

Discovering and Causally Validating Emotion-Sensitive Neurons in Large Audio-Language Models

Xiutian Zhao¹, Björn Schuller², Berrak Sisman¹

¹ Center for Language and Speech Processing (CLSP), Johns Hopkins University, USA

² Group on Language, Audio & Music (GLAM), Imperial College London, UK

Abstract

Emotion is a central dimension of spoken communication, yet, we still lack a mechanistic account of how modern large audio-language models (LALMs) encode it internally. We present the first neuron-level interpretability study of emotion-sensitive neurons (ESNs) in LALMs and provide causal evidence supporting the existence of such units in Qwen2.5-Omni, Kimi-Audio, and Audio Flamingo 3. Across these three widely used open-source models, we compare frequency-, entropy-, mean-deviation-, and contrast-based neuron selectors on multiple emotion recognition benchmarks. Using inference-time interventions, we reveal a consistent emotion-specific signature: deactivating neurons selected for a given emotion disproportionately degrades recognition of that emotion while largely preserving other classes, whereas targeted steering amplifies these units to bias predictions toward the target emotion. These effects arise with modest amounts of identification data and scale systematically with intervention strength. We further observe that ESNs exhibit non-uniform layer-wise clustering with partial cross-dataset transfer. Taken together, our results offer a causal, neuron-level account of emotion decisions in LALMs and highlight targeted neuron interventions as an actionable handle for controllable affective behaviors.

1 Introduction

The progress of large language models (LLMs) has accelerated the development of multimodal foundation models that jointly process text and other modalities (Wang et al., 2023a). Among them, large audio-language models (LALMs) (Liu et al., 2025; Xu et al., 2025b; KimiTeam et al., 2025; Ghosh et al., 2025), which operate on both speech and text, are increasingly prominent in applications such as conversational assistants, where affective competence is crucial for user trust and safety.

Despite strong empirical performance, however, we still lack a mechanistic understanding of how LALMs internally represent emotion and which components are responsible for emotion-related decisions (Gandhi et al., 2023).

Speech conveys not only linguistic content but also paralinguistic cues (e.g., intonation, pitch, energy) associated with a speaker’s affective state. While decades of affective speech research (Wani et al., 2021) have demonstrated the importance of these cues for tasks such as speech emotion recognition (Akçay and Oğuz, 2020; Ma et al., 2024; Ulgen et al., 2024; Naini et al., 2025) and expressive speech synthesis (Liu et al., 2021; Zhou et al., 2023), it remains unclear whether LALMs encode emotion through compact, intervention-sensitive neuron sets or through diffuse, non-specific mechanisms.

Neuron-level interpretability provides a natural lens for answering this question. Prior work has demonstrated that individual units can specialize to human-interpretable concepts in vision models (Bau et al., 2017, 2020) and exhibit selectivity for linguistic and other conceptual attributes in LLMs (Voita et al., 2024; Yu and Ananiadou, 2024; Alkhamissi et al., 2025). In multimodal settings, however, existing neuron-level studies have largely focused on modality- or task-specific patterns rather than affect, and causal validation remains limited (Wu et al., 2024; Huang et al., 2024; Fang et al., 2024; Xu et al., 2025a). Motivated by these gaps, we ask whether LALMs contain compact neuron subsets that function as emotion-sensitive units whose activation can be manipulated to selectively impair or steer emotion-related behavior.

We frame our study around the following questions:

- **Causality.** Do *emotion-sensitive neurons* (ESNs), i.e., neurons that preferentially activate on in-

puts tied to particular emotions when processing speech, exist in LALMs? Does deactivating these neurons lead to emotion-specific performance degradation, and can amplifying them systematically steer emotion-related model behavior?

- **Selectivity.** Which identification methods are most effective at isolating ESNs and are neurons associated with certain emotions intrinsically harder to detect than others?
- **Locality and Transferability.** How are ESNs distributed across decoder layers, and to what extent do these functional units generalize across datasets and acoustic conditions?

Empirically, we study three open-source LALMs that support direct speech input and text generation: Qwen2.5-Omni-7B (Xu et al., 2025b), Kimi-Audio (KimiTeam et al., 2025), and Audio Flamingo 3 (Ghosh et al., 2025). As probe testbeds, we use three established speech emotion recognition (SER) benchmarks: IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and MSP-PODCAST (Busso et al., 2026). To identify ESNs, we compare multiple selectors that operationalize frequency-, entropy-, mean-deviation-, and contrast-based notions of selectivity. Our results provide converging evidence that emotion-sensitive functional units exist in LALMs and can be causally validated.

First, selective deactivation exhibits clear self-cross structure: masking neurons identified for a given emotion disproportionately degrades recognition of that same emotion (*self-deactivation*), while producing substantially smaller average effects on other emotions (*cross-deactivation*). Critically, we find that not all identification criteria are equally effective: selectors based purely on baseline activation probability or entropy are often insufficient to isolate neurons with more emotion-specific causal effects, whereas mean-deviation- and contrast-based selectors yield neurons with markedly stronger such effects.

Second, we show actionability via activation steering. Amplifying the same neuron sets used for deactivation biases predictions toward the target emotion and can yield positive self-cross gaps, indicating that these units provide a controllable handle rather than being mere correlates. Beyond label-conditioned (targeted) steering, we also study label-free (agnostic) injection strategies that leverage the discovered ESNs without committing to

a chosen source emotion. We evaluate three variants: 2-Pass injection that reinforces the model’s initial prediction, Mix injection that softly weight emotion masks using internal evidence, and Union injection that simply boosts the union of all ESNs. Notably, the gap between reliable targeted steering and mixed agnostic outcomes suggests ESNs may interact non-additively under joint amplification.

Third, we analyze where ESNs reside and how well they transfer. We observe non-uniform layer-wise locality patterns, with ESNs clustering in earliest layer (0), early-mid (6–8), and later (19–22) decoder layers rather than being evenly distributed, and we find asymmetric, yet non-trivial cross-dataset transferability across emotions—suggesting partial robustness but also dataset- and category-dependent specificity.

Overall, this work contributes: (1) to our knowledge, the first neuron-level causal analysis of emotion representations in multiple LALMs via self-/cross deactivation and steering across multiple datasets; (2) a systematic comparison of identification methods showing which criteria best isolate causally emotion-sensitive units and the impacts of selecting parameters; and (3) evidence that these units have structured locality, non-trivial cross-dataset transfer, and can be leveraged for both targeted and label-free control of affective behavior in speech-enabled foundation models.

2 Related Work

Neuron Specialization and Unit Selectivity. Identifying neurons that respond selectively to specific features or concepts is a long-standing theme in interpretability. Prior work has shown that individual units in deep networks can align with human-interpretable concepts. In vision, Network Dissection quantified such alignments for CNN units (Bau et al., 2017, 2020), and recent work extends neuron-level interpretability to LLMs, including evidence that some neurons exhibit partial concept selectivity (Antverg and Belinkov, 2022; Huben et al., 2024; Voita et al., 2024; Yu and Ananiadou, 2024; Tang et al., 2024), especially through causal tracing style localization (Meng et al., 2022; Alkhamissi et al., 2025). Most closely related, studies on LLMs investigate affective mechanisms: Lee et al. (2025) report clustered emotion neurons with deactivation-based validation, and Wang et al. (2025) identify emotion circuits and demonstrate controllability via steering.

Neuron-Level Interpretability in Multimodal and Audio Models. Neuron-level analyses have also been applied to multimodal models, typically to characterize modality- or task-specific attributions (Huang et al., 2024; Fang et al., 2024; Xu et al., 2025a). In the audio domain, interpretability studies of audio and speech transformers often rely on layer-wise probing or attribution methods to reveal what information (e.g., phonetic, speaker, prosodic cues) is encoded across representations (Singla et al., 2022; Akman et al., 2025; Yang et al., 2025). Audio Network Dissection (Wu et al., 2024) labels acoustic units with natural language descriptors by summarizing responsive audio snippets, but it primarily targets generic acoustic/structural concepts and provides limited evidence for emotion-specific causal roles.

