Neuron-Level Emotion Control in Speech-Generative Large Audio-Language Models

Xiutian Zhao[1], Ismail Rasim Ulgen[1], Philipp Koehn[1], Björn Schuller[2], Berrak Sisman[1]
[1] Center for Language and Speech Processing (CLSP), Johns Hopkins University, USA
[2] Group on Language, Audio & Music (GLAM), Imperial College London, UK

Abstract

Speech-generative large audio-language models (LALMs) can produce expressive speech, yet reliable emotion control remains elusive: conversions often miss the target affect and may degrade linguistic fidelity through refusals, hallucinations, or paraphrase. We present, to our knowledge, one of the first neuron-level studies of emotional voice conversion (EVC) in LALMs and provide intervention-based evidence that compact sets of emotion-sensitive neuron (ESN) are functionally involved in, and can be exploited for, steering expressive speech generation at inference time. ESNs are identified via success-filtered activation aggregation enforcing both emotion realization and content preservation. Across three LALMs (Qwen2.5-Omni-7B, MiniCPM-o 4.5, Kimi-Audio), ESN interventions yield emotion-specific gains that generalize to unseen speakers. Controllability depends on selector design, mask sparsity, filtering, and intervention strength. Our results provide an intervention-based framework for training-free emotion control in LALMs for EVC.[1]

1 Introduction

Emotion is a core dimension of spoken communication: beyond lexical content, prosody, pitch, intensity, and speaking rate convey affect and shape perceived intent (Wani et al., 2021). Recent LALMs that jointly process speech and text (Wang et al., 2023; Défossez et al., 2024) enable expressive speech generation from natural-language instructions, opening new possibilities for conversational agents and speech editing. In practice, however, reliable emotion control remains challenging (Wang et al., 2025; Yang et al., 2025b): the same prompt may produce unintended affect, and even when the target emotion is realized, linguistic fidelity can degrade. Unlike conventional acoustic distortions that reduce intelligibility while preserving meaning, LALMs may also alter the underlying message through refusals, uncontrolled paraphrasing, and semantic hallucinations (Huang et al., 2025).

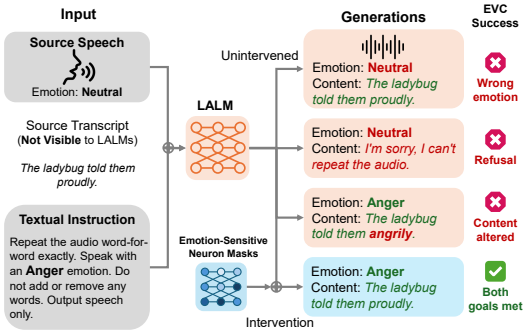


Figure 1: We investigate whether interventions on emotion-sensitive neurons in LALMs yield controllable emotion shifts without compromising content fidelity.

Emotional voice conversion (EVC) aims to modify the emotional style of a speech signal while preserving its linguistic content (Triantafyllopoulos et al., 2023; Yang et al., 2022). As illustrated in Figure 1, EVC is inherently multi-objective: successful conversion requires both target-emotion realization and linguistic content preservation. For speech-generative LALMs, this raises a natural question: are emotion-related control signals localized and accessible enough to support inference-time steering in generating expressive speech?

In large language model (LLM) research, training-free control through activation manipulation has emerged as an alternative to retraining (Turner et al., 2024). Recent work has begun identifying emotion-sensitive neurons (ESNs) in LALMs (Zhao et al., 2026b,c). However, we still lack an intervention-based understanding of whether compact neuron subsets contribute to emotional expressiveness during speech generation and can be manipulated for controllable EVC.

This paper investigates emotion control in speech-generative LALMs from a neuron-level per-

[1]Code is available at https://github.com/xiutian/neuron-emotion-speech-control

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1947–1961
November 6-10, 2026 ©2026 Association for Computational Linguistics

spective, focusing on instruction-following EVC. We hypothesize that compact ESNs exist in speech-generative LALMs and causally contribute to emotional expressiveness during generation. We center the study around the following research questions:

1. **Existence.** Do generation-related ESNs exist in LALMs, and does intervening on them causally affect EVC outcomes? If so, do these effects primarily alter emotional expressiveness or linguistic content?

2. **Identification.** Which neuron selection criteria most effectively isolate ESNs for EVC, and how do mask sparsity and identification data size influence results?

3. **Inference-Time Controllability.** How do intervention type and strength affect emotion control and content preservation in EVC?

4. **Localization.** Where do ESNs concentrate across modules and across layers in LALMs, and what does this imply for controllable EVC?

