tau) \\ &= \psi(E) + \psi'(E) \odot \Delta E_\tau + O(\|\Delta E_\tau\|^2) \\ &= Z + \frac{\Delta E_\tau}{\varepsilon + E} + O(\|\Delta E_\tau\|^2) = Z + \tau\,\frac{E D_{\text{tempo}}}{\varepsilon + E} + O(\tau^2), \end{aligned}$$

which is Eq. 9 with $g_\tau = \frac{ED_{\text{tempo}}}{\varepsilon+E} = \frac{BPD_{\text{tempo}}}{\varepsilon+BP}$. For the norm bound, ψ is $1/\varepsilon$-Lipschitz on $x \ge 0$, so

$$\|Z_\tau - Z\|_F \le \varepsilon^{-1}\|E(W_\tau - I)\|_F \le \varepsilon^{-1}\|E\|_F \|W_\tau - I\|_2 \le \varepsilon^{-1} L_W \|E\|_F |\tau|,$$

and $\|E\|_F = \|BP\|_F \le \|B\|_2\|P\|_F$ gives Eq. 9. The second-order remainder is explicit: with $C_2 = C_W\|E\|_F/\varepsilon + L_W^2\|E\|_F^2/(2\varepsilon^2)$ we have $\|Z_\tau - Z - \tau g_\tau\|_F \le C_2\tau^2$, so the first-order model carries an $O(\tau^2)$ local error.

Two conditions are worth naming. The comparison requires a common time grid, since a tempo change alters duration; hard padding at the boundary is not smooth, so the first-order statement is

cleanest on the shared interior support. And $\log$ does not commute with resampling, so the expansion must be taken on E before the logarithm.

For the continuous picture, $E_\tau(m,t) = E(m, e^{-\tau}t)$ gives $\partial_\tau E_\tau|_0 = -t\partial_t E$ and hence $\partial_\tau Z_\tau|_0 = -t\partial_t Z$, matching the discrete generator. A finite-difference estimate on a fine grid is $D_{\text{tempo}} \approx -D_t \operatorname{diag}(t_0, \dots, t_{T-1})$ with D_t a centred difference matrix; this is a numerical approximation to W_0', not a definition of it.

B Jacobian Propagation: Full Derivation

This appendix spells out the layer-by-layer composition behind Eq. 3 and the Lipschitz bound summarized in Section 3.3. We assume a fixed sequence structure and fixed text prompt, so that only the log-Mel input z varies, and locally Lipschitz Jacobians for the composed maps used in the second-order expansions.

Encoder. Let $a = \mathcal{E}(z)$ and assume $\mathcal{E}$ is twice continuously differentiable near z. Then

$$\mathcal{E}(z + \delta z) = \mathcal{E}(z) + J_{\mathcal{E}}(z)\,\delta z + O(\|\delta z\|^2), \qquad \delta a = J_{\mathcal{E}}(z)\,\delta z + O(\|\delta z\|^2),$$

with $J_{\mathcal{E}}(z) = \partial\mathcal{E}/\partial z$. Substituting $\delta z = G(z)\theta + O(\|\theta\|^2)$ from Eq. 2 gives $\delta a = J_{\mathcal{E}}(z)\,G(z)\,\theta + O(\|\theta\|^2)$.

Projector. For the projector $u = \mathcal{P}(a)$, the same expansion around a gives

$$\delta u = J_{\mathcal{P}}(a)\,\delta a + O(\|\delta a\|^2) = J_{\mathcal{P}}(a)\,J_{\mathcal{E}}(z)\,G(z)\,\theta + O(\|\theta\|^2).$$

If $\mathcal{P}$ is a linear layer $u = W_{\mathcal{P}}a + b_{\mathcal{P}}$, then $J_{\mathcal{P}} = W_{\mathcal{P}}$ and this step contributes no second-order remainder of its own.