3 Methods

We study whether LALMs contain neurons that are selectively active for emotions by coupling activation-based neuron analysis with causal interventions. Concretely, we follow an activation-based log-identify-intervene workflow (Huo et al., 2024; Huang et al., 2024; Fang et al., 2024; Tang et al., 2024) with SER as a diagnostic task: (1) we instrument decoder MLPs and collect neuron activations while the unintervened model answers correctly, (2) we score and select neurons for emotion sensitivity and construct masks, and (3) we intervene at inference time by manipulating identified neurons and quantifying their causal effects.

3.1 Activation Logging and Emotion-Sensitive Neuron Identification

We attach forward hooks to the decoder MLP feed-forward blocks and log internal activations on correctly solved SER items. The motivation is pragmatic: restricting to correct items reduces contamination from failure-mode generations and yields cleaner emotion-conditioned activation statistics.

Within each decoder MLP, we record the gating signal from the SwiGLU nonlinearity (Shazeer, 2020). Let u and v denote the two pre-activation streams, and let $g = \text{SiLU}(u)$ be the gated branch that modulates v . For layer l , neuron index n , and token position t , we denote the logged scalar gate activation by $a_{l,n,t}$ (the n -th coordinate of g at position t). These values serve as the basis for all subsequent statistics.

Activation Statistics. Let \mathcal{E} be the emotion set. For each identification example labeled $e \in \mathcal{E}$, we aggregate gate activations across valid token positions, using an indicator $m_t \in \{0, 1\}$ to exclude padding and other special markers. For every neuron (l, n) we maintain: (1) a positive-activation count $K_{l,n}^{(e)}$, and (2) the total number of valid token positions T_e contributed by emotion- e examples:

$$K_{l,n}^{(e)} = \sum_t m_t \mathbb{I}(a_{l,n,t}^{(e)} > 0)$$

$$T_e = \sum_t m_t$$

Intuitively, $K_{l,n}^{(e)}$ captures how often the unit is active under emotion e . Normalizing by T_e yields the emotion-conditioned activation probability profile, $P_{l,n}^{(e)}$, which serves as the sufficient statistic for all selectors below:

$$P_{l,n}^{(e)} = \frac{K_{l,n}^{(e)}}{T_e}$$

Here, $P_{l,n}^{(e)}$ reflects the empirical firing frequency (i.e., probability of positive activation) for the given neuron.

Identification Methods. Given $P_{l,n}^{(e)}$, we score neurons using the following four established methods, in addition to an emotion-independent random selection baseline.

- **Random Selection (RND)** is employed as an emotion-agnostic control. We construct a random mask by sampling the same total budget of neurons (i.e., the same r used by targeted selectors) uniformly over all decoder MLP neurons. We do *not* enforce layer-wise matching because it adds bookkeeping and compute overhead; in pilot checks, a layer-matched variant produced similar intervention effects within sampling variance. We therefore report the simpler global RND baseline throughout. Unless stated otherwise, we report RND results averaged over 5 independent random masks of different seeds.
- **Activation Probability (LAP)** (Huben et al., 2024; Gurnee et al., 2024) prioritizes neurons that are frequently active for a particular emotion, using only the frequency statistic $P_{l,n}^{(e)}$:

$$\text{LAP}_{l,n}^{(e)} = P_{l,n}^{(e)} = \frac{K_{l,n}^{(e)}}{T_e} \quad (1)$$

- **Activation Probability Entropy (LAPE)** (Tang et al., 2024; Namazifard and Poech, 2025) evaluates selectivity across emotions by forming a normalized distribution over $e \in \mathcal{E}$ for each neuron and computes its Shannon entropy. Lower entropy corresponds to more concentrated firing and thus stronger specialization:

$$\text{LAPE}_{l,n} = -\sum_{e \in \mathcal{E}} \tilde{P}_{l,n}^{(e)} \log \tilde{P}_{l,n}^{(e)}, \quad (2)$$

$$\tilde{P}_{l,n}^{(e)} = \frac{P_{l,n}^{(e)}}{\sum_{e'} P_{l,n}^{(e')}}. \quad (3)$$

Since LAPE assigns each neuron a single selectivity score $\text{LAPE}_{l,n}$ that is not conditioned on any specific emotion, we deterministically map neurons to emotions using the estimated firing probabilities $P_{l,n}^{(e)}$ to enable per-emotion interventions.

- **Mean Activation Difference (MAD)** (Bau et al., 2019; Dalvi et al., 2019) evaluates selectivity by measuring how much a neuron’s activation probability for a specific emotion deviates from its average activation probability across all emotions. Large absolute values indicate neurons that fire distinctively (either significantly more or significantly less frequently) for emotion e compared to their baseline behavior across all categories:

$$\text{MAD}_{l,n}^{(e)} = \left| P_{l,n}^{(e)} - \bar{P}_{l,n} \right|, \quad (4)$$

$$\bar{P}_{l,n} = \frac{1}{|\mathcal{E}|} \sum_{e' \in \mathcal{E}} P_{l,n}^{(e')}. \quad (5)$$

- **Contrastive Activation Margin (ConAct)** (Zhao et al., 2026a) is a margin-style criterion: for each neuron, it compares the top firing probability across emotions with the runner-up, and assigns the margin to the highest-scoring emotion while assigning zero to all other emotions. Concretely, using the firing probabilities $P_{l,n}^{(e)}$, define:

$$P_{l,n}^{(1)} = \max_{e \in \mathcal{E}} P_{l,n}^{(e)}, \quad e_{l,n}^{(1)} = \arg \max_e P_{l,n}^{(e)}$$

$$P_{l,n}^{(2)} = \max_{e \in \mathcal{E} \setminus \{e_{l,n}^{(1)}\}} P_{l,n}^{(e)}, \quad (6)$$

$$\text{ConAct}_{l,n}^{(e)} = \begin{cases} P_{l,n}^{(1)} - P_{l,n}^{(2)}, & \text{if } e = e_{l,n}^{(1)} \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Emotion-Sensitive Neurons Ranking. For each selector m and each emotion e , we obtain a global ranking of candidate neurons by their emotion- e score. We treat the intervention size as a hyperparameter and, for method comparability, always select a fixed fraction r of the highest-ranked ones as the *emotion-sensitive neurons (ESNs)*. Formally, let D_l be the width of the monitored MLP gate vector at decoder layer l (i.e., the number of gate units/neuron dimensions we log in that layer). For selector m and emotion e , we denote the chosen index set by $\mathcal{I}_l^{(m,e)} \subseteq \{1, \dots, D_l\}$ and use $\{\mathcal{I}_l^{(m,e)}\}_l$ as the mask support for deactivation and steering in §3.2.

3.2 Intervention: Deactivation, Targeted Steering and Agnostic Injection

To test whether the identified ESNs are not merely correlational but causally influential in emotion-related decisions, we intervene on their gate activations at inference time. Let $g_{l,t} \in \mathbb{R}^{D_l}$ denote the SwiGLU gate output at decoder layer l and token position t , i.e., $g = \text{SiLU}(u)$. Given ESN indices $\mathcal{I}_l^{(m,e_{\text{src}})}$, we build a layer-specific mask that either suppresses or amplifies exactly those indices while leaving all other parameters unchanged.