To answer these questions, we adopt an activation-based log-filter-intervention pipeline. Using the Emotional Speech Database (ESD) (Zhou et al., 2022), we evaluate three open speech-generative LALMs: Qwen2.5-Omni-7B (Xu et al., 2025a), MiniCPM-o 4.5 (Yao et al., 2024), and Kimi-Audio (KimiTeam, 2025). Despite a modest baseline, neuron-level intervention produces consistent and interpretable effects. Compact ESN masks induce structured self–cross patterns rather than arbitrary degradation, indicating that they functionally influence emotional rendering.

Contributions. This paper presents a systematic neuron-level study of emotion control in speech-generative LALMs under an EVC setting. We show that compact ESN masks support training-free emotion steering through inference-time intervention, providing evidence beyond passive correlation that is reflected in both automatic evaluation and human listening. We further analyze the effects of selector design, mask sparsity, identification quality, and intervention strength, finding that moderate steering best balances target-emotion gain and content preservation, whereas overly strong interventions degrade linguistic fidelity. Finally, identified ESNs concentrate in intermediate-to-late decoder MLP layers on the language-model (LM) side, analogous interventions in the downstream synthesis module are largely ineffective, and the control effects extend beyond the identification split.

2 Related Work

Affective speech synthesis and conversion study speech generation and transformation under paralinguistic affect control, most prominently emotion (Triantafyllopoulos et al., 2023). Emotional speech synthesis typically relies on explicit style or emotion representations (Skerry-Ryan et al., 2018; Wang et al., 2018), with later extensions incorporating semi- or weak supervision (Wu et al., 2019; Liu et al., 2021; Lei et al., 2022; Zhang et al., 2023). Other work explores continuous affect control (Lorenzo-Trueba et al., 2018; Lei et al., 2021; Schnell and Garner, 2021; Cho et al., 2024), while recent approaches investigate token- or activation-level steering for expressive TTS (Xie et al., 2025; Zhou et al., 2026; Wang et al., 2026b). EVC studies emotion transfer across speaking styles, including adversarial, diffusion-based, and representation-factorized approaches (Cao et al., 2020; He et al., 2021; Zhou et al., 2021; Fu et al., 2022; Prabhu et al., 2024; Kreuk et al., 2022; Zhang et al., 2026).

Neuron-level interpretability has enabled concept and style control through functional units in LLMs (Yu and Ananiadou, 2024; Rimsky et al., 2024; Lai et al., 2024; Da Silva et al., 2025), with growing interest in multimodal foundation models (Fang et al., 2024; Huo et al., 2024; Yang et al., 2025a). Recent studies further report affective circuits in text LLMs (Lee et al., 2025; Wang et al., 2026a; Sofroniew et al., 2026); however, attempts to identify and intervene on emotion-related functional units in speech-generative LALMs remain limited (Zhao et al., 2026c). Unlike prior work, we directly probe the speech-generation pathway in speech-generative LALMs and apply training-free neuron-level interventions for controllable EVC.

3 Methodology

We study emotion control in LALMs through a neuron-level intervention framework for instruction-following EVC, which presents two key challenges: (1) diverse failure modes from LM backbones, including hallucinations, silence, and refusals; and (2) the inherently multi-objective nature of EVC, which requires both emotional style transfer and linguistic content preservation. To address these challenges, we design a four-stage pipeline illustrated in Figure 2.

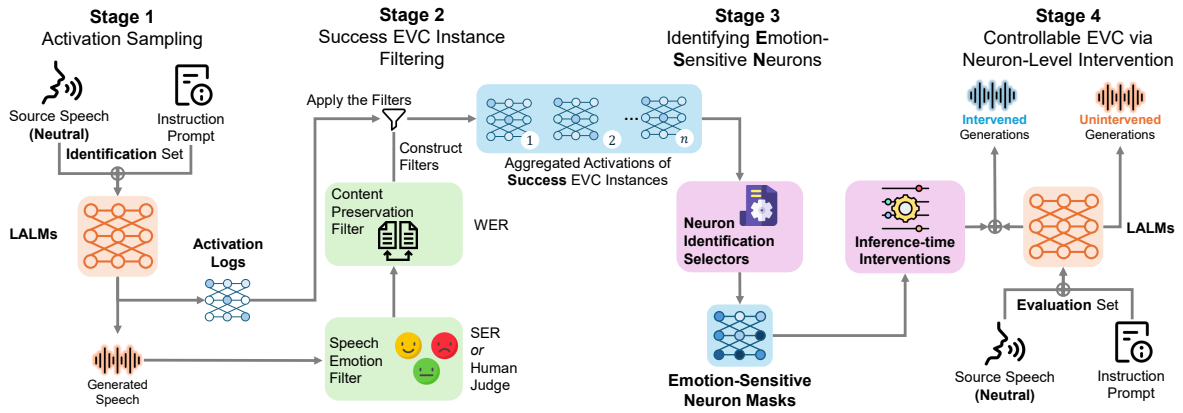


Figure 2: The overview of our 4-stage pipeline for identifying and manipulating emotion-sensitive neurons for EVC.