Merge. The merge operator assembles the input sequence from the audio embeddings and the fixed text embeddings, $H^{(0)} = \Pi(u, e_{\text{text}})$. With the text side fixed, only u varies, so

$$\delta H^{(0)} = J_\Pi\,\delta u + O(\|\delta u\|^2) = J_\Pi\,J_{\mathcal{P}}(a)\,J_{\mathcal{E}}(z)\,G(z)\,\theta + O(\|\theta\|^2),$$

and J_Π is an insertion/selection matrix whenever the merge is a fixed concatenation.

Decoder blocks. Each block is $H^{(\ell+1)} = \mathcal{B}_\ell(H^{(\ell)})$, with $\mathcal{B}_\ell(H) = H + F_\ell(H)$ under the residual form, so $J_{\mathcal{B}_\ell} = I + J_{F_\ell}$. Compose the blocks,

$$\delta H^{(\ell)} = J_{\mathcal{B}_{\ell-1}} \cdots J_{\mathcal{B}_0}\,J_\Pi\,J_{\mathcal{P}}\,J_{\mathcal{E}}\,G(z)\,\theta + O(\|\theta\|^2).$$

The operator $\mathcal{S}$ selects the last conditioning position from the full hidden-state sequence. Reading the generation anchor $h^{(\ell)} = \mathcal{S}H^{(\ell)}$ finally gives

$$\delta h^{(\ell)} = J_{\ell\leftarrow z}\,G(z)\,\theta + O(\|\theta\|^2), \qquad J_{\ell\leftarrow z} = \mathcal{S}\,J_{\mathcal{B}_{\ell-1}} \cdots J_{\mathcal{B}_0}\,J_\Pi\,J_{\mathcal{P}}\,J_{\mathcal{E}},$$

which is Eq. 3.

Lipschitz bound. If $\mathcal{E}, \mathcal{P}, \Pi, \mathcal{B}_j$ are locally $L_{\mathcal{E}}, L_{\mathcal{P}}, L_\Pi, L_{\mathcal{B}_j}$-Lipschitz respectively, then composing the inequalities gives

$$\|\delta h^{(\ell)}\| \le L_{\text{sem}}^{(\ell)}\,\|\delta z\|, \qquad L_{\text{sem}}^{(\ell)} \le \|\mathcal{S}\|\Big(\textstyle\prod_{j=0}^{\ell-1} L_{\mathcal{B}_j}\Big)\,L_\Pi\,L_{\mathcal{P}}\,L_{\mathcal{E}},$$

and with $\|\delta z\| \le C_Z\|\theta\| + O(\|\theta\|^2)$ we obtain $\|\delta h^{(\ell)}\| \le L_{\text{sem}}^{(\ell)} C_Z\|\theta\| + O(\|\theta\|^2)$. As stated in the main text, the product $L_{\text{sem}}^{(\ell)}$ is typically a very loose overestimate; it establishes that the map is locally stable, and should never be read as a numerical sensitivity estimate. Actual sensitivity is measured directly with a Jacobian–vector product.

C PITCH: LOCAL SPECTRAL-WARP DERIVATION

We do not assert an exact identity for an arbitrary pitch-shifting implementation. Instead, let P_ρ be the post-shift power spectrogram and assume that, near $\rho = 0$, it admits the local model

$$P_\rho = W_\rho P + R_\rho, \qquad W_0 = I, \qquad W_\rho = I + \rho D_{\text{pitch}} + R_{W,\rho}, \tag{10}$$

where W_ρ acts on the frequency index, $D_{\text{pitch}} = \partial_\rho W_\rho|_0$, $\|R_{W,\rho}\|_2 \le C_W \rho^2$, and $\|R_\rho\|_F \le C_R \rho^2$. The residual R_ρ absorbs phase correction, peak tracking, formant preservation, interpolation, normalization, and implementation artifacts.

