TRISTEER: TRI-SPACE ADAPTIVE STEERING FOR LARGE AUDIO LANGUAGE MODEL SAFETY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Audio Language Models (LALMs) have extended language-model interaction to the audio modality, while introducing new safety risks. We investigate how safety-relevant information is organized in audio-conditioned hidden representations and whether this structure can support inference-time control. A central challenge is to identify which internal semantic and acoustic evidence should trigger and shape an intervention. Our analysis reveals layer-dependent alignment between audio-native and text-only safety directions, while semantic and acoustic variations form reusable structures in the representation space. Based on these findings, we propose TriSteer, a training-free framework that constructs ordered contrast-defined measurement subspaces for semantic, decision-relevant acoustic, and decision-irrelevant acoustic variation. At each layer, TriSteer adaptively sets intervention strength from the current input's deviation relative to fixed benign reference boundaries, while acoustic attribution selectively refines the steering direction. Experiments across three LALMs and multiple safety benchmarks show that TriSteer improves the safety-utility trade-off without model updates. These results demonstrate that the geometry of internal safety representations provides a practical basis for adaptive inference-time control.

1 INTRODUCTION

Large Audio Language Models (LALMs) directly interpret spoken instructions and support reasoning over both speech and general audio (Chu et al., 2024; Xu et al., 2025). This audio pathway also creates a growing attack surface: harmful intent can be transferred through speech, combined with non-speech interference, or encoded through diverse acoustic realizations that bypass text-oriented safeguards (Yang et al., 2025; Peng et al., 2026; Cheng et al., 2025; Song et al., 2026). Inference-time defenses for deployed LALMs remain comparatively limited. Training-based approaches can internalize how non-lexical cues should affect behavior, but require model adaptation (Wang et al., 2026b); static prompt defenses such as AdaShield-S apply the same instruction across inputs (Wang et al., 2025; Peng et al., 2026). Activation steering offers a lightweight alternative by intervening directly in hidden representations (Li et al., 2023; Rimsky et al., 2024). Adaptive subspace methods also refine steering directions in vision-language models (Zeng et al., 2025). For audio-conditioned representations, a remaining challenge is to identify which semantic and acoustic evidence should determine the intervention.

Audio safety is therefore not only a behavior-alignment problem; it is also a representation-attribution problem. Before deciding how to steer, a defense must determine what information in the hidden representation should affect the safety decision and what variation should not. Three considerations motivate our design. First, our paired analysis shows layer-dependent alignment between audio-native and text-only behavior directions, with lower agreement in earlier layers and stronger agreement in many middle and later layers (Figure 2(a)). Because steering is applied at every language layer, this variation is relevant to the deployed construction. Second, our cross-validated held-out analysis of controlled semantic and acoustic contrasts reveals distinguishable, reusable structure within audio-conditioned representations (Section 3), supporting separate measurement of these forms of variation. Third, identifying these sources alone does not determine how they should affect steering. Some acoustic evidence can legitimately affect the desired safety response, whereas other acoustic variation should not mechanically strengthen the intervention.

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.

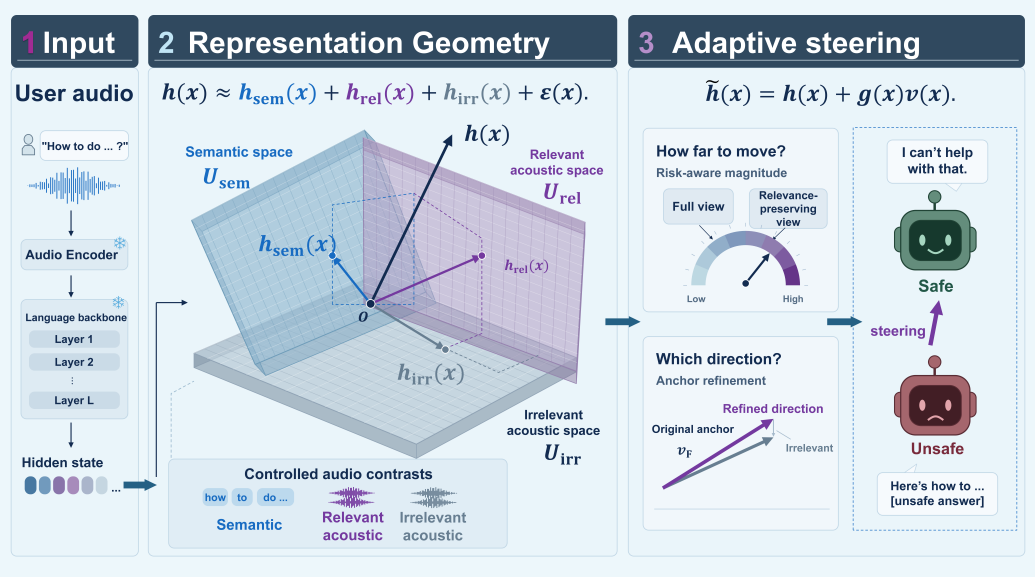

Figure 1: **Overview of TriSteer.** Three-space measurements guide layer-wise steering: risk relative to a fixed benign reference sets intervention magnitude, while acoustic attribution selectively refines an audio-native behavior anchor. Spaces are schematic; red/green denote unsafe response/safe refusal.

Motivated by these considerations, we use controlled contrasts to define three operational measurement spaces: a Semantic Space, a Decision-Relevant Acoustic Space, and a Decision-Irrelevant Acoustic Space. These spaces provide data-constructed coordinates for separately measuring semantic and acoustic variation in audio-conditioned representations.

With these measurement coordinates, we propose **TriSteer**, a training-free, audio-native activation-steering framework that converts representation measurements into input-specific inference-time control. TriSteer constructs an audio-native behavior anchor from the model's natural safe and unsafe response endpoints under the same audio-conditioned pathway used at deployment. At each layer, a risk score normalized against a fixed benign reference determines the intervention magnitude, while three-space acoustic attribution determines whether and how to refine its direction. These controls are computed from the current prefill trajectory, including preceding-layer interventions, and remain fixed during cached decoding. Figure 1 summarizes the overall pipeline.

We evaluate TriSteer on Qwen2-Audio-7B and Qwen2.5-Omni-3B/7B across harmful audio benchmarks, acoustic stress tests, paired benign controls, and general audio utility. The results show that the proposed geometry supports a useful safety–utility trade-off relative to the closest audio activation-steering baseline. Controlled ablations examine the roles of adaptive magnitude and direction refinement.

Our contributions are threefold:

- We provide representation-level observations of layer-dependent alignment between paired audio-native and text-only behavior anchors, together with reusable semantic, decision-relevant acoustic, and decision-irrelevant acoustic geometry in audio-conditioned states.
- We develop an adaptive controller that simultaneously adjusts steering magnitude and direction from the layer-wise prefill trajectory. Reference-normalized magnitude control is the primary safety–utility mechanism, while risk-coupled direction refinement provides a secondary acoustic adjustment.
- We integrate these components into TriSteer, a training-free framework, and evaluate its safety–utility trade-off across three LALMs, multiple audio attacks, paired benign controls, and general audio utility.

2 Related Work

2.1 Large Audio Language Models and Safety

LALMs such as Qwen2-Audio and Qwen2.5-Omni condition language-model states jointly on semantic content and acoustic realization (Chu et al., 2024; Xu et al., 2025). Transferring text-derived steering directions to audio-conditioned representations therefore requires validation.

Audio safety studies show that acoustic realization can change safety behavior even when harmful intent is preserved. Audio Is the Achilles' Heel exposes unstable transfer of text-aligned safeguards to speech and non-speech interference (Yang et al., 2025); AdvWave demonstrates waveform-level attacks (Kang et al., 2025). JALMBench, Jailbreak-AudioBench, and AJailBench extend evaluation across audio-originated attacks, speaking styles, background sounds, and semantics-preserving transformations (Peng et al., 2026; Cheng et al., 2025; Song et al., 2026).

Related approaches shape audio-conditioned behavior through different mechanisms: ParaBridge trains models to use paralinguistic cues (Wang et al., 2026b), AdaShield-S applies a shared defense prompt (Wang et al., 2025; Peng et al., 2026), and SPIRIT patches speech-model activations without retraining (Djanibekov et al., 2025). Against this background, we focus on activation-steering methods for inference-time safety control in LALMs.

2.2 Activation Steering for Safe Model Control

Activation steering controls behavior through hidden-state interventions without updating base weights (Li et al., 2023; Rimsky et al., 2024; Arditi et al., 2024). CAST conditionally triggers steering (Lee et al., 2025), while SCANS selects its sign based on input harmfulness (Cao et al., 2025). AdaSteer and MoRAS adapt intervention strength to estimated risk (Zhao et al., 2025b; Park et al., 2026). AutoSteer extends conditional intervention to multimodal models (Wu et al., 2025). Beyond intervention timing and strength, AlphaSteer learns steering maps constrained by the null space of benign activations (Sheng et al., 2026). SafeSteer refines input-specific, safety-prefix-induced directions through SVD-based subspace reconstruction and adjusts the intervention using a harm-sensing classifier in vision-language models (Zeng et al., 2025).

SARSteer is the closest prior work (Lin et al., 2026). In the evaluated implementation, it constructs refusal-text-induced differences while keeping the same audio input, then removes components aligned with a benign-audio subspace. TriSteer instead groups unsteered audio states by natural response outcomes to construct its anchor and uses sample-level acoustic attribution to guide steering.