3.1 Stage 1: Activation Sampling

For each source utterance x in the identification set, we instruct the LALM to perform EVC and generate speech with target emotion e_{tgt}. The instruction prompt is fixed across conditions, and activations are recorded from modules involved in generation.

Hooking Position. We instrument the decoder-side feed-forward blocks (MLPs) because they offer a stable, high-bandwidth location for neuron-level interventions (Bau et al., 2019; Turner et al., 2024). Many LALMs employ gated MLPs (e.g., SwiGLU variants) (Shazeer, 2020), which offer a natural position for hooking. Concretely, let $u_{l,t}, v_{l,t} \in \mathbb{R}^{D_l}$ be the two pre-activation streams at layer l and decoding step/token position t, and let $g_{l,t} = \phi(u_{l,t}) \in \mathbb{R}^{D_l}$ denote the gated branch after the nonlinearity $\phi(\cdot)$ (e.g., SiLU). The MLP output is $h_{l,t} = W_o\,(g_{l,t} \odot v_{l,t})$, where $\odot$ is element-wise product. We log the scalar neuron activations $a_{l,n,t} = [g_{l,t}]_n$, for neuron index $n \in \{1, \ldots, D_l\}$. These logged values serve as the basis for our identification statistics and intervention hooks. In addition to the MLPs in the LM backbone of LALMs, we also examine whether the synthesis MLP layers contain ESNs, which we analyze later in §5.5.

3.2 Stage 2: Successful EVC Instance Filtering

A key design choice is to record activations only from conversions that plausibly satisfy EVC. As illustrated in Figure 1, raw generations contain diverse failure modes, and including them in activation statistics can obscure emotion-related signals. For each generated audio y, we therefore apply a two-axis filter aligned with the standard decomposition of voice conversion evaluation into speaking style and linguistic content (Zhou et al., 2022).

Emotion check via external judge. We employ automatic judge models, external to the generating LALM, to perform speech emotion recognition (SER): they predict an emotion label $\hat{e}(y)$. We accept a generation as emotion-matched if $\hat{e}(y) = e_{\text{tgt}}$.

Content preservation via transcript matching. We use the LALM-emitted decoded text as a text-side proxy for linguistic content preservation, comparing it with the source transcript of x. We note that this proxy does not independently verify the words actually realized in the generated waveform. After normalizing both texts with whisper-normalizer, we accept a generation as content-preserving if its word error rate (WER) is at most 0.15, its character error rate (CER) is at most 0.10, and its sequence-similarity ratio is at least 0.85. These thresholds are fixed across all experiments.

Success instance sets. For each target emotion $e \in \mathcal{E}$, we denote the filtered success set as $\mathcal{S}_e$. The size of this set is capped by a parameter c (the set is uniformly sampled if $|\mathcal{S}_e| > c$) to balance emotions (later analyzed in §5.2.3).

3.3 Stage 3: Identifying ESNs

3.3.1 Activation Statistics

Using only the filtered success set $\mathcal{S}_e$, our method computes neuron-wise activation probabilities from positive-activation counts. For each neuron (l, n) and emotion e, we accumulate

$$K^{(e)}_{l,n} \leftarrow K^{(e)}_{l,n} + \sum_{(x,y)\in\mathcal{S}_e} \sum_{t} \mathbb{I}(a^{(x\to y)}_{l,n,t} > 0), \quad (1)$$

$$T_e \leftarrow T_e + \sum_{(x,y)\in\mathcal{S}_e} \sum_{t} 1. \quad (2)$$

We then define the activation probability $P_{l,n}^{(e)} = \frac{K_{l,n}^{(e)}}{T_e}$. Thus, $P_{l,n}^{(e)}$ measures how often neuron (l, n) is active on successful EVC instances of target emotion e.

Let $N = \sum_{l=1}^{L} D_l$ denote the total number of neurons across all instrumented layers, and let $r \in (0, 1)$ be the selection rate. In all four methods below, we select $N_{\text{sel}} = \lfloor rN \rfloor$ neurons per emotion.

3.3.2 Neuron Ranking Methods

LAP (Activation Probability) (Huben et al., 2024; Gurnee et al., 2024; Voita et al., 2024). For each emotion e, neurons are ranked directly by their activation probability, i.e., $\text{LAP}_{l,n}^{(e)} = P_{l,n}^{(e)}$, and the top N_{sel} are selected.