Deactivation. We evaluate necessity by zeroing the selected neurons through an elementwise mask:

$$r_{l,n}^{(m,e_{\text{src}})} = \begin{cases} 0, & n \in \mathcal{I}_l^{(m,e_{\text{src}})} \\ 1, & \text{otherwise,} \end{cases} \quad (8)$$

and applying it to the gate vector:

$$\tilde{g}_{l,t}^{\text{abl}} = g_{l,t} \odot r_l^{(m,e_{\text{src}})}. \quad (9)$$

Targeted (emotion-specific) Steering. By scaling the same coordinates with a gain factor $\alpha \geq 0$ using a per-layer scale vector $s_l(\alpha)$ (Turner et al., 2024), yielding the steered gate $\tilde{g}_{l,t}^{\text{steer}}$. This intervention increases the contribution of ESN dimensions associated with e_{src} without modifying weights and is applied to evaluate controllability.

$$s_{l,n}(\alpha) = \begin{cases} 1 + \alpha, & n \in \mathcal{I}_l^{(m,e_{\text{src}})} \\ 1, & \text{otherwise,} \end{cases} \quad (10)$$

$$\tilde{g}_{l,t}^{\text{steer}} = g_{l,t} \odot s_l(\alpha). \quad (11)$$

Agnostic Injection. Targeted steering requires specifying a source emotion e_{src} (hence selecting $\mathcal{I}_l^{(m, e_{\text{src}})}$). In many settings, however, a *label-free* intervention that leverages the discovered ESNs without committing to a chosen emotion is demanded. We therefore implement three agnostic injection strategies, including 2-Pass, Mix, and Union injections. These strategies are inspired by classic self-training / bootstrapping ideas (Yarowsky, 1995; McClosky et al., 2006; Zelikman et al., 2022), as well as soft routing / mixture weighting mechanisms (Shazeer et al., 2017). Implementation details are provided in Appendix B.

Beyond serving as label-free control baselines, these agnostic strategies also act as a probe of inter-emotion interactions: unlike targeted steering, they may jointly amplify multiple emotion-linked neuron sets (or amplify a mispredicted set), which can induce interference if affective circuits share downstream bottlenecks or exert competing influences on the final decision.

4 Experiment Setup

Models and Datasets. We study three open-source LALMs that accept speech input and demonstrate strong general audio understanding, including Qwen2.5-Omni-7B (Xu et al., 2025b), Kimi-Audio (KimiTeam et al., 2025) and Audio Flamingo 3 (Ghosh et al., 2025). The model versions and dataset splits are listed in Appendix A.1 (Table 3, 4).

We evaluate the models on three widely used SER benchmarks: IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and MSP-PODCAST (Busso et al., 2026). We focus on overlapping subsets of discrete emotion categories that are consistently annotated across these datasets. For evaluation, we construct *class-balanced* test subsets with a fixed number of utterances per emotion; for identification, we additionally *cap* the pool of correctly answered items per emotion to make selectors comparable across categories.

Evaluation Protocol (Self vs Cross Effects). After selecting ESNs for each source emotion $e_{\text{src}} \in \mathcal{E}$ (for every selector in §3.1), we evaluate their speech emotion-sensitivity by running the model on held-out utterances with and without intervention. We report two complementary settings. In the *self-effect* condition ($e_{\text{src}} = e_{\text{eval}}$), the intervention targets neurons identified from the same emotion as the evaluated subset. In the *cross-effect*

condition ($e_{\text{src}} \neq e_{\text{eval}}$), we reuse an emotion- e_{src} mask while evaluating on a different emotion subset e_{eval} . For deactivation, we expect performance to decrease; for targeted steering, we expect increases toward e_{src} . Comparing self vs. cross effects isolates whether a mask primarily modulates a specific emotion pathway rather than causing broad, non-specific degradation or global changes in affective processing.

Prompting and Decoding. All models are evaluated in a controlled multiple-choice SER format using a single instruction template (Appendix A.3). To reduce known multiple-choice artifacts such as label/position preferences (Zheng et al., 2024; Zhao et al., 2024), we randomize the option-number \leftrightarrow emotion mapping for every evaluation item, so a preference for a particular number cannot systematically inflate any single emotion. We decode deterministically (greedy; temperature 0 with sampling disabled) and cap generation at 20 tokens, which is sufficient for the required short-form response. Since some instruction-tuned models may still emit extra text, we post-process generations with a lightweight parsing routine described in Appendix A.2.

5 Results

5.1 Deactivation

The deactivation section (left half) in Table 1 shows that masks produced by MAD and ConAct consistently yield a strong separation between self- and cross-effects: performance drops sharply when deactivating ESNs tied to the evaluated emotion, while the average cross-emotion changes are smaller in magnitude, yielding substantial self–cross gaps. Across the three LALMs, this manifests as large negative self-effects (11–15 accuracy points) paired with near-zero cross-effects, producing substantial self–cross gaps. Deactivating the same-sized random masks (RND) produces smaller and less structured changes, supporting that we are not merely removing generic capacity. LAP/LAPE do not reliably produce a clean diagonal signature (often yielding weak, noisy shifts or broader degradation). Figure 1 visualizes this difference: MAD/ConAct show pronounced diagonal dominance, with limited off-diagonal spillover, consistent with emotion-specific units rather than purely correlated cues.

LALM	Acc. Δ	Deactivation					Targeted Steering				
		RND	LAP	LAPE	MAD	ConAct	RND	LAP	LAPE	MAD	ConAct
Qwen2.5-Omni-7B	Self-Effect	0.32	-7.62	1.04	-13.09	-13.50	0.01	0.12	0.36	2.48	2.73
	Cross-Effect Avg.	-	-7.33	0.04	0.19	1.75	-	0.00	-0.04	-0.25	-0.60
	Self-Cross Gap	-	-0.29	1.00	-13.28	-15.25	-	0.12	0.40	2.73	3.33
Kimi-Audio	Self-Effect	-0.51	0.27	-0.81	-13.63	-11.65	-0.19	-0.99	-0.38	2.25	1.94
	Cross-Effect Avg.	-	-0.55	-0.92	-1.27	0.44	-	-0.78	-0.25	-0.66	-0.39
	Self-Cross Gap	-	0.83	0.10	-12.36	-12.09	-	-0.21	-0.13	2.91	2.78
Audio Flamingo 3	Self-Effect	-0.13	-34.62	-6.49	-15.17	-14.63	-0.20	-0.18	1.06	2.97	3.35
	Cross-Effect Avg.	-	-35.30	-1.88	-1.96	0.70	-	-0.04	-0.30	-0.36	-0.72
	Self-Cross Gap	-	0.68	-4.61	-13.21	-15.33	-	-0.14	1.36	3.33	4.07

Table 1: Macro-averaged effects of deactivation (left) and targeted steering (right) across three datasets ($r = 0.5\%$), using ESNs produced by five identification methods. For each method, we report two evaluation settings: *self-effect* ($e_{\text{src}} = e_{\text{eval}}$) and *cross-effect* (averaged over $e_{\text{src}} \neq e_{\text{eval}}$). We quantify emotion specificity via the *self-cross gap* (self minus cross). All entries are *accuracy-changes* relative to the corresponding full model. Random selection (RND) samples neurons without emotion conditioning and therefore has no self/cross distinction. Per-dataset breakdowns are provided in Appendix C.1 and C.2 (Table 5, 6 and 7 for deactivation, Table 8, 9 and 10 for steering).