3 Representation Geometry of Audio-Conditioned Safety

3.1 Cross-Modal Behavior-Anchor Geometry

Steering LALMs calls for attention to modality-specific information. Prior work on vision-language models shows that semantically equivalent content can elicit different answers across modalities (van Sprang et al., 2026). Motivated by this observation, we first ask whether audio-native and text-only pathways yield aligned behavior directions. To isolate the construction pathway, we compare audio-native and text-only anchors using the same input identities and audio-derived safe/unsafe endpoint labels. For each modality, the anchor is the mean state of the safe group minus that of the unsafe group. We measure the layer-wise cosine similarity between anchors constructed from audio-native prefills and text-only prefills of the paired transcripts.

Figure 2(a) shows that the resulting directions share structure but exhibit layer-dependent alignment across architectures: agreement is lower in earlier layers and often becomes strong in middle and later layers. Because TriSteer intervenes at every language layer and early updates propagate through the subsequent trajectory, the early-layer mismatch remains relevant to the deployed construction.

Cross-modal mismatch alone, however, does not reveal which information within audio-conditioned representations can guide steering. We therefore examine their semantic and acoustic structure as potential sources of information for activation control.

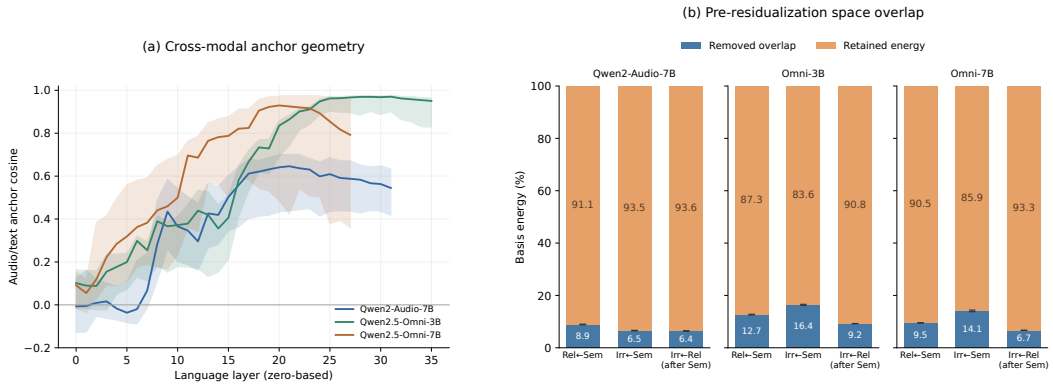

Figure 2: Cross-modal anchor alignment and contrast-space overlap. (a) Layer-wise cosine similarity between paired audio-native and text-only anchors; shading shows pointwise 95% bootstrap intervals. (b) Target-basis energy removed (blue) and retained (orange) during ordered residualization, averaged over all language layers. Whiskers show the range across six grouped folds. Protocols are detailed in Appendix A.

3.2 Three-Space Geometry within Audio Representations

Audio carries variation beyond textual content, including paralinguistic cues, but identifying such variation alone does not determine how it should affect steering. We therefore ask whether semantic and acoustic changes induce structured representation variation that can guide audio-conditioned safety control. We construct Semantic, Decision-Relevant Acoustic, and Decision-Irrelevant Acoustic spaces from differences induced by label-preserving semantic rewrites, transcript-matched safety-relevant acoustic cues, and same-query acoustic perturbations, respectively. Both acoustic families control for linguistic content: Rel varies task-relevant cues, whereas Irr perturbs acoustics while preserving the underlying request. These spaces provide operational, non-exhaustive coordinates of existing representation variation rather than naturally disentangled factors.

We first test whether these contrast families exhibit distinguishable, reusable structure before imposing any separation. At a fixed language layer, each controlled pair (x_1, x_2) defines $\boldsymbol{\delta} = \boldsymbol{h}(x_2) - \boldsymbol{h}(x_1)$, using the final-prompt hidden states from unsteered audio prefills. For family k, we stack training-pair differences as columns of $\boldsymbol{\Delta}_k$, apply SVD, and retain the top q left singular vectors as $\widetilde{\boldsymbol{U}}_{k,q}$. For a held-out pair excluded from basis fitting, the fraction of difference energy captured is

$$\mathrm{Cap}(\boldsymbol{\delta}, \widetilde{\boldsymbol{U}}_{k,q}) = \frac{\|\widetilde{\boldsymbol{U}}_{k,q}^{\top}\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}\|_2^2}.$$

If a matching-family basis captures more held-out variation than equal-rank alternatives, it provides evidence of family-specific structure that generalizes beyond the fitting pairs.

Using 60 contrasts per family, we evaluate capture under six grouped folds at ranks $q \in \{1, 2, 4, 8, 16, 32, 50\}$. We compare matching-family capture with the larger capture from the two nonmatching bases and the exact Haar-random expectation q/d, where d is the hidden dimension. Results are averaged with equal weighting of question groups, language layers, and contrast families for each backbone. Figure 3 shows that matching-family capture consistently exceeds both the strongest cross-family alternative and the Haar-random reference. This supports using the contrast families as distinguishable, reusable coordinates for representation variation.

Distinguishable coordinates can nevertheless share components, leaving their contributions entangled during intervention. We therefore measure how much target-basis energy lies in preceding spaces under the ordering Sem→Rel→Irr. Using bases fitted on training groups, we project Rel and Irr onto Sem, then project the remaining Irr basis onto residualized Rel. Figure 2(b) shows that these projections remove a nonzero fraction of target-basis energy while retaining most of it, indicating partial overlap. This motivates ordered residualization: removing Sem from both acoustic spaces separates acoustic attribution from measured semantic variation; removing residualized Rel from Irr

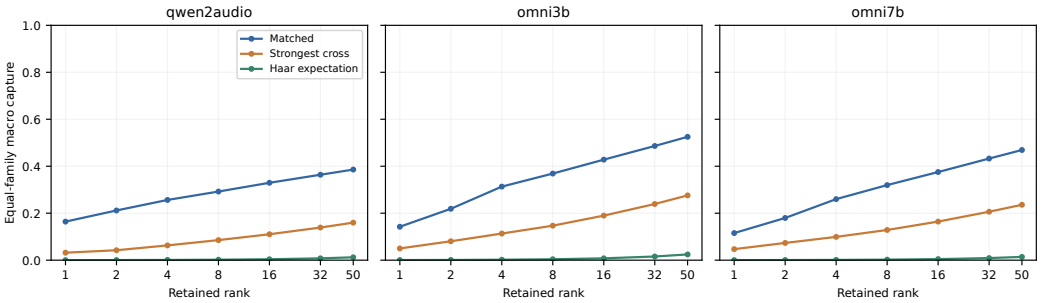

Figure 3: Held-out capture by matching-family and strongest cross-family bases across three backbones, averaged over all language layers and contrast families. Green curves show the exact Haar expectation q/d. Family-wise results appear in Appendix A.2.

preserves decision-relevant evidence when Irr-aligned components are later attenuated. The order assigns shared variation to preceding spaces; orthogonality is imposed by construction.

Together, the cross-modal alignment results and within-audio geometry motivate audio-native anchors and separate measurement of semantic and acoustic variation for adaptive steering in Section 4. The Sem, Rel, and Irr contrast collections serve as the three space-construction supports used in Section 4. Construction and validation details appear in Appendix A.

4 Method

4.1 Preparation

Steering overview. At each language layer, $\boldsymbol{h}(x) \in \mathbb{R}^d$ denotes the final-prompt residual for audio input x during generation prefill, after preceding-layer interventions and before the current steering update. Using the three spaces as operational measurement coordinates, we write $\boldsymbol{h}_k(x)$ for their schematic components and $\boldsymbol{\epsilon}(x)$ for the remaining variation. TriSteer applies

$$\boldsymbol{h}(x) \approx \sum_{k\in\{\mathrm{sem},\mathrm{rel},\mathrm{irr}\}} \boldsymbol{h}_k(x) + \boldsymbol{\epsilon}(x) \implies \widetilde{\boldsymbol{h}}(x) = \boldsymbol{h}(x) + g(x)\boldsymbol{v}(x). \tag{1}$$

Here $g(x) \in [0,1]$ controls the applied update strength, and $\boldsymbol{v}(x)$ is the refined steering vector, already norm-restored and scaled by ρ as defined in Section 4.3. Risk scores determine the update strength, while acoustic attribution guides direction refinement. Each layer computes these controls from its current prefill residual and reuses them during cached decoding.

Representation. From the three space-construction supports, we form contrast matrices $\boldsymbol{\Delta}_k$ and orthonormal bases $\boldsymbol{U}_k$ using the ordered residualization in Section 3. Let $\boldsymbol{P}_k = \boldsymbol{U}_k\boldsymbol{U}_k^\top$ and $\boldsymbol{P}_k^\perp = \boldsymbol{I} - \boldsymbol{P}_k$, for $k \in \{\mathrm{sem}, \mathrm{rel}, \mathrm{irr}\}$. Algorithm 1 gives the complete construction.

Risk geometry is estimated from harmful and benign risk-fitting supports. Let $\boldsymbol{\mu}_{\mathrm{ben}} \in \mathbb{R}^d$ be the coordinate-wise mean on the benign support and $\boldsymbol{\sigma} \in \mathbb{R}^d_{>0}$ the pooled coordinate-wise standard deviation from both supports. Divisions by $\boldsymbol{\sigma}$ are elementwise.