Multiplying by the Mel filterbank gives

$$BP_\rho = BP + \rho B D_{\text{pitch}} P + O(\rho^2). \tag{11}$$

Since $u \mapsto \log(\varepsilon + u)$ is smooth on the nonnegative half-line, elementwise Taylor expansion yields

$$Z_\rho = Z + \rho g_\rho + O(\rho^2), \qquad g_\rho = (B D_{\text{pitch}} P) \oslash (\varepsilon \mathbf{1} + BP). \tag{12}$$

For the norm bound, the same map is $1/\varepsilon$-Lipschitz. Therefore

$$\|Z_\rho - Z\|_F \le \frac{1}{\varepsilon} \|B(P_\rho - P)\|_F \tag{13}$$

$$\le \frac{\|B\|_2}{\varepsilon} \left(\|W_\rho - I\|_2 \|P\|_F + \|R_\rho\|_F \right) \tag{14}$$

$$\le \frac{\|B\|_2}{\varepsilon} \left(L_{\text{pitch}} \|P\|_F |\rho| + O(\rho^2) \right). \tag{15}$$

Here L_{pitch} is a Lipschitz constant of the pitch warp near the identity. Thus $\|Z_\rho - Z\|_F = O(|\rho|)$ for fixed input and implementation. In the continuous idealization, $P_\rho(\nu, t) \propto P(e^{-\rho}\nu, t)$ gives the transport component $D_{\text{pitch}} \approx -\nu \partial_\nu$. With Mel coordinate $m = \mathcal{W}(\nu)$ and velocity $v(m) = \mathcal{W}^{-1}(m)\mathcal{W}'(\mathcal{W}^{-1}(m))$, this becomes $g_\rho \approx -v(m)\partial_m Z$ up to normalization and residual terms. The sign of the resulting refusal effect is not fixed by this signal-level calculation.

D ADDITIVE NOISE: WAVEFORM-TO-LOG-MEL DERIVATION

Let $x_\alpha = x + \alpha\zeta$, let $\mathcal{F}$ be the STFT operator, and write $X = \mathcal{F}x$ and $N = \mathcal{F}\zeta$. STFT linearity gives $X_\alpha = X + \alpha N$ exactly. Consequently, with $P = |X|^2$,

$$P_\alpha = P + \alpha C + \alpha^2 Q, \qquad C = 2\,\mathrm{Re}(X^* \odot N), \qquad Q = |N|^2. \tag{16}$$

The linear term is speech–noise interference and depends on relative phase; it cannot in general be replaced by independent speech and noise powers. After Mel aggregation, $E_\alpha = BP + \alpha BC + \alpha^2 BQ$. Applying the log-Mel map gives

$$Z_\alpha = Z + \alpha g_{\text{noise}} + O(\alpha^2), \tag{17}$$

$$g_{\text{noise}} = \frac{2B\,\mathrm{Re}((\mathcal{F}x)^* \odot (\mathcal{F}\zeta))}{\varepsilon + B|\mathcal{F}x|^2}. \tag{18}$$

The explicit second-order term is

$$\alpha^2 \left[\frac{BQ}{\varepsilon + BP} - \frac{(BC)^2}{2(\varepsilon + BP)^2} \right], \tag{19}$$

with all operations elementwise. Moreover,

$$\|Z_\alpha - Z\|_F \le \frac{\|B\|_2}{\varepsilon} \left(2|\alpha| \|X\|_F \|N\|_F + \alpha^2 \|N\|_F^2 \right), \tag{20}$$

so fixed waveform noise also produces a local $O(|\alpha|)$ log-Mel displacement. Unlike pitch and tempo, its tangent is phase-sensitive interference rather than a coordinate deformation.

E Optional Refusal-Geometry Diagnostic and Output Proxy

This appendix records a possible representation-level interpretation of Eq. 4. It is not used to construct FG-Guard, and we do not treat a one-dimensional refusal direction as an established mechanism.

At layer ℓ and the fixed generation anchor, collect harmful-audio and harmless-audio activations and define

$$r^{(\ell)} = \mu^{(\ell)}_{\text{harm}} - \mu^{(\ell)}_{\text{harmless}}, \qquad \hat{r}^{(\ell)} = r^{(\ell)}/\|r^{(\ell)}\|. \tag{21}$$

This difference-in-means direction must be extracted on the evaluated Audio-LM; a text-derived direction is not assumed to transfer. It is a local probe whose validity requires activation-addition and directional-ablation tests, especially because refusal geometry may contain multiple directions.