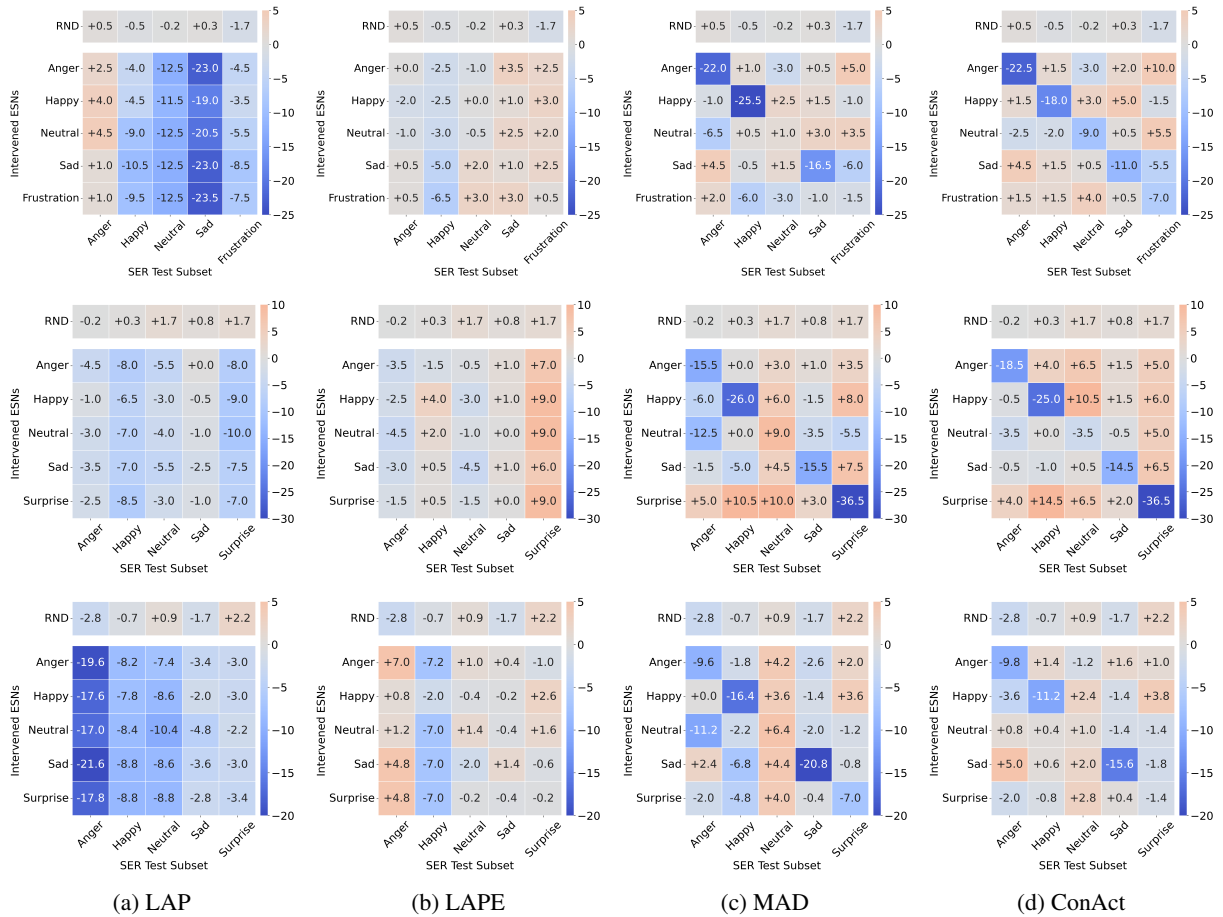


Figure 1: Per-emotion accuracy-change heatmaps for Qwen2.5-Omni-7B under neuron *deactivation*, reported on IEMOCAP (top), MELD (middle), and MSP-PODCAST (bottom). Rows index the *source emotion* used to identify the ESN mask; columns index the *evaluation emotion subset*. All values are absolute accuracy differences with respect to the unintervened model. Diagonals correspond to self-effects, while off-diagonals reflect cross-effects.

Effect of ESN Set Size. Figure 2(a–c) varies the fraction of deactivated neurons (r) while fixing the model and selector, revealing a trade-off between selectivity and intervention strength. For

small r , deactivation already yields clear diagonal patterns, indicating that a small subset of neurons can suffice to induce emotion-specific degradation. As r increases, self-deactivation effects strengthen,

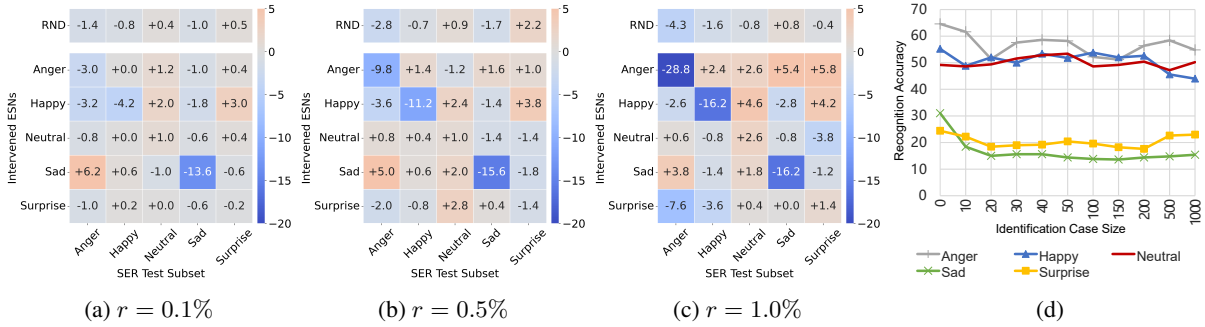


Figure 2: Sensitivity to intervention budget and identification-pool size. (a–c) Accuracy-change heatmaps as we vary the deactivated fraction r of ESNs. (d) Accuracies as we vary the number of correctly answered identification examples per emotion used to construct the ESN masks (Qwen2.5-Omni-7B, ConAct-selected ESNs, MSP-Podcast).

but off-diagonal changes also grow, reflecting increased collateral disruption of shared circuitry and a shift toward broader capacity loss. Thus, larger masks amplify intervention strength but reduce interpretability by mixing emotion-specific and emotion-general effects. This motivates using a moderate r (i.e., 0.5%) in subsequent experiments to balance causal potency with clean self–cross dis-sociation.

Effect of Identification Pool Size. Figure 2(d) examines how many correctly answered instances per emotion are required to obtain stable ESNs. The curves plateau rapidly: a small identification pool already produces intervention effects comparable to those obtained with hundreds of instances, with larger pools yielding diminishing returns. This indicates that once the model observes a modest number of representative utterances, neuron rankings and downstream intervention behavior become largely stable. Overall, these results suggest that stable neuron identification does not require extremely large pools, enabling efficient analysis even for low-resource emotion categories.

5.2 Activation Steering

Targeted Steering. Table 1 right half shows that amplifying the same ESNs identified for deactivation yields complementary, constructive effects. Across all three LALMs, MAD and ConAct produce consistent self-steering gains (approximately +2 to +3 accuracy points) while leaving cross-steering effects largely unchanged on average, resulting in the largest self–cross gaps. In contrast, RND is effectively neutral, and LAP/LAPE yield only small or unstable improvements, mirroring their weaker selectivity under deactivation. This symmetry between deactivation and steering

strengthens the causal interpretation: neurons that are important for recognizing a target emotion can be sufficient to bias predictions toward that emotion when amplified, without broadly affecting others.

Figure 3 further illustrates a strength–specificity trade-off as the steering gain α increases. At low α , effects remain strongly diagonal, while larger α amplifies the diagonal gain but can introduce modest off-diagonal spillover. We view spillover not merely as noise, but as potential evidence that ESNs are not fully independent: sufficiently strong amplification can perturb shared downstream computation, revealing coupling (and potential competition) between affective pathways. Overall, targeted steering demonstrates that ESNs provide an actionable handle for controlled, emotion-specific behavior modulation.

Agnostic Injection. Unlike targeted steering, agnostic injection does not condition on a source emotion. As summarized in Table 2 provides dataset-wise results), gains are modest and model-dependent: Mix and Union improve Qwen2.5-Omni-7B (up to +0.9 for Mix) and Audio Flamingo 3 (up to +1.0 for Union), but fail to consistently benefit Kimi-Audio, where all strategies slightly underperform the unmasked baseline. In contrast to the strong and consistent targeted steering gains, this suggests that the naive Union injection over all ESNs can trigger inter-emotion interference. Concretely, 2-Pass may reinforce early mistakes by amplifying the ESNs associated with the model’s first-pass prediction, while Union injects a broad ESN set that can push multiple affective directions simultaneously, reducing decisiveness when the activated units are misaligned with the true affect. Mix offers a softer compromise, but still lacks consistency across models, consistent

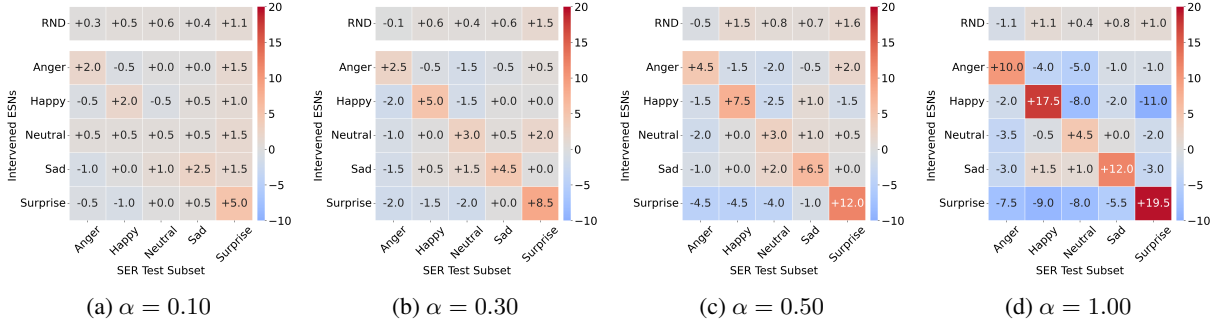


Figure 3: Accuracy-change Δ heatmaps on MELD for different steering strengths α (ConAct, Qwen2.5-Omni-7B).