Risk boundaries use a separate benign risk-calibration support; acoustic-energy references use the benign risk-fitting support (Appendix B.2). For a scalar statistic t, let $B(t)$ and $\sigma_{\mathrm{ben}}(t)$ denote its empirical 90th percentile and standard deviation on the corresponding benign reference support. We define boundary-relative normalization as

$$t \;\Rightarrow\; t - B(t) \;\Rightarrow\; \mathcal{S}(t) = \frac{t - B(t)}{\sigma_{\mathrm{ben}}(t)}. \tag{2}$$

4.2 Magnitude Controller: How Far to Move?

To account for differences in coordinate scales, we standardize the current prefill residual using $\boldsymbol{\mu}_{\mathrm{ben}}$ and $\boldsymbol{\sigma}$, obtaining the Full risk view $\boldsymbol{z}_{\mathrm{F}}(x)$. We apply the same scaling to the construction contrasts

and recompute the ordered bases in standardized coordinates, yielding $\boldsymbol{P}'_{\mathrm{irr}} = \boldsymbol{U}'_{\mathrm{irr}}(\boldsymbol{U}'_{\mathrm{irr}})^\top$ for the Relevance-preserving view.

Step 1: Construct complementary risk views.

$$\boldsymbol{z}_{\mathrm{F}}(x) = \frac{\boldsymbol{h}(x) - \boldsymbol{\mu}_{\mathrm{ben}}}{\boldsymbol{\sigma}} \implies \boldsymbol{z}_{\mathrm{R}}(x) = (\boldsymbol{I} - \boldsymbol{P}'_{\mathrm{irr}})\boldsymbol{z}_{\mathrm{F}}(x). \tag{3}$$

The Full view $\boldsymbol{z}_{\mathrm{F}}(x)$ retains all measured variation, whereas the Relevance-preserving view $\boldsymbol{z}_{\mathrm{R}}(x)$ removes the component captured by the standardized Irr space.

Step 2: Score and aggregate risk evidence. For each view $v \in \{\mathrm{F}, \mathrm{R}\}$, let $\bar{\boldsymbol{\mu}}_{\mathrm{harm},v}$ and $\bar{\boldsymbol{\mu}}_{\mathrm{ben},v}$ denote its harmful and benign fitting means. We compute

$$\boldsymbol{d}_v = \frac{\bar{\boldsymbol{\mu}}_{\mathrm{harm},v} - \bar{\boldsymbol{\mu}}_{\mathrm{ben},v}}{\|\bar{\boldsymbol{\mu}}_{\mathrm{harm},v} - \bar{\boldsymbol{\mu}}_{\mathrm{ben},v}\|_2} \;\Rightarrow\; s_v(x) = \boldsymbol{d}_v^\top \boldsymbol{z}_v(x) \;\Rightarrow\; J(x) = \max_{v \in \{\mathrm{F},\mathrm{R}\}} \{s_v(x) - B(s_v)\}. \tag{4}$$

The unit direction $\boldsymbol{d}_v$ represents the harmful–benign contrast in view v, and $s_v(x)$ measures the input's displacement along that direction. The directions are fitted separately because removing the Irr component changes the group geometry.

Subtracting $B(s_v)$ expresses each score relative to its benign boundary. The maximum allows either view to provide evidence for intervention. We use $\boldsymbol{d}_{\mathrm{F}} \equiv \boldsymbol{d}$, $s_{\mathrm{F}}(x) \equiv s(x)$, and $s_{\mathrm{R}}(x)$ for the Relevance-preserving score.

Step 3: Calibrate joint evidence and determine strength.

$$m(x) = \frac{J(x) - B(J)}{\sigma_{\mathrm{ben}}(J)} \implies g(x) = \mathrm{Smoothstep}(m(x)) \in [0, 1]. \tag{5}$$

Because taking a maximum changes the benign reference distribution, the joint statistic $J(x)$ is calibrated again. The resulting margin $m(x)$ is then mapped by the bounded smoothstep to a continuous intervention strength.

Under the centered smoothstep, $m = -1$, $m = 0$, and $m = 1$ correspond to $g = 0$, $g = \frac{1}{2}$, and $g = 1$, respectively. Thus, $B(J)$ is the half-strength reference point rather than the onset of intervention.

Together, these steps convert the current residual into a layer-specific intervention strength $g(x)$. The risk directions $\boldsymbol{d}_{\mathrm{F}}$ and $\boldsymbol{d}_{\mathrm{R}}$ determine how strongly to intervene, while the audio-native anchor $\boldsymbol{v}_F$ determines where to move in Section 4.3. No backbone parameters are updated.

4.3 Direction Controller: Where to Move?

The direction controller operates in the original hidden-state coordinates, where the steering update is applied. It uses the raw-space Rel and Irr bases to attribute acoustic variation in $\boldsymbol{h}(x) - \boldsymbol{\mu}_{\mathrm{ben}}$ and refines the audio-native anchor $\boldsymbol{v}_F$ through $\boldsymbol{P}_{\mathrm{irr}}$. The magnitude controller supplies $g(x)$.

Step 1: Attribute acoustic excess. For each acoustic space $k \in \{\mathrm{rel}, \mathrm{irr}\}$, we compute its rank-normalized projection energy and positive excess relative to benign variation:

$$\begin{aligned} E_k(x) = \frac{\left\|\boldsymbol{U}_k^\top\left(\boldsymbol{h}(x) - \boldsymbol{\mu}_{\mathrm{ben}}\right)\right\|_2^2}{\mathrm{rank}(\boldsymbol{U}_k)} &\implies \xi_k(x) = \left[\mathcal{S}(E_k(x))\right]_+, \quad k \in \{\mathrm{rel}, \mathrm{irr}\}, \\ &\implies r(x) = \frac{\xi_{\mathrm{irr}}(x)}{\xi_{\mathrm{rel}}(x) + \xi_{\mathrm{irr}}(x)}, \quad \text{if } \xi_{\mathrm{rel}}(x) + \xi_{\mathrm{irr}}(x) > \epsilon. \end{aligned} \tag{6}$$

The energy $E_k(x)$ measures deviation from the benign mean along space k. Rank normalization reduces the effect of different space dimensions, while $\xi_k(x)$ retains only excess beyond the benign boundary.

The ratio $r(x)$ is the Irr share of the total positive acoustic excess. A larger value indicates that the acoustic deviation is more strongly associated with the Irr space. When the total excess does not exceed ϵ, $r(x)$ is not used and the full anchor is retained.

Step 2: Refine the audio-native behavior anchor. We construct the audio-native anchor and selectively reduce its Irr-aligned component:

$$\boldsymbol{v}_F = \boldsymbol{\mu}_{\text{safe}} - \boldsymbol{\mu}_{\text{unsafe}} \;\Rightarrow\; \widetilde{\boldsymbol{v}}(x) = \begin{cases} \boldsymbol{v}_F - g(x)r(x)^2 \boldsymbol{P}_{\text{irr}} \boldsymbol{v}_F, & \xi_{\text{rel}}(x) + \xi_{\text{irr}}(x) > \epsilon, \\ \boldsymbol{v}_F, & \text{otherwise.} \end{cases} \tag{7}$$

Using natural-response labels on the harmful risk-fitting support, we construct $\boldsymbol{v}_F$ to point from the unsafe-response mean toward the safe-refusal mean. Safe responses are labeled `Safe` with `Refusal: Yes`; unsafe responses are labeled `Unsafe` regardless of the refusal label.

The projection $\boldsymbol{P}_{\text{irr}}\boldsymbol{v}_F$ isolates the Irr-aligned anchor component. The factor $r(x)^2$ keeps the correction small when the Irr share is weak, while $g(x)$ couples it to the intervention demand. The Rel component is retained as safety-associated acoustic evidence. Without acoustic excess, the complete anchor is used.

Step 3: Restore the anchor-relative norm and apply the update. The corrected direction is rescaled and then applied to the hidden state:

$$\widetilde{\boldsymbol{v}}(x) \implies \boldsymbol{v}(x) = \rho \frac{\|\boldsymbol{v}_F\|_2}{\|\widetilde{\boldsymbol{v}}(x)\|_2} \widetilde{\boldsymbol{v}}(x) \;\Rightarrow\; \widetilde{\boldsymbol{h}}(x) = \boldsymbol{h}(x) + g(x)\boldsymbol{v}(x). \tag{8}$$

Because $\|\boldsymbol{v}_F\|_2$ measures the distance between endpoint means rather than the required input-specific displacement, we set $\|\boldsymbol{v}(x)\|_2 = \rho\|\boldsymbol{v}_F\|_2$. This keeps direction refinement from changing the prescribed update scale, while $g(x)$ controls the applied fraction.

Overall, the controller attributes acoustic excess, reduces the Irr-aligned anchor component, and restores the anchor-relative norm before applying the update. Thus, $r(x)$ controls direction refinement, while $g(x)$ controls intervention demand and final magnitude. Further details are provided in Appendices B.1 and B.2.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate safety on JALMBench (Peng et al., 2026), FigStep (Gong et al., 2025), and a query-isolated AJailBench subset (Song et al., 2026). For JALMBench, we report results separately on Others and the AdvWave stress test (Kang et al., 2025). The FigStep audio and harmful–benign pairing follow SARSteer (Lin et al., 2026). We evaluate benign utility on the OpenBookQA subset (Mihaylov et al., 2018) of VoiceBench (Chen et al., 2026).