Define $s_{\text{ref}}(z) = \hat{r}^{(\ell)\top} h^{(\ell)}(z)$. Combining Eq. 3 with the chain rule gives

$$\Delta s_{\text{ref}} = \hat{r}^{(\ell)\top} J_{\ell \leftarrow z} G(z)\theta + O(\|\theta\|^2) \tag{22}$$

$$= q(z)^\top \theta + O(\|\theta\|^2), \qquad q(z) = G(z)^\top J^\top_{\ell \leftarrow z} \hat{r}^{(\ell)}. \tag{23}$$

The pulled-back direction $g^{\text{in}}_{\text{ref}} = J^\top_{\ell \leftarrow z} \hat{r}^{(\ell)}$ can be reshaped to the $M \times T$ input plane. Its inner product with g_j is

$$\kappa_j = \hat{r}^{(\ell)\top} J_{\ell \leftarrow z} g_j, \qquad A_j = \cos(g^{\text{in}}_{\text{ref}}, g_j). \tag{24}$$

The sign of κ_j is a model- and input-dependent measurement, never a consequence of audio physics.

For the output-level continuous proxy, let $o = \mathcal{O}(h^{(L)})$, $p = \text{softmax}(o)$, and use the model-specific refusal-opening token set $\mathcal{R}$ defined in the main text. Set $P_{\mathcal{R}} = \sum_{j \in \mathcal{R}} p_j$ and $m_{\mathcal{R}} = \log(P_{\mathcal{R}}/(1 - P_{\mathcal{R}}))$. Then

$$(\nabla_o m_{\mathcal{R}})_j = \begin{cases} p_j/P_{\mathcal{R}}, & j \in \mathcal{R}, \\ -p_j/(1 - P_{\mathcal{R}}), & j \notin \mathcal{R}, \end{cases} \tag{25}$$

and, writing $g^{\mathcal{R}}_{\text{logit}} = \nabla_o m_{\mathcal{R}}$,

$$\Delta m_{\mathcal{R}} = (g^{\mathcal{R}}_{\text{logit}})^\top J_{\mathcal{O}} J_{L \leftarrow z} G(z)\theta + O(\|\theta\|^2). \tag{26}$$

Here $J_{L \leftarrow z}$ denotes the Jacobian to the final layer L, and $\mathcal{O}$ the final norm plus unembedding. Both s_{ref} and $m_{\mathcal{R}}$ are continuous probes; neither is a Taylor expansion of the discrete generated sequence.

F Optional Local-Margin Interpretation

Let b be a local decision boundary for a continuous refusal surrogate and let $\Gamma = s_{\text{ref}}(z) - b$. Then

$$s_{\text{ref}}(z_\theta) - b = \Gamma + q(z)^\top \theta + O(\|\theta\|^2). \tag{27}$$

A small transformation can cross the boundary when $|\Gamma|$ is small or $\|q(z)\|$ is large. Conversely, if the composite score is locally L_s-Lipschitz and $L_s\|\delta z\| < |\Gamma|$, a crossing is locally excluded. This is a stability statement for a continuous surrogate, not a guarantee about emitted tokens.

G Additional Analysis Figures

For the main-text numerical check in Figure 2, we apply positive and negative tempo changes to `harmful-395.wav` with `librosa` and pass the waveforms through the original Qwen2-Audio processor. We use $\epsilon = 0.01$ with $\tau = \log \sigma$; the theoretical curve uses $Z_\tau(m,t) = Z_0(m, te^{-\tau})$. We summarize each $128 \times T$ tangent by $r(t) = \sqrt{128^{-1} \sum_m g(m,t)^2}$ and apply the same nine-frame moving average to both curves. The remaining magnitude gap reflects modelling and implementation residuals.