Dataset	Model	Unmasked	$\alpha = 0.1$				$\alpha = 0.3$				$\alpha = 1.0$			
			RND	2-Pass	Mix	Union	RND	2-Pass	Mix	Union	RND	2-Pass	Mix	Union
IEMOCAP	Qwen2.5-Omni-7B	48.4	48.2	48.1	48.2	48.3	47.9	48.7	48.4	48.9	47.6	49.1	49.2	49.0
	Kimi-Audio	64.6	64.3	62.0	64.2	64.1	66.2	61.7	64.2	64.5	63.4	60.6	65.1	60.7
	Audio Flamingo 3	59.5	59.7	59.8	59.9	60.0	59.5	59.6	60.2	60.2	59.4	59.3	60.4	61.6
MELD	Qwen2.5-Omni-7B	45.3	45.9	45.9	46.1	46.0	45.9	45.7	45.6	46.4	45.7	45.5	46.7	44.5
	Kimi-Audio	57.4	57.3	58.2	57.5	57.1	57.1	58.0	57.5	56.8	57.1	56.3	56.8	53.2
	Audio Flamingo 3	49.4	49.1	49.3	49.6	49.2	49.0	49.3	49.1	49.6	48.7	49.1	49.7	48.3
MSP-Podcast	Qwen2.5-Omni-7B	44.9	44.7	44.9	44.8	45.2	44.8	45.0	44.9	45.9	45.4	45.4	45.3	45.9
	Kimi-Audio	47.9	47.9	47.8	47.9	47.8	48.0	47.6	47.9	47.9	47.7	46.7	47.7	46.4
	Audio Flamingo 3	51.9	51.9	52.3	52.2	52.4	51.7	52.6	52.5	53.0	51.0	52.6	52.9	54.0

Table 2: Agnostic injection results using different injection strategies across three α values. For the Mix method, we are showing the results for $\tau = 0.5$.

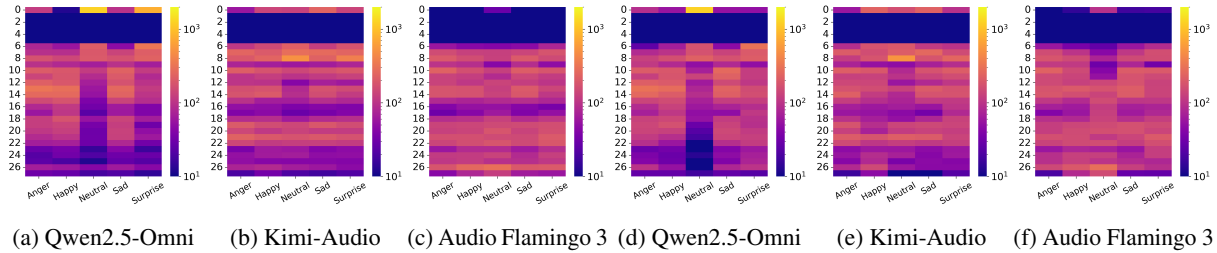


Figure 4: Layer-wise distribution of identified ESNs by MAD (subfigure a, b, c) and ConAct (subfigure d, e, f). All three models have 28-layer decoders. The color is log-scaled for better readability.

with partial cancellation among competing affective pathways. Overall, these results indicate that basic agnostic injection is a weaker and less reliable control mechanism than targeted steering, and they hint at a non-trivial competitive structure among ESNs.

5.3 Locality and Transferability

Locality. Figure 4 shows the layer-wise distribution of ESNs identified by MAD and ConAct on MSP-PODCAST across the three models. ESNs consistently cluster in the earliest layer (layer 0), early-mid layers (6–8), and later layers (19–22), with relatively sparse presence in central blocks (15–18). Both methods largely bypass these middle layers. Interestingly, the Neutral category exhibits the strongest emotion-specific deviations, for

which both methods pick unique patterns. Overall, these results indicate that the layer distribution of ESNs depends on both the identification method and the emotion category, suggesting that different affective states engage different depths of the network.

Cross-Dataset Transferability. We further investigate the dataset-independent generalization of ESNs by evaluating whether ESNs identified on one dataset remain causally effective when deactivated on another. Figure 5 showcases four source-target dataset transfers. Across all six transfer directions, we observe recurring diagonal structure for shared emotions, indicating that many ESNs encode more dataset-robust affective computations rather than corpus-specific artifacts. However,

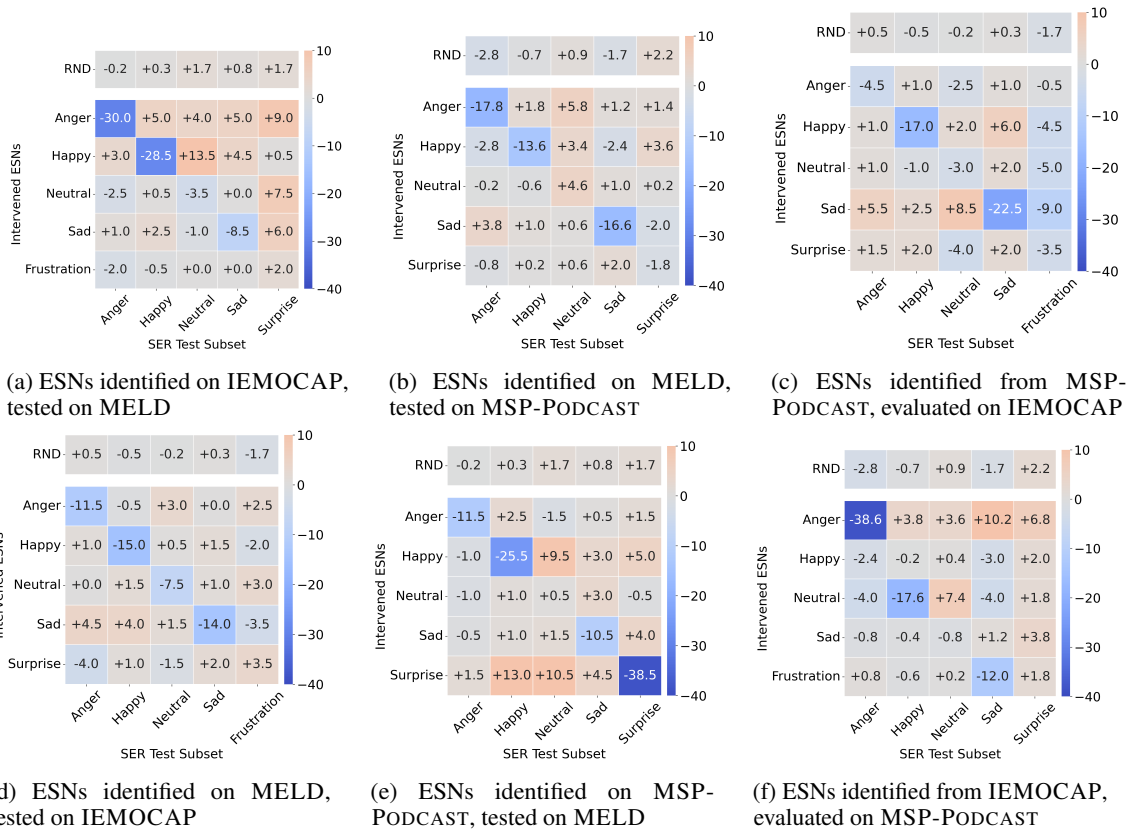


Figure 5: Accuracy-change heatmaps on **cross-dataset deactivation**. The results of Qwen2.5-Omni-7B using ConAct selector are shown. Note that while all datasets contain Anger, Joy/Happiness, Neutral and Sadness; MELD and MSP-PODCAST additionally share Surprise; IEMOCAP has Frustration.

transfer strength is uneven and sometimes asymmetric: certain source–target pairs preserve strong self-deactivation effects, while others degrade, consistent with differences in speaking style, acoustic conditions, and annotation practices across datasets. Among all emotions, Neutral exhibits the least stable transfer by often showing smaller or non-diagonal effects, which suggests that Neutral may rely more on dataset-dependent decision heuristics (or the absence-of-evidence boundary) than on a single portable neuron subset. These results point to a mixed but encouraging picture: ESNs show partial transfer, but their strength and selectivity depend on both the emotion category and the source–target distribution, motivating multi-dataset identification for robust control.