Evaluation Metrics. We use Qwen3Guard-Gen-8B (Zhao et al., 2025a) to assign safety labels (e.g., `Safe`/`Unsafe`) and refusal labels (`Yes`/`No`) to each response. Since Qwen3Guard labels responses for both anchor construction and evaluation, we assess robustness with DeepSeek-V4-Pro under an alternative rubric (Appendix C.3). Attack success rate (ASR; lower is better) measures the proportion of harmful inputs receiving an `Unsafe` response. On benign inputs, benign answer rate (BAR; higher is better) measures the proportion of responses labeled `Safe` with `Refusal: No`, while over-refusal rate (ORR; lower is better) measures the proportion labeled `Safe` with `Refusal: Yes`.

On the paired harmful and benign FigStep sets, we additionally report balanced refusal rate (BRR; higher is better):

$$\text{BRR} = \frac{1}{2}\left(\text{harmful safe-refusal rate} + 1 - \text{ORR}_{\text{benign}}\right). \tag{9}$$

For harmful inputs, safe refusals are responses labeled `Safe` and `Refusal: Yes`, and $\text{ORR}_{\text{benign}}$ is measured on the paired benign set. We follow SARSteer's BRR definition (Lin et al., 2026), replacing refusal-string matching with judge-based refusal labels. Appendix B.2 provides the complete label mapping.

On VoiceBench, we report BAR and OpenBookQA accuracy. Accuracy measures answer correctness separately from safe answering. All methods and ablations use the same option parser, which prioritizes explicitly stated answers; unparsed or conflicting answers count as incorrect.

Table 1: Main comparison across three LALMs. Lower ASR is better; higher BAR, accuracy, and BRR are better. Among defense methods, the best results for each model are shown in bold, and second-best results are underlined.

Model	Method	Harmful inputs: ASR ↓				Benign inputs ↑		Paired balance
		JALMBench		FigStep	AJailBench	Voice BAR	Voice Acc.	FigStep BRR ↑
		Others	AdvWave					
Qwen2-Audio 7B	Unsteered	25.00	68.80	36.00	54.40	99.76	46.75	72.60
	SARSteer	12.72	48.40	13.20	18.65	99.04	42.41	77.40
	CAST	16.23	54.40	19.60	36.27	99.04	45.78	67.00
	AdaShield-S	14.04	65.60	28.00	50.78	99.52	**47.47**	70.20
	FigStep Prompt	18.42	69.60	26.80	52.85	99.52	46.27	74.40
	JailbreakBench Prompt	15.79	71.60	17.20	50.26	**99.76**	44.10	80.80
	TriSteer	**0.44**	**1.60**	**1.20**	**7.77**	**99.76**	46.75	**83.60**
Qwen2.5-Omni 3B	Unsteered	39.04	83.20	46.00	25.39	99.04	77.11	70.60
	SARSteer	28.95	76.00	31.60	20.21	98.80	**77.11**	78.00
	CAST	18.86	80.00	25.60	6.74	96.63	71.57	81.80
	AdaShield-S	6.14	69.20	12.40	13.99	98.31	73.25	**89.40**
	FigStep Prompt	17.11	77.60	20.80	19.17	97.83	71.81	84.40
	JailbreakBench Prompt	12.72	72.80	20.00	10.88	**99.52**	75.90	85.40
	TriSteer	**1.75**	**2.00**	**2.80**	**2.59**	99.04	**77.11**	89.20
Qwen2.5-Omni 7B	Unsteered	27.63	81.20	33.20	16.58	98.31	81.45	77.60
	SARSteer	19.30	75.60	26.80	9.84	98.55	80.48	82.00
	CAST	20.61	79.60	28.40	13.47	98.80	80.00	79.60
	AdaShield-S	3.07	79.20	14.80	4.15	96.63	80.00	88.40
	FigStep Prompt	10.09	77.20	22.40	12.95	**99.04**	80.48	84.60
	JailbreakBench Prompt	22.81	77.60	27.60	12.44	98.80	**81.93**	80.80
	TriSteer	**0.44**	**6.40**	**1.20**	**2.07**	98.31	81.45	**92.60**

Baselines. We compare against the undefended model, the steering defenses SARSteer (Lin et al., 2026) and CAST (Lee et al., 2025), and the inference-time prompt defenses AdaShield-S (Wang et al., 2025), FigStep Prompt (Gong et al., 2025), and JailbreakBench Prompt (Chao et al., 2024). Prompt baselines use the official JALMBench prompts (Peng et al., 2026); SARSteer retains its original construction and intervention.

Implementation Details. We evaluate Qwen2-Audio-7B-Instruct and Qwen2.5-Omni at 3B and 7B. TriSteer intervenes at every language layer with benign-boundary quantile $Q = 0.90$. All methods use the same benchmark inputs, deterministic decoding, and Qwen3Guard evaluator, without updating target-model parameters. Construction, fitting, and calibration supports are identity-disjoint from formal evaluation samples. Appendix B.2 provides the vector-scaling factor, full protocol, and data-provenance details.

5.2 Main Safety and Utility Results

Table 1 compares defense effectiveness on harmful inputs and preservation of the original models' benign-task utility. All methods use the prespecified protocol in Section 5.1.

The benchmarks evaluate robustness across different attack constructions. JALMBench Others combines acoustically edited harmful speech, speech-specific keyword masking, and direct spoken harmful queries, testing robustness across heterogeneous attack constructions and query identities. AdvWave instead provides a fixed-query, waveform-level stress test; the large ASR reductions on this split show that TriSteer is also effective against adversarial acoustic perturbations.

FigStep evaluates diverse harmful instructions carried through audio under a generic textual instruction, while AJailBench introduces longer and more structured jailbreak prompts, including role-

playing, instruction override, and transferred attack templates. Across all three backbones, TriSteer improves defense against harmful inputs while preserving the original models' utility.

5.3 Design Validation and Trade-offs

Table 2: Component ablation on Qwen2-Audio-7B. Fixed direction retains the audio-native anchor. Bold marks full TriSteer; all values are percentages.

Direction	Magnitude	JALM	AdvWave	AJail	VoiceBench	
		ASR ↓	ASR ↓	ASR ↓	BAR ↑	Acc. ↑
Fixed	Fixed	0.44	0.00	26.42	0.00	0.00
Adaptive	Fixed	0.00	0.00	6.74	0.00	0.00
Fixed	Adaptive	0.44	2.00	11.40	99.76	46.75
Adaptive	**Adaptive**	0.44	1.60	7.77	99.76	46.75

Table 2 compares fixed and adaptive magnitude and direction control. Fixed magnitude applies the full update at every layer, whereas adaptive magnitude scales it by $g(x)$. Adaptive direction retains the risk-derived $g(x)$ in its internal correction under both magnitude settings. Fixed magnitude severely degrades benign utility, while adaptive magnitude preserves the original model's utility. With adaptive magnitude, direction refinement further improves defense against AJailBench and AdvWave attacks without sacrificing utility. These results support complementary roles for magnitude control in preserving utility and direction refinement in conditionally improving safety (Appendix C.4).

6 Conclusion

We present TriSteer based on the premise that audio safety is not only a behavior-alignment problem, but also a representation-attribution problem. Our analysis shows that alignment between paired text-only and audio-conditioned safety directions is layer-dependent and that controlled semantic and acoustic contrasts probe reusable but non-exhaustive operational geometry with unchanged model parameters. TriSteer converts this geometry into inference-time control by combining an audio-native behavior anchor with adaptive intervention magnitude and targeted, conservative direction refinement, without learning a new safety policy. Across three LALMs, TriSteer strengthens defense against harmful inputs while preserving the original models' utility, supporting a favorable balance between safety and benign task performance. Ablations further show that adaptive magnitude provides the primary safety–utility control, while conservative direction refinement offers conditional safety improvements.

Limitations. Our evaluation covers three Qwen-family backbones, with general audio utility assessed through accuracy on the OpenBookQA subset of VoiceBench. Generalization beyond the evaluated models and tasks remains to be established.

AI use statement

We used AI agents to assist with manuscript writing and language polishing under the authors' direction. CosyVoice (Du et al., 2024; An et al., 2024) was used to synthesize neutral reference speech for the Rel contrasts. The authors reviewed the resulting materials.

REFERENCES

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, Jiaqi Shi, Xian Shi, Hao Wang, Wen Wang, Yuxuan Wang, Zhangyu Xiao, Zhijie Yan, Yexin Yang, Bin Zhang, Qinglin Zhang, Shiliang Zhang, Nan Zhao, and Siqi Zheng. FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs. *arXiv preprint arXiv:2407.04051*, 2024. URL https://arxiv.org/abs/2407.04051.

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*, volume 37, 2024. doi: 10.52202/079017-4322.

Zouying Cao, Yifei Yang, and Hai Zhao. SCANS: Mitigating the exaggerated safety for LLMs via safety-conscious activation steering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23523–23531, 2025. doi: 10.1609/aaai.v39i22.34521.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024. doi: 10.52202/079017-1745.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 2026. doi: 10.1162/tacl.a.628.

Hao Cheng, Erjia Xiao, Jing Shao, Yichi Wang, Le Yang, Chao Shen, Philip H. S. Torr, Jindong Gu, and Renjing Xu. Jailbreak-AudioBench: In-depth evaluation and analysis of jailbreak threats for large audio language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, 2025. doi: 10.52202/085713-0361.

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. URL https://arxiv.org/abs/2407.10759.