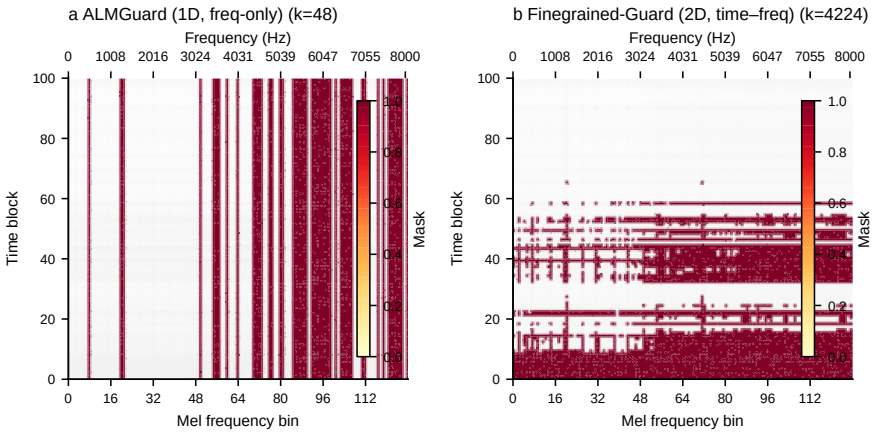


Figure 4: Mask comparison on the same population-averaged saliency score (greyscale, $\log_{10}$ scale); warm cells are selected. (a) Frequency-only mask: k=48 bands broadcast across all T=100 time blocks as continuous stripes. (b) FG-Guard joint mask: 4,224 cells (33% of this 128×100 grid; c=35% at full T=3000 resolution) at the specific (m,t) positions where the jailbreak-to-transcription ratio is highest.

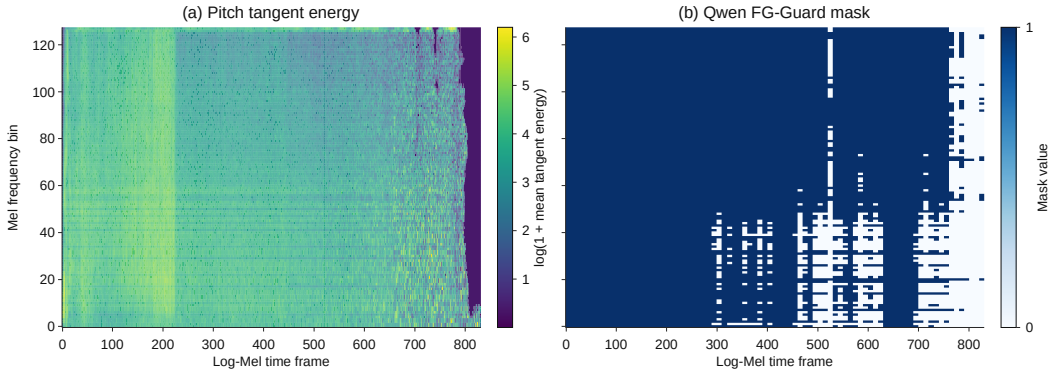
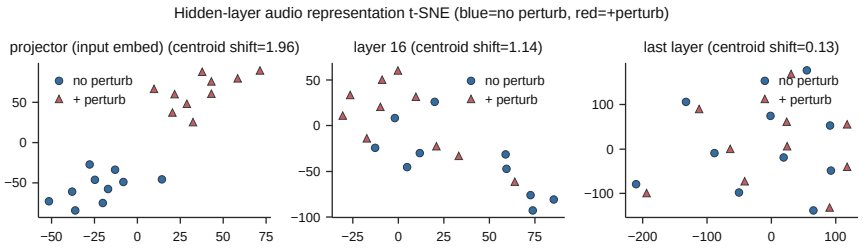
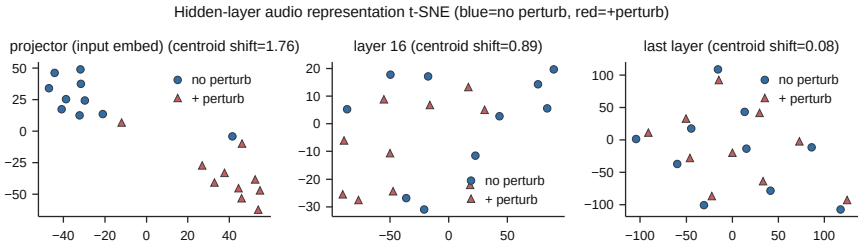