6 Conclusion

In this work, we presented a neuron-level causal study of emotion-related decisions in LALMs, and found consistent evidence that compact ESNs exist across Qwen2.5-Omni-7B, Kimi-Audio, and Audio Flamingo 3. Across three benchmarks, we observe

clear self–cross intervention signatures that are difficult to attribute to generic capacity loss: deactivating MAD/ConAct-selected ESNs produces strong emotion-specific drops while largely preserving other emotions. Methodologically, we find that ESN identification is highly method-dependent: MAD/ConAct consistently yields more selective and stable ESN sets than LAP/LAPE or random baselines. Actionably, amplifying the same ESNs yields reliable targeted steering gains with minimal cross-emotion spillover. Beyond targeted control, we evaluated agnostic injection strategies and found mixed outcomes, which hints at competitive interaction among ESNs. We further find that ESNs stabilize with modest identification pools, exhibit non-uniform layer-wise locality and uneven, yet non-trivial cross-dataset transfer.

Together, these findings provide evidence from causal interventions that compact, emotion-sensitive functional units exist in LALMs and that neuron-level interventions offer a practical handle for interpreting and controlling affective behavior in speech-enabled foundation models.

Limitations

While our results consistently support the existence and controllability of ESNs in multiple LALMs, several aspects remain outside the current study’s scope. Methodologically, we operationalize neuron behavior through decoder SwiGLU MLP gate activations and evaluate causality via targeted inference-time deactivation and gain-based amplification. These interventions are intentionally lightweight and comparable across architectures, but they do not fully characterize how parameterized emotion cues are distributed across other components (e.g., attention and audio–text fusion) or how multiple units compose into higher-level circuits. Additionally, we study transfer primarily across datasets within the SER setting. Understanding whether emotion-sensitive units generalize across tasks, such as expressive speech generation (Zhao et al., 2026b), and how to make steering more uniformly reliable remains an important direction for future work. Finally, while the weaker and less stable outcomes of agnostic injection hint at inter-emotion interference among ESNs, we do not yet provide a dedicated causal decomposition of these interactions (e.g., pairwise co-steering or controlled multi-emotion activation studies); a systematic characterization of competitive vs. cooperative affective circuitry is an important direction for future work.

Ethical Considerations

Our experiments are conducted on public datasets and open-sourced models, and we emphasize that our results should not be interpreted as validating emotion inference as a reliable proxy for human mental state, intent, or truthfulness.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under CAREER Award IIS-2533652.

References

- Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76.
- Alican Akman, Qiyang Sun, and Björn W. Schuller. 2025. [Improving audio explanations using audio](#)

[language models](#). *IEEE Signal Processing Letters*, 32:741–745.

- Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025. [The LLM language network: A neuroscientific approach for identifying causally task-relevant units](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10887–10911, Albuquerque, New Mexico. Association for Computational Linguistics.
- Omer Antverg and Yonatan Belinkov. 2022. [On the pitfalls of analyzing individual neurons in language models](#). In *International Conference on Learning Representations*.
- Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. [Identifying and controlling important neurons in neural machine translation](#). In *International Conference on Learning Representations*.
- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. [Network Dissection: Quantifying Interpretability of Deep Visual Representations](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, Los Alamitos, CA, USA. IEEE Computer Society.
- David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. [Understanding the role of individual units in a deep neural network](#). *Proceedings of the National Academy of Sciences*, 117(48):30071–30078.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.
- Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N. Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, Huang-Cheng Chou, and Pravin Mote. 2026. [The msp-podcast corpus](#). *IEEE Transactions on Affective Computing*, pages 1–19.
- Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. [What is one grain of sand in the desert? analyzing individual neurons in deep nlp models](#). In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI’19/IAAI’19/EAAI’19. AAAI Press.
- Junfeng Fang, Zongze Bi, Ruipeng Wang, Houcheng Jiang, Yuan Gao, Kun Wang, An Zhang, Jie Shi, Xi-ang Wang, and Tat-Seng Chua. 2024. Towards neuron attributions in multimodal large language models. In *Proceedings of the 38th International Conference*

- on *Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. [Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity](#). *Journal of Machine Learning Research*, 23(120):1–39.
- Kanishk Gandhi, J.-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.
- Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. [Audio flamingo 3: Advancing audio intelligence with fully open large audio language models](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Yves Grandvalet and Yoshua Bengio. 2004. [Semi-supervised learning by entropy minimization](#). In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.
- Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. [Universal neurons in GPT2 language models](#). *Transactions on Machine Learning Research*.
- Kaichen Huang, Jiahao Huo, Yibo Yan, Kun Wang, Yutao Yue, and Xuming Hu. 2024. [Miner: Mining the underlying pattern of modality-specific neurons in multimodal large language models](#). *Preprint*, arXiv:2410.04819.
- Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. [Sparse autoencoders find highly interpretable features in language models](#). In *The Twelfth International Conference on Learning Representations*.
- Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. [MMNeuron: Discovering neuron-level domain-specific interpretation in multimodal large language model](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6816, Miami, Florida, USA. Association for Computational Linguistics.
- KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. [Kimi-audio technical report](#). *Preprint*, arXiv:2504.18425.
- Jaewook Lee, Woojin Lee, Oh-Woog Kwon, and Harksoo Kim. 2025. [Do large language models have “emotion neurons”? investigating the existence and role](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15617–15639, Vienna, Austria. Association for Computational Linguistics.
- Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. [Voxtral](#). *Preprint*, arXiv:2507.13264.
- Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. 2021. [Expressive tts training with frame and style reconstruction loss](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1806–1818.
- Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. [emotion2vec: Self-supervised pre-training for speech emotion representation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand. Association for Computational Linguistics.
- David McClosky, Eugene Charniak, and Mark Johnson. 2006. [Effective self-training for parsing](#). In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.
- Abinay Reddy Naini, Lucas Goncalves, Ali N. Salman, Pravin Mote, Ismail R. Ulgen, Thomas Thebaud, Laureano Moro Velazquez, Leibny Paola Garcia, Najim Dehak, Berrak Sisman, and Carlos Busso. 2025. [The Interspeech 2025 Challenge on Speech Emotion Recognition in Naturalistic Conditions](#). In *Interspeech 2025*, pages 4668–4672.
- Danial Namazifard and Lukas Galke Pösch. 2025. [Isolating culture neurons in multilingual large language models](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 768–785, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. [MELD: A multimodal multi-party dataset for emotion recognition in conversations](#). In

- Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.
- Noam Shazeer. 2020. [Glu variants improve transformer](#). *Preprint*, arXiv:2002.05202.
- Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. [Outrageously large neural networks: The sparsely-gated mixture-of-experts layer](#). In *International Conference on Learning Representations*.
- Yaman Kumar Singla, Jui Shah, Changyou Chen, and Rajiv Ratn Shah. 2022. [What do audio transformers hear? probing their representations for language delivery & structure](#). In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 910–925.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. [Language-specific neurons: The key to multilingual capabilities in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. [Steering language models with activation engineering](#). *Preprint*, arXiv:2308.10248.
- Ismail Rasim Ulgen, Zongyang Du, Carlos Busso, and Berrak Sisman. 2024. [Revealing emotional clusters in speaker embeddings: A contrastive learning strategy for speech emotion recognition](#). In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12081–12085.
- Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2024. [Neurons in large language models: Dead, n-gram, positional](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1288–1301, Bangkok, Thailand. Association for Computational Linguistics.
- Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia, Huishuai Zhang, Dongyan Zhao, and Xiuying Chen. 2025. [Do llms "feel"? emotion circuits discovery and control](#). *Preprint*, arXiv:2510.11328.
- Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023a. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.
- Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. 2021. A comprehensive review of speech emotion recognition systems. *IEEE access*, 9:47795–47814.
- Tung-Yu Wu, Yu-Xiang Lin, and Tsui-Wei Weng. 2024. [And: audio network dissection for interpreting deep acoustic models](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Jiaqi Xu, Cuiling Lan, and Yan Lu. 2025a. [Deciphering functions of neurons in vision-language models](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3173–3181.
- 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. 2025b. [Qwen2.5-omni technical report](#). *Preprint*, arXiv:2503.20215.
- Chih-Kai Yang, Neo Ho, Yi-Jyun Lee, and Hung-Yi Lee. 2025. [Audiolens: A closer look at auditory attribute perception of large audio-language models](#). In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.
- David Yarowsky. 1995. [Unsupervised word sense disambiguation rivaling supervised methods](#). In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.
- Zeping Yu and Sophia Ananiadou. 2024. [Neuron-level knowledge attribution in large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280, Miami, Florida, USA. Association for Computational Linguistics.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. [Star: Bootstrapping reasoning with reasoning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488. Curran Associates, Inc.
- Xiutian Zhao, Rochelle Choenni, Rohit Saxena, and Ivan Titov. 2026a. [Finding culture-sensitive neurons in vision-language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3366–3381, Rabat, Morocco. Association for Computational Linguistics.
- Xiutian Zhao, Ismail Rasim Ulgen, Philipp Koehn, Björn Schuller, and Berrak Sisman. 2026b. [Neuron-level emotion control in speech-generative large audio-language models](#). *Preprint*, arXiv:2603.17231.

Xiutian Zhao, Ke Wang, and Wei Peng. 2024. [Measuring the inconsistency of large language models in preferential ranking](#). In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 171–176, Bangkok, Thailand. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. [Large language models are not robust multiple choice selectors](#). In *The Twelfth International Conference on Learning Representations*.

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W. Schuller, and Haizhou Li. 2023. [Speech synthesis with mixed emotions](#). *IEEE Transactions on Affective Computing*, 14(4):3120–3134.

A Reproducibility Details

A.1 Datasets and Models

Models	Sources
Qwen2.5-Omni-7B	https://huggingface.co/Qwen/Qwen2.5-Omni-7B
Kimi-Audio	https://huggingface.co/moonshotai/Kimi-Audio-7B-Instruct
Audio Flamingo 3	https://huggingface.co/nvidia/audio-flamingo-3

Table 3: Sources of the evaluated models.

We employed the following three models: **Qwen2.5-Omni-7B** (Xu et al., 2025b) is an end-to-end multimodal model with a streaming Thinker–Talker design. **Kimi-Audio** (KimiTeam et al., 2025) is an audio foundation model supporting audio understanding, generation, and conversational interaction. **Audio Flamingo 3** (Ghosh et al., 2025) provides reasoning capabilities over speech, sound, and music, with support for long-form audio. The specific versions are listed in Table 3.

Regarding the datasets, we curate balanced held-out test sets with 200 utterances per emotion for IEMOCAP and MELD, and 500 for MSP-Podcast, as shown in Table 4. All remaining utterances are used for neuron identification, with the number of correctly answered samples per emotion controlled to ensure comparability across categories. Note that the maximum identification set sizes are determined by the lowest number of correctly answered instances per emotion per model: for example, since all models only correctly answered a little above 200 for Joy/Happiness subsets, then the maximum identification set size is set to 200. The counts in Table 4 denote the number of correctly answered instances available for identification after excluding the held-out evaluation set (and after applying caps).

Emotion	Qwen2.5-Omni-7B	Kimi-Audio	Audio Flamingo 3
<i>IEMOCAP</i>			
Anger	360	500	500
Frustration	500	500	500
Joy/Happiness	210	202	209
Neutral	500	500	500
Sadness	320	500	500
Maximum			
Identification Set (Per Emotion)	200	200	200
Evaluation Set (Per Emotion)	200	200	200
<i>MELD</i>			
Anger	500	500	500
Joy/Happiness	500	500	500
Neutral	500	500	500
Sadness	228	326	404
Surprise	500	500	500
Maximum			
Identification Set (Per Emotion)	200	200	200
Evaluation Set (Per Emotion)	200	200	200
<i>MSP-Podcast</i>			
Anger	1000	1000	1000
Joy/Happiness	1000	1000	1000
Neutral	1000	1000	1000
Sadness	1000	1000	1000
Surprise	1000	810	1000
Maximum			
Identification Set (Per Emotion)	1000	800	1000
Evaluation Set (Per Emotion)	500	500	500

Table 4: Correctly-answered pool size (per emotion, per model) and evaluation/identification subsampling. The per-emotion counts are (i) after holding out the evaluation set, and (ii) capped to save computation resources because we only care about the lower bounds (to determine the maximum identification set size).

A.2 Answer Normalization

We parse model outputs into a single discrete option to make evaluation robust to minor formatting variations. Because our SER prompt requests an *option number* (Appendix A.3), we primarily extract an integer in $\{1, \dots, |\mathcal{E}|\}$ from the generation.

Concretely, we normalize the decoded string by lowercasing, collapsing whitespace, and stripping surrounding punctuation. We then apply the following cascade:

- Direct numeric parse.** If the output contains one or more integers in $\{1, \dots, |\mathcal{E}|\}$, we take the *last* such integer as the predicted option (models may mention alternatives before concluding).
- Spelled-out numbers.** If no digit is found, we map common textual forms (e.g., “one”,

“two”) to the corresponding option index when unambiguous.

3. **Fallback emotion-string match.** As a last resort, we match emotion names against the per-item option list (with the same normalization) and again take the last matched option if multiple appear.

If none of the above succeeds, we mark the prediction as invalid for that item.

A.3 Prompt Template for SER

Listing 1: Prompt template used for SER generation, selected emotions are randomly assigned to option indices (e.g., “1”) for each instance.

Listen to the speech clip and choose the correct emotion of the speaker:

1: {emotion 1}
 2: {emotion 2}
 3: {emotion 3}
 4: {emotion 4}
 5: {emotion 5}

Answer with the option index only.

B Agnostic Injection Methods

(1) 2-Pass Self-Consistent Injection. We first run the model without any intervention (Pass 1) and extract the predicted option, which we map to a predicted emotion \hat{e} . In Pass 2, we attach the corresponding mask $\{I_l^{(m,\hat{e})}\}_l$ and apply the corresponding gain-based amplification. This procedure mirrors bootstrapping/self-training in that it uses the model’s own first-pass decision as a pseudo-label, and reinforces it via a second-pass intervention (Yarowsky, 1995; McClosky et al., 2006; Zelikman et al., 2022):

$$\tilde{g}_{l,t}^{2\text{pass}} = g_{l,t} \odot s_l^{(\hat{e})}(\alpha), \quad (12)$$

$$s_{l,n}^{(\hat{e})}(\alpha) = \begin{cases} 1 + \alpha, & n \in \mathcal{I}_l^{(m,\hat{e})}, \\ 1, & \text{otherwise.} \end{cases} \quad (13)$$

We use the Pass 2 output as the final prediction. Intuitively, 2-Pass aims to make the model more *self-consistent* by amplifying ESNs associated with its own inferred affect (Wang et al., 2023b).

(2) Mix Injection. Mix can be viewed as a soft, label-free compromise between no intervention and full targeted steering, guided by the model’s instantaneous internal evidence; the temperature τ regulates how confidently the method concentrates

on a single emotion versus spreading mass across multiple emotions (Grandvalet and Bengio, 2004; Fedus et al., 2022).