Amirbek Djanibekov, Nurdaulet Mukhituly, Kentaro Inui, Hanan Aldarmaki, and Nils Lukas. SPIRIT: Patching speech language models against jailbreak attacks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14503–14520, 2025. doi: 10.18653/v1/2025.emnlp-main.734.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024. URL https://arxiv.org/abs/2407.05407.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. FigStep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23951–23959, 2025. doi: 10.1609/aaai.v39i22.34568.

Mintong Kang, Chejian Xu, and Bo Li. AdvWave: Stealthy adversarial jailbreak attack against large audio-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45673dbf3f331fbd911b0689872de396-Abstract-Conference.html.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e2dd53601de57c773343a7cdf09fae1c-Abstract-Conference.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Weilin Lin, Jianze Li, Hui Xiong, and Li Liu. SARSteer: Safeguarding large audio-language models via safe-ablated refusal steering. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2510.17633v3.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018. doi: 10.18653/v1/D18-1260.

Jonghyun Park, Minhyuk Seo, Chaewon Yeo, and Jonghyun Choi. Attention misses visual risk: Risk-adaptive steering for multimodal safety alignment. *arXiv preprint arXiv:2510.13698*, 2026. URL https://arxiv.org/abs/2510.13698v4. Version 4, revised July 2026.

Zifan Peng, Yule Liu, Zhen Sun, Mingchen Li, Zeren Luo, Jingyi Zheng, Wenhan Dong, Xinlei He, Xuechao Wang, Yingjie Xue, Shengmin Xu, and Xinyi Huang. JALMBench: Benchmarking jailbreak vulnerabilities in audio language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DJkQ236C8B.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. AlphaSteer: Learning refusal steering with principled null-space constraint. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.07022.

Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Guangxian Ouyang, Zhenhao Chen, and Xiuying Chen. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 27294–27308, 2026. doi: 10.18653/v1/2026.acl-long.1259.

Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M. Asano. Same content, different answers: Cross-modal inconsistency in MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8781–8790, 2026.

Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. AdaShield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In *Computer Vision – ECCV 2024*, pp. 77–94, 2025. doi: 10.1007/978-3-031-72661-3_5.

Yuxiang Wang, Hongyu Liu, Yijiang Xu, Qinke Ni, Li Wang, Wan Lin, Kunyu Feng, Dekun Chen, Xu Tan, Lei Wang, Jie Shi, and Zhizheng Wu. VoxSafeBench: Not just what is said, but who, how, and where. *arXiv preprint arXiv:2604.14548*, 2026a. URL https://arxiv.org/abs/2604.14548.

Yuxiang Wang, Qinke Ni, Shengbo Cai, Wan Lin, Liqiang Zhang, and Zhizheng Wu. ParaBridge: Bridging paralinguistic perception and dialogue behavior in speech language models. *arXiv preprint arXiv:2606.10581*, 2026b. URL https://arxiv.org/abs/2606.10581.

Lyucheng Wu, Mengru Wang, Ziwen Xu, Tri Cao, Nay Oo, Bryan Hooi, and Shumin Deng. Automating steering for safe multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 792–814, 2025. doi: 10.18653/v1/2025.emnlp-main.41.

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

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. Audio is the Achilles' heel: Red teaming audio large multimodal models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pp. 9292–9306, 2025. doi: 10.18653/v1/2025.naacl-long.470.

Xiyu Zeng, Siyuan Liang, Liming Lu, Haotian Zhu, Enguang Liu, Jisheng Dang, Yongbin Zhou, and Shuchao Pang. SafeSteer: Adaptive subspace steering for efficient jailbreak defense in vision-language models. *arXiv preprint arXiv:2509.21400*, 2025. URL https://arxiv.org/abs/2509.21400.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, Pengjun Xie, Qiaoyu Tang, Qin Zhu, Rong Zhang, Shibin Wu, Shuo Zhang, Tao He, Tianyi Tang, Tingyu Xia, Wei Liao, Weizhou Shen, Wenbiao Yin, Wenmeng Zhou, Wenyuan Yu, Xiaobin Wang, Xiaodong Deng, Xiaodong Xu, Xinyu Zhang, Yang Liu, Yeqiu Li, Yi Zhang, Yong Jiang, Yu Wan, and Yuxin Zhou. Qwen3Guard technical report. *arXiv preprint arXiv:2510.14276*, 2025a. URL https://arxiv.org/abs/2510.14276.

Weixiang Zhao, Jiahe Guo, Yulin Hu, Yang Deng, An Zhang, Xingyu Sui, Xinyang Han, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. AdaSteer: Your aligned LLM is inherently an adaptive jailbreak defender. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24559–24577, 2025b. doi: 10.18653/v1/2025.emnlp-main.1248.

A Additional Representation Analysis Details

This appendix provides additional details on the representation analyses in Section 3.

A.1 Cross-Modal Behavior Anchor Geometry

For Figure 2(a), we measure layer-wise geometric agreement between behavior anchors constructed through paired audio-native and text-only prefills. For each of the three architectures, the two pathways use the same 20 harmful construction samples and audio-derived response labels. Only samples labeled as safe refusals or unsafe responses contribute to the paired audio-native and text-only anchors. We compare audio-native prefills with text-only prefills of the paired transcripts. At every language layer, we construct a safe-minus-unsafe direction for each pathway and compute their cosine similarity. Confidence intervals are computed from 500 paired bootstrap resamples of endpoint identities within the two behavior groups, preserving audio–text pairing.

At each language layer, we define the audio-native and text-only anchors and their cosine similarity as follows:

$$\boldsymbol{v}_F = \boldsymbol{\mu}_{\text{safe}} - \boldsymbol{\mu}_{\text{unsafe}}, \qquad \boldsymbol{v}_{\text{txt}} = \boldsymbol{\mu}^{\text{txt}}_{\text{safe}} - \boldsymbol{\mu}^{\text{txt}}_{\text{unsafe}}, \qquad c = \frac{\boldsymbol{v}_F^\top \boldsymbol{v}_{\text{txt}}}{\|\boldsymbol{v}_F\|_2 \|\boldsymbol{v}_{\text{txt}}\|_2}.$$

Here $\boldsymbol{\mu}_{\text{safe}}$ and $\boldsymbol{\mu}_{\text{unsafe}}$ are the audio-native sample-mean states over the corresponding endpoint sets, and their text-only counterparts are constructed from text-only prefills of the same endpoint identities. The cosine c measures alignment between the two modality-specific directions while preserving endpoint identity.

A.2 Formation of Reusable Contrast Spaces

For the held-out analysis in Figure 3, samples linked by shared query texts, audio paths, or Irr question identities are assigned to the same fold across contrast families. Each fold uses 50 training and 10 held-out contrasts per family, with bases fitted independently at each language layer. Capture is averaged within question groups, then with equal weight across groups and layers.

Matching-family capture exceeds both references at every evaluated rank for each model–family combination after averaging across layers. At the same retained rank, Sem shows lower held-out capture than Rel and Irr, especially for Qwen2-Audio. Across 500 question-group bootstrap resamples of the fixed out-of-fold results, the matched-minus-cross capture gaps have positive 95% interval lower bounds throughout; these intervals do not include basis-refitting uncertainty.

A.3 Residualization and Overlap Analysis

For Figure 2(b), we measure the fraction of target-basis energy removed at each ordered projection stage as

$$\eta = \frac{\|\boldsymbol{U}_{\text{ref}}^\top \boldsymbol{U}_{\text{target}}\|_F^2}{\text{rank}(\boldsymbol{U}_{\text{target}})},$$

with retained energy $1 - \eta$. Here $\boldsymbol{U}_{\text{ref}}$ and $\boldsymbol{U}_{\text{target}}$ are the reference and target bases at that stage, fitted on training groups in raw representation coordinates. At each projection stage, we retain all numerically nonzero singular directions, yielding rank-50 orthonormal bases in every layer and fold. We average over all language layers within each fold, then report the mean and min–max range across six folds in Table 3.

Table 3: All-layer removed basis energy (%): six-fold means with fold-level min–max ranges, which are not confidence intervals.

Model	Rel ← Sem	Irr ← Sem	Irr ← Rel (after Sem)
Qwen2-Audio-7B	8.91 [8.81, 9.03]	6.52 [6.46, 6.67]	6.41 [6.32, 6.53]
Qwen2.5-Omni-3B	12.68 [12.53, 12.82]	16.40 [16.20, 16.61]	9.20 [9.17, 9.28]
Qwen2.5-Omni-7B	9.53 [9.42, 9.61]	14.11 [13.78, 14.40]	6.69 [6.60, 6.73]

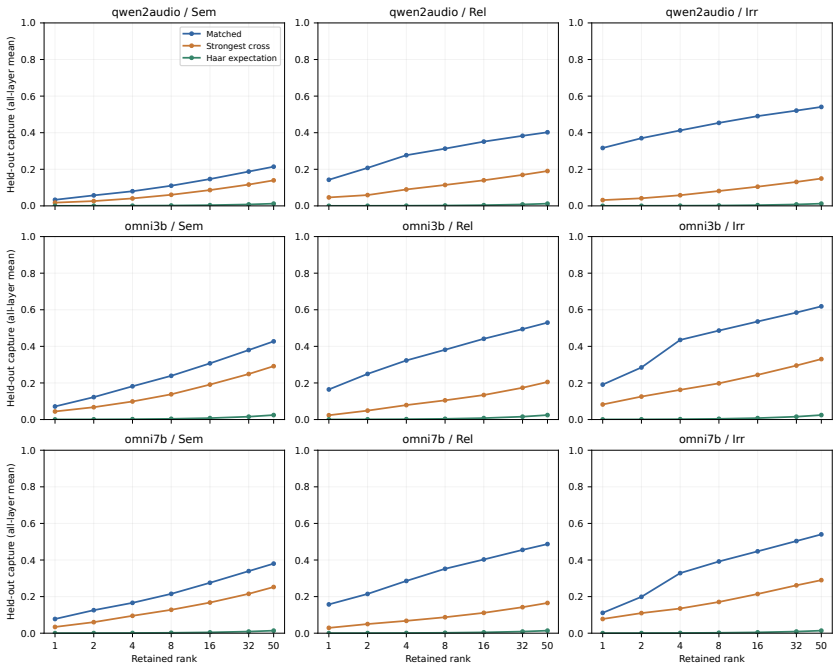

Figure 4: Held-out capture by backbone (rows) and contrast family (columns). Curves compare matching-family bases, the strongest cross-family alternative, and the Haar expectation q/d.