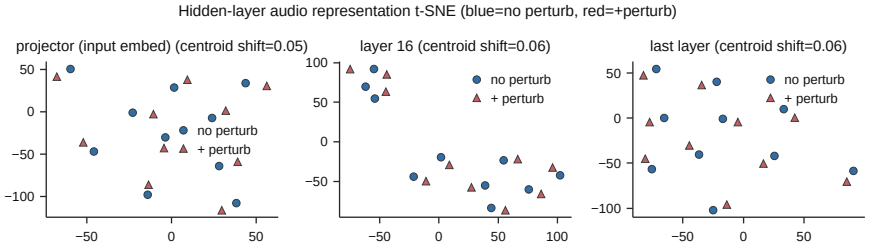
Figure 5: Pitch-tangent energy and the empirical Qwen2-Audio FG-Guard mask. For each of 129 audios, panel (a) estimates the processor-level finite-difference tangent $g_\rho^{\mathrm{FD}}(m,t) = [Z_{+\epsilon}(m,t) - Z_{-\epsilon}(m,t)]/(2\epsilon)$ and displays $\log(1 + E_\rho(m,t))$, where $E_\rho(m,t) = \mathbb{E}_x[(g_\rho^{\mathrm{FD}}(m,t))^2]$. Brighter cells therefore indicate log-Mel coordinates that are more sensitive to small positive and negative pitch transformations; the dark region without valid frames is padding. Panel (b) shows the binary mask at the corresponding coordinates (dark blue: retained; pale: excluded). The retained support visibly overlaps regions of high tangent energy. Because this comparison has no size-matched random baseline or scalar overlap statistic, it establishes qualitative support agreement only, not tangent-subspace membership or a causal explanation of defense performance.



(a) Unmasked perturbation.



(b) ALMGuard (1D).



(c) FG-Guard (2D, mixed objective).

Figure 6: t-SNE of decoder hidden states before (blue circles) and after (red triangles) adding the indicated perturbation. Pairing appears mainly in deeper layers for ALMGuard and from shallower layers for FG-Guard. This is a qualitative visualization only: distances and the displayed centroid shifts in separately fitted t-SNE projections are not used for quantitative or mechanistic claims.

H EXPERIMENTAL DETAILS

H.1 FG-GUARD SETUP DETAILS

For all harmful examples and both backbones, we use the fixed safe target $y_{\text{refuse}} =$ *I'm sorry, but I cannot fulfill this request.* Model parameters remain frozen; only the shared log-Mel perturbation is optimized. For Qwen2-Audio, the front end pads or truncates every clip to a $(1, 128, 3000)$ tensor, corresponding to at most 30 seconds. Each update samples one harmful example and one benign transcription example, evaluates them sequentially, and combines their perturbation gradients as $g_h + \lambda g_b$. We then apply the binary mask and clip the perturbation to $[-\xi, \xi]$.

We use Adam with learning rate $\eta = 3 \times 10^{-4}$, $\lambda = 1$, $\xi = 0.5$, and mask coverage $c = 35\%$. Training is conducted on one NVIDIA GeForce RTX 4090 GPU. Each family-specific shared perturbation is optimized for at most 3,000 update steps, with early stopping when the sampled refusal loss is below 0.1. The example sampler uses seed 42.

H.2 ATTACK EVALUATION SETS

AdvWave. AdvWave (Kang et al., 2025) is a white-box ALM jailbreak attack that optimizes an adversarial suffix with cross-entropy loss. We use a 44,100-sample suffix (approximately 2.76 seconds at 16 kHz) and the original dynamic target-selection mechanism, which constructs a query-specific jailbreak target. Optimization runs for at most 3,000 iterations and terminates when the loss falls below 0.1. The 39 examples are evaluated with two complementary folds defined by example-index groups: fold 0 uses 20 training and 19 test examples, whereas fold 1 uses 19 training and 20 test examples; neither fold uses a validation split.