For each layer l and emotion $e \in \mathcal{E}$, we compute an evidence score from the current gate activations by averaging over neurons in that emotion’s mask and over token positions:

$$q_l^{(e)} = \mathbb{E}_t \left[\mathbb{E}_{n \in I_l^{(m,e)}} [g_{l,t,n}] \right]. \quad (14)$$

We convert these scores into mixture weights with a temperature-controlled softmax:

$$w_l^{(e)} = \frac{\exp(q_l^{(e)}/\tau)}{\sum_{e' \in \mathcal{E}} \exp(q_l^{(e')}/\tau)}, \quad (15)$$

where $\tau > 0$ controls sharpness (smaller τ yields a more peaked distribution). Finally, we apply a per-emotion scaled gain:

$$\tilde{g}_{l,t,n}^{Mix} = \begin{cases} g_{l,t,n} \cdot (1 + \alpha w_l^{(e)}), & n \in I_l^{(m,e)}, \\ g_{l,t,n}, & \text{otherwise.} \end{cases} \quad (16)$$

Overlapping masks. If a neuron index n belongs to multiple emotion-specific sets at layer l (i.e., $n \in I_l^{(m,e)}$ for more than one e), we apply the strongest multiplicative gain:

$$\tilde{g}_{l,t,n}^{(m)} = g_{l,t,n} \cdot \max_{e: n \in I_l^{(m,e)}} (1 + \alpha w_l^{(e)}),$$

and otherwise $\tilde{g}_{l,t,n}^{(m)} = g_{l,t,n}$.

(3) Union Injection. Union injection is a single-pass label-free baseline that amplifies all ESNs regardless of emotion identity. It corresponds to using no disambiguating routing signal (cf. routing vs. dense activation in mixture-style models) (Fedus et al., 2022). We first form the layer-wise union set:

$$U_l = \bigcup_{e \in \mathcal{E}} I_l^{(m,e)}, \quad (17)$$

and then apply the same gain to every neuron in U_l :

$$\tilde{g}_{l,t,n}^{Union} = \begin{cases} g_{l,t,n} \cdot (1 + \alpha), & n \in U_l, \\ g_{l,t,n}, & \text{otherwise.} \end{cases} \quad (18)$$

Compared to Mix, Union does not attempt to infer which emotion is currently active; it provides a simple way to globally boost emotion-related circuitry in one forward pass, at the cost of reduced specificity.

C Additional Results

C.1 Dataset-Specific Deactivation Results

LALM	Acc.Δ	RND	LAP	LAPE	MAD	ConAct
Qwen2.5-Omni-7B	Self-Deactivation	-0.32	-9.00	-0.30	-12.90	-13.50
	Cross-Deactivation Avg.	-	-8.97	0.25	-0.12	1.43
	Self-Cross Gap	-	-0.03	-0.55	-12.78	-14.92
Kimi-Audio	Self-Deactivation	-0.62	0.40	-1.80	-16.00	-13.00
	Cross-Deactivation Avg.	-	-0.65	-0.60	-2.60	0.12
	Self-Cross Gap	-	1.05	-1.20	-13.40	-13.12
Audio Flamingo 3	Self-Deactivation	-0.30	-32.90	-6.10	-14.60	-13.70
	Cross-Deactivation Avg.	-	-33.07	-0.81	-0.65	0.43
	Self-Cross Gap	-	0.17	-5.25	-12.72	-14.07

Table 5: Deactivation results on IEMOCAP using ESNs selected by five identification methods.

LALM	Acc.Δ	RND	LAP	LAPE	MAD	ConAct
Qwen2.5-Omni-7B	Self-Deactivation	0.86	-4.90	1.90	-16.90	-19.60
	Cross-Deactivation Avg.	-	-4.72	0.68	1.33	3.40
	Self-Cross Gap	-	-0.18	1.23	-18.23	-23.00
Kimi-Audio	Self-Deactivation	-0.62	0.30	-1.20	-12.50	-10.70
	Cross-Deactivation Avg.	-	0.03	-1.53	-0.98	0.10
	Self-Cross Gap	-	0.28	0.32	-11.53	-10.80
Audio Flamingo 3	Self-Deactivation	-0.22	-36.40	-5.40	-15.00	-15.70
	Cross-Deactivation Avg.	-	-36.80	-1.87	-0.15	1.10
	Self-Cross Gap	-	1.40	-3.52	-14.85	-16.80

Table 6: Deactivation results on MELD.

LALM	Acc.Δ	RND	LAP	LAPE	MAD	ConAct
Qwen2.5-Omni-7B	Self-Deactivation	0.41	-8.96	1.52	-9.48	-7.40
	Cross-Deactivation Avg.	-	-8.29	-0.81	-0.65	0.43
	Self-Cross Gap	-	-0.67	2.33	-8.83	-7.83
Kimi-Audio	Self-Deactivation	-0.30	0.12	0.56	-12.40	-11.24
	Cross-Deactivation Avg.	-	-1.04	-0.63	-0.24	1.10
	Self-Cross Gap	-	1.16	1.19	-12.16	-12.34
Audio Flamingo 3	Self-Deactivation	0.14	-34.56	-7.96	-15.92	-14.48
	Cross-Deactivation Avg.	-	-35.04	-2.91	-3.85	0.64
	Self-Cross Gap	-	0.48	-5.05	-12.07	-15.12

Table 7: Deactivation results on MSP-PODCAST

C.2 Dataset-Specific Targeted Steering Results

LALM	Acc.Δ	RND	LAP	LAPE	MAD	ConAct
Qwen2.5-Omni-7B	Self-Steering	-0.54	-0.20	-0.90	0.90	2.20
	Cross-Steering Avg.	-	-0.50	-1.05	-0.08	-0.90
	Self-Cross Gap	-	0.30	0.15	0.98	3.10
Kimi-Audio	Self-Steering	-0.38	-1.50	-0.10	2.10	2.10
	Cross-Steering Avg.	-	-1.10	0.50	-0.90	-1.23
	Self-Cross Gap	-	-0.40	0.40	3.00	3.32
Audio Flamingo 3	Self-Steering	0.02	-0.50	1.30	2.90	3.10
	Cross-Steering Avg.	-	0.00	-0.65	-0.68	-0.53
	Self-Cross Gap	-	-0.50	1.95	3.58	3.63

Table 8: Targeted steering results on IEMOCAP.

LALM	Acc.Δ	RND	LAP	LAPE	MAD	ConAct
Qwen2.5-Omni-7B	Self-Steering	0.60	0.40	-0.10	4.30	4.70
	Cross-Steering Avg.	-	0.30	0.57	-0.05	-0.45
	Self-Cross Gap	-	0.10	-0.67	4.35	5.15
Kimi-Audio	Self-Steering	-0.28	-1.00	-0.60	2.10	1.40
	Cross-Steering Avg.	-	-1.05	-0.15	-0.35	-0.62
	Self-Cross Gap	-	0.05	-0.45	2.45	2.02
Audio Flamingo 3	Self-Steering	-0.36	-0.20	-0.20	2.10	3.00
	Cross-Steering Avg.	-	-0.30	-0.60	-0.57	-1.00
	Self-Cross Gap	-	0.10	-0.40	2.67	4.00

Table 9: Targeted steering results on MELD.

LALM	Acc.Δ	RND	LAP	LAPE	MAD	ConAct
Qwen2.5-Omni-7B	Self-Steering	-0.04	0.16	2.08	2.24	1.28
	Cross-Steering Avg.	-	0.19	0.35	-0.62	-0.46
	Self-Cross Gap	-	-0.03	1.73	2.86	1.74
Kimi-Audio	Self-Steering	0.10	-0.48	-0.44	2.56	2.32
	Cross-Steering Avg.	-	-0.19	-0.09	-0.72	0.68
	Self-Cross Gap	-	-0.29	-0.35	3.28	3.00
Audio Flamingo 3	Self-Steering	-0.27	0.16	2.08	3.92	3.96
	Cross-Steering Avg.	-	0.19	0.35	0.17	-0.63
	Self-Cross Gap	-	-0.03	1.73	3.75	4.59

Table 10: Targeted steering results on MSP-PODCAST.