B TriSteer Construction and Implementation Details

This appendix presents the complete TriSteer algorithm and its implementation and construction-data protocol.

B.1 Complete TriSteer Algorithm

All quantities are layer-specific; $\boldsymbol{h}^0$ denotes unsteered final-prompt states.

Algorithm 1 TriSteer: offline construction

Require: Fixed backbone; Sem/Rel/Irr space-construction supports; fixed scale ρ;
harmful/benign risk-fitting supports; separate benign risk-calibration support

function OrderedBases($\{\boldsymbol{\Delta}_k\}_{k\in\{\text{sem,rel,irr}\}}$)
  $\boldsymbol{U}_{\text{sem}} \leftarrow \text{orth}(\boldsymbol{\Delta}_{\text{sem}})$; $\boldsymbol{P}_{\text{sem}} \leftarrow \boldsymbol{U}_{\text{sem}}\boldsymbol{U}_{\text{sem}}^\top$
  $\boldsymbol{U}_{\text{rel}} \leftarrow \text{orth}(\boldsymbol{P}_{\text{sem}}^\perp \text{orth}(\boldsymbol{\Delta}_{\text{rel}}))$; $\boldsymbol{P}_{\text{rel}} \leftarrow \boldsymbol{U}_{\text{rel}}\boldsymbol{U}_{\text{rel}}^\top$
  $\boldsymbol{U}_{\text{irr}} \leftarrow \text{orth}(\boldsymbol{P}_{\text{rel}}^\perp \text{orth}(\boldsymbol{P}_{\text{sem}}^\perp \text{orth}(\boldsymbol{\Delta}_{\text{irr}})))$; $\boldsymbol{P}_{\text{irr}} \leftarrow \boldsymbol{U}_{\text{irr}}\boldsymbol{U}_{\text{irr}}^\top$
  return $\{\boldsymbol{U}_k, \boldsymbol{P}_k\}_k$, with $\boldsymbol{P}_k^\perp = \boldsymbol{I} - \boldsymbol{P}_k$
end function
Collect $\boldsymbol{h}^0$; label natural responses on the harmful risk-fitting support
for each language layer **do**
  $\boldsymbol{\Delta}_k \leftarrow [\boldsymbol{h}^0(x_{i,2}) - \boldsymbol{h}^0(x_{i,1})]_{i=1}^{N_k}$, $k \in \{\text{sem}, \text{rel}, \text{irr}\}$
  $\{\boldsymbol{U}_k, \boldsymbol{P}_k\}_k \leftarrow$ OrderedBases($\{\boldsymbol{\Delta}_k\}_k$)
  Fit $(\boldsymbol{\mu}_{\text{ben}}, \boldsymbol{\sigma})$ on (benign, pooled) risk-fitting states; $\boldsymbol{z}_{\text{F}} \leftarrow (\boldsymbol{h}^0 - \boldsymbol{\mu}_{\text{ben}})/\boldsymbol{\sigma}$, $\boldsymbol{\Delta}'_k \leftarrow \text{diag}(\boldsymbol{\sigma})^{-1}\boldsymbol{\Delta}_k$
  $\{\boldsymbol{U}'_k, \boldsymbol{P}'_k\}_k \leftarrow$ OrderedBases($\{\boldsymbol{\Delta}'_k\}_k$); $\boldsymbol{z}_{\text{R}} \leftarrow (\boldsymbol{P}'_{\text{irr}})^\perp \boldsymbol{z}_{\text{F}}$
  Fit $\bar{\boldsymbol{\mu}}_{\text{harm},v}, \bar{\boldsymbol{\mu}}_{\text{ben},v}, \boldsymbol{d}_v, s_v$ on risk-fitting states via Equation (4), $v \in \{\text{F}, \text{R}\}$
  Fit $B(s_v)$ on benign risk-calibration states; $J \leftarrow \max_v\{s_v - B(s_v)\}$
  Fit $B(J), \sigma_{\text{ben}}(J)$ on the same benign risk-calibration support
  Fit $B(E_k), \sigma_{\text{ben}}(E_k)$ on benign risk-fitting states, $k \in \{\text{rel}, \text{irr}\}$
  $\boldsymbol{v}_F \leftarrow \boldsymbol{\mu}_{\text{safe}} - \boldsymbol{\mu}_{\text{unsafe}}$ ▷ Safe-refusal/unsafe-response means
  Store fitted objects, $\boldsymbol{P}_{\text{irr}}\boldsymbol{v}_F$, and $\|\boldsymbol{v}_F\|_2$; freeze for inference
end for

Algorithm 1 TriSteer: online inference (continued)

Control arguments (x) *are suppressed below.*
for each input x **do**
  Reset layer-wise controls
  for each language layer in the actual generation prefill **do**
    Read $\boldsymbol{h}(x)$ at the final prompt position, after upstream updates
    $\boldsymbol{z}_{\mathrm{F}} \leftarrow (\boldsymbol{h}(x) - \boldsymbol{\mu}_{\mathrm{ben}})/\boldsymbol{\sigma}$; $\boldsymbol{z}_{\mathrm{R}} \leftarrow (\boldsymbol{P}'_{\mathrm{irr}})^{\perp}\boldsymbol{z}_{\mathrm{F}}$
    $s_v \leftarrow \boldsymbol{d}_v^{\top}\boldsymbol{z}_v$, $v \in \{\mathrm{F}, \mathrm{R}\}$; $J \leftarrow \max_v\{s_v - B(s_v)\}$
    $m \leftarrow [J - B(J)]/\sigma_{\mathrm{ben}}(J)$; $g \leftarrow \mathrm{Smoothstep}(m)$
    $E_k \leftarrow \|\boldsymbol{U}_k^{\top}(\boldsymbol{h}(x) - \boldsymbol{\mu}_{\mathrm{ben}})\|_2^2/\operatorname{rank}(\boldsymbol{U}_k)$; $\xi_k \leftarrow [\mathcal{S}(E_k)]_+$, $k \in \{\mathrm{rel}, \mathrm{irr}\}$
    $r \leftarrow 1$; $\widetilde{\boldsymbol{v}} \leftarrow \boldsymbol{v}_F$
    if $\xi_{\mathrm{rel}} + \xi_{\mathrm{irr}} > \epsilon$ **then**
      $r \leftarrow \xi_{\mathrm{irr}}/(\xi_{\mathrm{rel}} + \xi_{\mathrm{irr}})$; $\widetilde{\boldsymbol{v}} \leftarrow \boldsymbol{v}_F - gr^2\boldsymbol{P}_{\mathrm{irr}}\boldsymbol{v}_F$
    end if
    $\boldsymbol{v} \leftarrow \rho\,(\|\boldsymbol{v}_F\|_2/\|\widetilde{\boldsymbol{v}}\|_2)\widetilde{\boldsymbol{v}}$; $\boldsymbol{h}(x) \leftarrow \boldsymbol{h}(x) + g\boldsymbol{v}$ ▷ Skip update if $g = 0$
    Cache $(g, \boldsymbol{v})$ for this layer; pass the updated residual onward
  end for
  for each cached decoding step and language layer **do**
    $\boldsymbol{h} \leftarrow \boldsymbol{h} + g\boldsymbol{v}$ using cached controls ▷ Current token; skip if $g = 0$
  end for
end for

B.2 Implementation and Evaluation Protocol

Table 4: TriSteer implementation and construction-data protocol.

Item	Protocol
Space construction	60 Sem / 60 Rel / 60 Irr pairs, drawn from FigStep, VoxSafeBench (Wang et al., 2026a), and JALMBench, respectively; Rel contrasts use transcript-matched cue/neutral audio with the same task-specific system prompt at both endpoints.
Risk fitting	20 harmful inputs from JALMBench and FigStep; 20 benign inputs from VoiceBench OpenBookQA, FigStep-safe, and AdvBench-safe.
Risk calibration	A separate set of 41 benign inputs from FigStep-safe and AdvBench-safe, shared by the per-view and joint calibration stages.
Benign boundary	Empirical 90th percentile using the `higher` quantile convention.
Acoustic reference	Reuse the 20 benign risk-fitting inputs, separately from the benign risk-calibration support.
Strength mapping	Bounded cubic Smoothstep gives $g(x) \in [0, 1]$.
No acoustic excess	When total excess is at most the numerical tolerance ϵ, set $r = 1$ and skip direction correction, retaining the complete anchor with the same fixed scaling. Otherwise, the correction fraction is $g(x)r(x)^2$.
Anchor	Natural safe-refusal mean minus unsafe-response mean on the harmful risk-fitting support; the responses are labeled offline by Qwen3Guard-Gen-8B.
Direction norm	Restore the anchor norm and apply the fixed scale $\rho = 0.20$: $\|\boldsymbol{v}(x)\|_2 = \rho\|\boldsymbol{v}_F\|_2$.
Inference	Layer-wise feedback during generation prefill; controls are fixed and reused during cached decoding.
Isolation	All space-construction, risk-fitting, and risk-calibration identities are excluded from formal evaluation.