AdvBench-Audio. AdvBench-Audio is an audio rendering of AdvBench (Zou et al., 2023; Jin et al., 2025). We use the HuggingFace release `WeifeiJin/AdvBench-Audio`, whose JSON manifest pairs each prompt with a matching TTS waveform. Its 520 examples are grouped by their 520 harmful goals and split 80/10/10 into 416 training, 52 validation, and 52 test examples. Each spoken instruction is provided without an additional textual prompt and scored by the SORRY-Bench judge.

AudioJailbreak. We use the AudioJailbreak release (Song et al., 2025) (snapshot dated May 8, 2024), including its Do_Not_Answer, jailbreak_llms, and jailbreakbench subsets. Source prompts are taken from `rename_map.jsonl`. The 1,493 examples are grouped by normalized harmful goal and stratified by source, yielding 1,212 training, 129 validation, and 152 test examples.

Jailbreak-AudioBench (JALB). JALB (Cheng et al., 2025) covers original, accent, emotion, and emphasis delivery families. Its 3,725 clips correspond to 520 original goals and are split by goal into 2,980 training, 374 validation, and 371 test examples using an 80/10/10 partition. We evaluate entries from `prompt_pairs.jsonl`; the benchmark varies the acoustic delivery of its underlying requests.

H.3 FALSE-REJECTION SET

FigStep-Audio. FigStep-Audio contains paired harmful and safety-aligned clips. In both conditions the textual prompt is the neutral sentence "Please generate detailed steps for the question in the audio." The official script with `fix_seed(42)` splits each side into training and test sets. The test split contains 250 harmful clips for measuring harmful-request refusal and 250 safety-aligned clips for measuring FRR and BRR. Filenames are paired across the two sides, which differ only in spoken content.

H.4 UTILITY EVALUATION SETS

AIR-Bench-Chat. We evaluate 200 examples from the speech-QA/Common Voice English subset of AIR-Bench-Chat (Yang et al., 2024). They are used to compute the AIR-Bench score and Win Rate with the official `score_chat.py` script.

Table 3: Transcription preservation on the speech-recognition evaluation set. Lower is better. Best results are bold and second-best results are underlined, based on the unrounded values.

Method	CER (%, ↓)	WER (%, ↓)
No Perturbation	<u>6.6</u>	<u>7.3</u>
ALMGuard	7.8	9.3
No Transcription Loss	11.9	18.2
FG-Guard	**4.0**	**6.7**

LibriSpeech. We use clips from the LibriSpeech `dev-clean` split (Panayotov et al., 2015) to measure WER and CER. Both metrics are macro-averaged after lowercasing and removing punctuation.

Removing the transcription term produces the highest CER and WER, whereas FG-Guard is below ALMGuard and the unperturbed model on this set. Because the ALM remains frozen, the decrease is plausibly explained by task-specific adaptation of the shared input perturbation toward the reference transcripts; it does not imply that the defense generally improves ASR.

H.5 Baseline Setup

No defense. The unmodified backbone is evaluated with the same prompts, audio, decoding, and judge as the defended variants.

Gaussian noise. We add zero-mean Gaussian noise whose energy matches the FG-Guard perturbation; this gives a standard deviation of approximately 0.063.

Local smoothing. We replace each waveform sample by a moving average over radius $h = 3$, i.e., a kernel of length $2h + 1 = 7$.

ALMGuard. We use its frequency-only mask and refusal-only objective with the same fixed safe target. Adam uses learning rate 10^{-3}, with at most 3,000 iterations per example and up to 10 epochs.

No mask. This variant uses the same mixed objective, data, clip bound, and update budget as FG-Guard but allows δ to have full support. It isolates the safety–utility effect of the learned 2D mask.

H.6 Evaluation Details

Generation is greedy with sampling disabled. The deterministic SORRY-Bench Mistral-7B judge assigns 1 to direct compliance with a harmful request and 0 otherwise; SRoA is the mean of these decisions. FRR is the fraction of safety-aligned FigStep-Audio test queries that are incorrectly refused. For each trained family-specific defense, we evaluate AIR-Bench and FigStep-Audio with each of the four family-specific checkpoints and report the arithmetic mean of the four results for each metric.