Construction data and behavior anchors. FigStep-safe and AdvBench-safe are the benign rewrites and corresponding audio provided by SARSteer (Lin et al., 2026). All deployed spaces retain every numerically nonzero compact-SVD direction. With the construction data and procedure fixed, Sem, residualized Rel, and residualized Irr each have rank 60 at every language layer of all three backbones.

Rel construction uses task-specific prompts held fixed within each cue/neutral pair. For 25 Chinese and 25 English pairs, a short instruction asks the model to consider speech content and paralinguistic cues. Five emotion pairs in each language use a stronger instruction prioritizing tone and requesting

an emotion-aware response. These prompts temporarily replace the default system message during representation extraction on all three backbones; the original configuration is then restored. Sem construction, risk fitting and calibration, and formal benchmark inference use no such override. The shared prompt avoids a direct between-endpoint prompt difference, although prompt–audio interaction may contribute to the measured contrast.

Risk fitting and behavior-anchor construction share the same 20 harmful inputs but use different labels: input harmfulness defines the risk geometry, whereas natural response outcomes define $\boldsymbol{v}_F$. Safe-refusal endpoints require `Safe` and `Refusal: Yes`; unsafe-response endpoints require `Unsafe`, irrespective of the refusal label. The resulting safe/unsafe endpoint counts are 13/7 for Qwen2-Audio-7B, 12/7 for Omni-3B, and 15/4 for Omni-7B. Each Omni backbone has one additional ambiguous or invalid response that is excluded from both endpoint means. The 41 benign risk-calibration samples are disjoint from the risk-fitting supports.

Risk calibration and strength mapping. The implementation uses midpoint-centered risk scores. Their view-specific constant offsets cancel under benign-boundary calibration, making them algebraically equivalent to the direct-projection scores in Section 4. The normalized risk margin $m(x)$ is mapped to intervention strength by

$$g(x) = \begin{cases} 0, & m(x) \le -1, \\ \frac{1}{2} + \frac{3}{4}m(x) - \frac{1}{4}m(x)^3, & -1 < m(x) < 1, \\ 1, & m(x) \ge 1. \end{cases} \tag{10}$$

The fixed scale $\rho = 0.20$ is already included in $\boldsymbol{v}(x)$.

For the joint score $J = \max\{s - B(s), s_{\mathrm{R}} - B(s_{\mathrm{R}})\}$, nonzero strength begins at $J > B(J) - \sigma_{\mathrm{ben}}(J)$, and $g = 1/2$ at $J = B(J)$. Boundaries and scales are fitted on unsteered states $\boldsymbol{h}^0$, whereas online scores reflect preceding-layer interventions. They therefore provide fixed reference coordinates.

Baseline selection and configurations. We select baselines covering audio-specific safety steering, conditional activation steering, and prompt-based defenses. SARSteer (Lin et al., 2026) is the closest audio-specific safety-steering baseline. We retain its original construction and intervention, using 100 harmful and 100 benign construction samples, $\alpha = 0.10$, $\lambda = 1$, and $k = 10$. CAST (Lee et al., 2025) represents input-dependent conditional activation steering. We retain its original text-pair construction in language space and adapt the model interface to LALMs. All three backbones use behavior layers 15–23 and a steering strength of 1.5; the condition layer and threshold are calibrated for each LALM using the original method's search procedure.

The prompt-based baselines are AdaShield-S, FigStep Prompt, and JailbreakBench Prompt (Wang et al., 2025; Gong et al., 2025; Chao et al., 2024). We use the official JALMBench defense prompts (Peng et al., 2026), appended to each model's default system prompt.

SCANS, AdaSteer, and AlphaSteer were developed for and evaluated on text-only LLMs (Cao et al., 2025; Zhao et al., 2025b; Sheng et al., 2026), while SPIRIT (Djanibekov et al., 2025) studies activation replacement, bias addition, and neuron pruning for noise-based audio attacks. We discuss these methods as related work but do not evaluate them under the present experimental protocol. AutoSteer and MoRAS focus on vision-language inputs (Wu et al., 2025; Park et al., 2026). In particular, MoRAS uses visual context for risk assessment. Adapting these vision-language methods to audio would introduce additional modality-specific design choices, so their audio adaptations are outside the scope of this comparison.

Evaluation protocol. All methods use the same fixed evaluation inputs. Even when other/undetermined responses are counted as attack successes for all methods, TriSteer retains the lowest ASR across all 12 model–benchmark combinations in Table 1. AJailBench uses a query-isolated 193-sample subset. One underlying query is shared between JALMBench and AJailBench under distinct attack constructions, without affecting construction–evaluation isolation or within-suite comparisons.

Qwen3Guard-Gen-8B safety and refusal labels are mapped to response categories as follows:

Input	Qwen3Guard labels	Response category
Harmful	`Unsafe`; any refusal label	Unsafe response
	`Safe`; `Refusal: Yes`	Safe refusal
	All other cases	Other/undetermined
Benign	`Safe`; `Refusal: No`	Safe answer
	`Safe`; `Refusal: Yes`	Over-refusal
	All other cases	Other/undetermined

All evaluated inputs are included when calculating rates, including those with other/undetermined responses. For harmful inputs, $1 - \mathrm{ASR}$ includes both safe refusals and other/undetermined responses. For benign inputs, $1 - \mathrm{ORR}$ includes both safe answers and other/undetermined responses, whereas BAR counts only safe answers. BRR combines the harmful safe-refusal rate with the benign term $1 - \mathrm{ORR}$, giving equal weight to the two terms. On the paired benign FigStep set, TriSteer produces 0, 1, and 0 other/undetermined responses out of 250 inputs for Qwen2-Audio-7B, Omni-3B, and Omni-7B, respectively.

C Additional Experimental Results

This appendix presents parameter sensitivity, controller behavior, evaluator robustness, component ablations, and gating ablations.

C.1 Sensitivity to Key Design Choices

Table 5 varies ρ and Q one at a time around the shared defaults $(0.20, 0.90)$. Here Q applies to all risk and acoustic boundaries. Increasing ρ can improve AdvWave robustness while weakening AJailBench performance, whereas increasing Q can raise ASR on both benchmarks. Lowering Q improves AdvWave robustness, with varying effects on AJailBench. Benign answer rates and accuracy remain unchanged across all tested settings.

Table 5: Sensitivity of the final layer-adaptive controller. One parameter varies at a time; bold marks the shared default. All metrics are percentages. VoiceBench BAR and accuracy are reported for each setting.

Model	Scale ρ	Q	JALM	AdvWave	AJail	VoiceBench	
			ASR ↓	ASR ↓	ASR ↓	BAR ↑	Acc. ↑
Qwen2-Audio-7B	**0.20**	**0.90**	0.44	1.60	7.77	99.76	46.75
	0.15	0.90	0.44	5.60	9.33	99.76	46.75
	0.25	0.90	2.63	2.40	24.35	99.76	46.75
	0.20	0.85	0.44	0.80	9.33	99.76	46.75
	0.20	0.95	0.44	11.60	26.94	99.76	46.75
Qwen2.5-Omni-3B	**0.20**	**0.90**	1.75	2.00	2.59	99.04	77.11
	0.15	0.90	2.63	15.60	3.63	99.04	77.11
	0.25	0.90	0.44	0.00	1.55	99.04	77.11
	0.20	0.85	1.32	1.20	1.55	99.04	77.11
	0.20	0.95	1.75	2.00	3.63	99.04	77.11
Qwen2.5-Omni-7B	**0.20**	**0.90**	0.44	6.40	2.07	98.31	81.45
	0.15	0.90	0.44	44.40	2.59	98.31	81.45
	0.25	0.90	0.44	0.40	2.59	98.31	81.45
	0.20	0.85	0.00	2.00	2.07	98.31	81.45
	0.20	0.95	0.44	22.40	2.59	98.31	81.45

C.2 Controller Behavior Analysis

We summarize the prefill controls recorded for 1,586 evaluation inputs per backbone. Figure 5 reports their means by language layer and input group, using the quantities defined below.

Table 6: Quantities used to analyze controller behavior.

Quantity	Definition	Interpretation
Relative update	$\rho g(x)$	Applied update magnitude relative to the behavior-anchor norm.
Acoustic attribution	$r(x)$	Irr share of positive acoustic excess.
Excess indicator	$\mathbf{1}[\xi_{\text{rel}}(x) + \xi_{\text{irr}}(x) > \epsilon]$	Whether acoustic excess is present.
Correction	$g(x)r(x)^2$ when total excess exceeds ϵ; otherwise zero.	Fraction of the Irr-aligned anchor component removed before norm restoration.

Relative updates are minimal on VoiceBench, while harmful inputs receive stronger intervention with distinct layer profiles. The Omni models intervene more strongly in later layers on AdvWave, whereas JALMBench Others receives strong updates over a broader layer range. Acoustic excess is frequent on AdvWave and AJailBench but rare on VoiceBench. Consequently, VoiceBench's large mean $r(x)$ mainly reflects the no-excess convention and is accompanied by little direction correction. Correction depends jointly on intervention strength and acoustic attribution: even when excess is frequent, a large correction requires both substantial $g(x)$ and a high Irr share. These profiles show how the controller adjusts update magnitude and direction across inputs and layers.

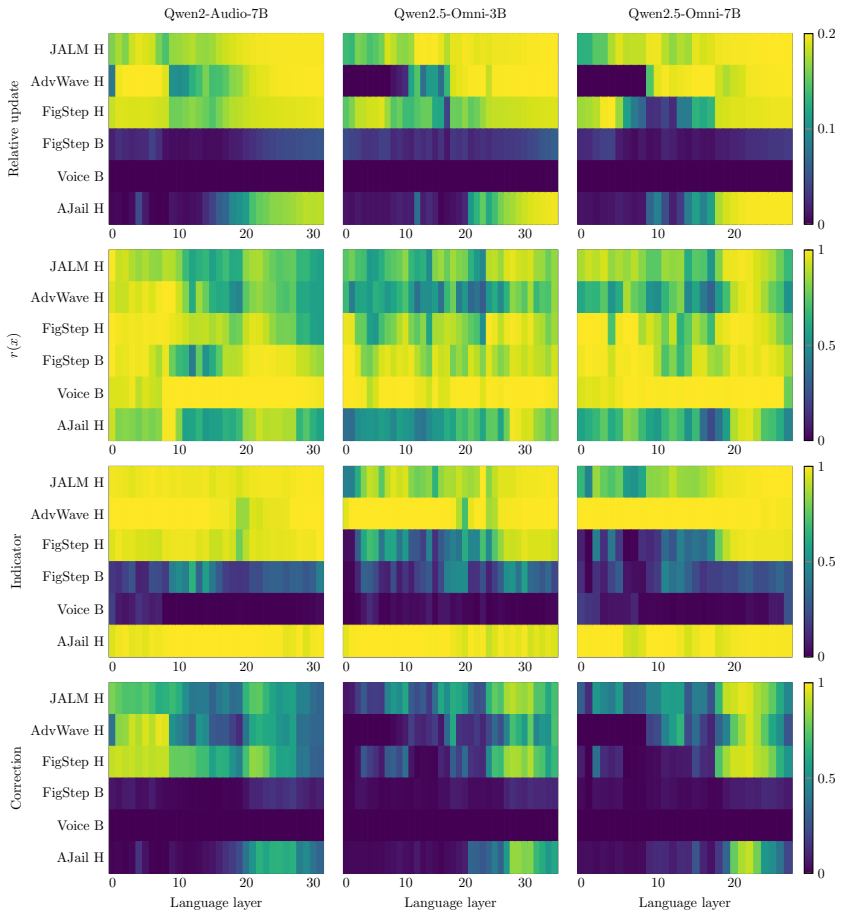

Figure 5: Mean prefill controls by language layer and input group, as defined in Table 6. H and B denote harmful and benign inputs. The mean indicator is the fraction of inputs with acoustic excess. Layer indices are zero-based and model-specific.

C.3 Evaluator Robustness Analysis

We re-evaluate saved responses from all three backbones using DeepSeek-V4-Pro, comparing TriSteer, SARSteer, and AdaShield-S on JALMBench Others, AdvWave, AJailBench, and paired FigStep. This covers 1,171 responses per model–method combination, 10,539 in total. Each response is classified based on its original query and answer using the rubric below; target-model generations are unchanged, and Qwen3Guard labels are not reused.

Table 7: DeepSeek-V4-Pro evaluation rubric.

Input	Response category	Criterion
Harmful	Safe refusal	Refuses, redirects safely, or gives non-enabling high-level information.
	Unsafe compliance	Fulfills the harmful request or provides actionable harmful assistance.
	Other/undetermined	All other cases.
Benign	Safe answer	Meaningfully attempts to answer the benign request.
	Over-refusal	Refuses the benign request on safety grounds.
	Other/undetermined	All other cases.

ASR counts unsafe compliance, ORR counts benign over-refusal, and BRR follows the same aggregation as in the main evaluation. Other/undetermined responses remain separate from explicit refusals. The safe-refusal category is broader than Qwen3Guard's `Safe` and `Refusal: Yes` criterion, so this experiment evaluates robustness to both the evaluator and the rubric.

Table 8 shows that TriSteer achieves the lowest harmful ASR across all evaluated model–benchmark combinations. Its FigStep BRR is highest on two backbones and close to that of AdaShield-S on the third, although benign ORR is not uniformly lowest. These results support the consistency of TriSteer's safety advantage under an alternative evaluation framework.

Table 8: DeepSeek-V4-Pro evaluation of the saved main-experiment responses. All values are percentages; bold marks the best value within each backbone. FigStep reports harmful ASR, benign ORR, and BRR.

Backbone	Method	JALM ASR ↓	AdvWave ASR ↓	AJail ASR ↓	FigStep ASR ↓	FigStep ORR ↓	FigStep BRR ↑
Qwen2-Audio-7B	TriSteer	**0.00**	**0.00**	**2.07**	**2.40**	21.60	**86.80**
	SARSteer	10.09	26.00	4.66	13.60	23.20	79.00
	AdaShield-S	11.40	47.60	24.87	26.80	**4.00**	78.40
Qwen2.5-Omni-3B	TriSteer	**1.75**	**0.00**	**1.04**	**4.00**	12.00	91.40
	SARSteer	28.07	56.00	11.92	29.60	**0.40**	82.60
	AdaShield-S	7.02	49.20	5.70	12.40	2.00	**91.80**
Qwen2.5-Omni-7B	TriSteer	**0.44**	**3.20**	**0.52**	**1.60**	8.80	**94.20**
	SARSteer	19.74	59.60	8.81	25.60	**0.00**	85.80
	AdaShield-S	4.82	59.20	2.59	13.60	0.80	91.60

C.4 Additional Component Ablations

Role of adaptive magnitude. Both Omni models show substantially better VoiceBench answer rates and accuracy with adaptive magnitude than with the fixed-magnitude controls. This utility advantage holds with either fixed or adaptive direction, indicating that magnitude adaptation is the main component preserving benign behavior.

Secondary role of direction refinement. Under adaptive magnitude, direction refinement provides additional safety gains for Omni-7B but yields mixed results for Omni-3B, including a small regression on AdvWave. VoiceBench utility remains effectively unchanged for both models. Refinement under fixed magnitude is similarly inconsistent across safety and utility metrics. These results position direction refinement as a secondary, setting-dependent component. More reliable criteria for when acoustic attribution leads to beneficial direction refinement remain to be investigated.

Table 9: Additional component ablations on Qwen2.5-Omni-3B and -7B, using the same controls and metric definitions as Table 2. The adaptive-direction/fixed-magnitude variant likewise retains the risk-derived $g(x)$ for direction correction. All anchors are audio-native; boldface control labels identify the complete TriSteer configuration. All metrics are percentages.

Model	Direction	Magnitude	JALM	AdvWave	AJail	VoiceBench	
			ASR ↓	ASR ↓	ASR ↓	BAR ↑	Acc. ↑
Omni-3B	Fixed	Fixed	0.88	2.00	0.52	23.37	32.77
	Adaptive	Fixed	1.32	1.60	0.52	23.86	30.60
	Fixed	Adaptive	1.75	1.60	2.59	99.04	77.11
	Adaptive	**Adaptive**	1.75	2.00	2.59	99.04	77.11
Omni-7B	Fixed	Fixed	0.00	0.80	1.55	54.94	66.75
	Adaptive	Fixed	0.00	1.60	1.04	60.24	68.19
	Fixed	Adaptive	0.44	7.20	2.59	98.31	81.45
	Adaptive	**Adaptive**	0.44	6.40	2.07	98.31	81.45

C.5 Supplementary Gating Ablations

We compare TriSteer with Full-only, which independently calibrates full-view gating and retains the original anchor without constructing the three spaces, and Full Irr removal, which retains both risk views but unconditionally removes $\boldsymbol{P}_{\text{irr}}\boldsymbol{v}_F$ from the anchor before norm restoration. Both use adaptive magnitude, $\rho = 0.20$, $Q = 0.90$, and the same supports, evaluation inputs, and scoring protocol. The TriSteer rows reuse the main results in Table 1.

Table 10 shows that full TriSteer improves over Full-only on both AdvWave and AJailBench across all three backbones, with mixed effects on JALMBench Others. Full Irr removal gives lower ASR on Qwen2-Audio AJailBench and Omni-3B AdvWave than full TriSteer, but higher ASR in several other settings, indicating that direction-refinement gains depend on the model and attack. VoiceBench answer rates and accuracy remain unchanged across all three variants.

Table 10: Supplementary gating ablations with adaptive magnitude. Bold method labels mark full TriSteer; all values are percentages.

Model	Method	JALM	AdvWave	AJail	VoiceBench	
		ASR ↓	ASR ↓	ASR ↓	BAR ↑	Acc. ↑
Qwen2-Audio-7B	Full-only	0.44	2.40	10.88	99.76	46.75
	Full Irr removal	0.44	2.40	6.74	99.76	46.75
	TriSteer	0.44	1.60	7.77	99.76	46.75
Omni-3B	Full-only	1.32	4.80	3.11	99.04	77.11
	Full Irr removal	2.19	0.00	3.11	99.04	77.11
	TriSteer	1.75	2.00	2.59	99.04	77.11
Omni-7B	Full-only	0.44	8.40	2.59	98.31	81.45
	Full Irr removal	0.44	8.80	2.59	98.31	81.45
	TriSteer	0.44	6.40	2.07	98.31	81.45