LAPE (Activation Probability Entropy) (Tang et al., 2024; Huo et al., 2024; Namazifard and Poech, 2025). This method applies a two-stage procedure. First, for each neuron, activation probabilities are normalized across emotions, $\tilde{P}_{l,n}^{(e)} = P_{l,n}^{(e)} / \sum_{e'} P_{l,n}^{(e')}$, and an entropy score $H_{l,n} = -\sum_e \tilde{P}_{l,n}^{(e)} \log \tilde{P}_{l,n}^{(e)}$ is computed, where lower entropy indicates stronger emotion selectivity. Neurons that are globally too inactive are then removed using a probability threshold τ_p computed from a global percentile over all $P_{l,n}^{(e)}$ values; a neuron is kept only if it exceeds τ_p for at least one emotion. Among the remaining neurons, the lowest-entropy $N_{\text{cand}} = \lfloor \min(1, 5r)N \rfloor$ form a candidate pool. Finally, for each target emotion e, the method selects the top N_{sel} neurons from this pool by $P_{l,n}^{(e)}$.

MAD (Mean Activation Deviation) (Bau et al., 2019; Dalvi et al., 2019; Jorgensen et al., 2024). For each neuron, we first compute its mean activation probability across emotions, $\bar{P}_{l,n} = \frac{1}{|\mathcal{E}|} \sum_{e'} P_{l,n}^{(e')}$. For target emotion e, the score is then the absolute deviation from this mean: $\text{MAD}_{l,n}^{(e)} = \left| P_{l,n}^{(e)} - \bar{P}_{l,n} \right|$. Hence, this method favors neurons whose activation probability for emotion e differs strongly, in either direction, from their average behavior across emotions.

ConAct (Contrastive Activation Margin) (Zhao et al., 2026a). For each neuron, let $P_{l,n}^{(1)}$ and $P_{l,n}^{(2)}$ denote its largest and second-largest activation probabilities across emotions, respectively, and let $e_{l,n}^{(1)}$ be the corresponding top emotion. The margin is $\Delta_{l,n} = P_{l,n}^{(1)} - P_{l,n}^{(2)}$. For target emotion e, the method ranks a neuron by this margin only when $e = e_{l,n}^{(1)}$; otherwise, its score is set to 0. It then selects the top N_{sel} neurons for each emotion.

Random baseline. As a control, we additionally sample $N_{\text{sel}} = \lfloor rN \rfloor$ neurons uniformly at random without replacement from all N hooked neurons, and map them back to their layers to form a mask with the same sparsity as the other selectors.

3.4 Stage 4: Controllable EVC via Neuron-Level Intervention

To test whether the identified neurons are functionally involved in EVC behavior, we intervene on the post-activation gate values of hooked MLPs during generation, without modifying model weights. For a hooked layer l, let $g_{l,t} \in \mathbb{R}^{D_l}$ denote the activated gate vector at generation step t, and let $\mathcal{I}_l^{(m,e)}$ denote the selected neuron indices for method m and target emotion e. Interventions are applied only to dimensions in $\mathcal{I}_l^{(m,e)}$, before the elementwise product and output projection.

Gain scaling (steering) (Turner et al., 2024). For selected neurons, we multiply the activated gate values by a constant gain $1+\alpha$ ($\alpha \geq 0$); we refer to this multiplicative intervention simply as steering throughout the paper:

$$[\tilde{g}_{l,t}]_n = \begin{cases} (1+\alpha)[g_{l,t}]_n, & n \in \mathcal{I}_l^{(m,e)}, \\ [g_{l,t}]_n, & \text{otherwise.} \end{cases} \quad (3)$$

Additive shift (Rimsky et al., 2024). As an alternative intervention, we add a constant offset α ($\alpha \geq 0$) to the selected gate dimensions:

$$[\tilde{g}_{l,t}]_n = \begin{cases} [g_{l,t}]_n + \alpha, & n \in \mathcal{I}_l^{(m,e)}, \\ [g_{l,t}]_n, & \text{otherwise.} \end{cases} \quad (4)$$

Unlike steering, this shift is independent of the neuron's current activation.

Floor clamping (Templeton et al., 2024). We also test a minimum-activation constraint, in which each selected dimension is clamped from below by a scalar threshold α ($\alpha \geq 0$):

$$[\tilde{g}_{l,t}]_n = \begin{cases} \max\left([g_{l,t}]_n, \alpha\right), & n \in \mathcal{I}_l^{(m,e)}, \\ [g_{l,t}]_n, & \text{otherwise.} \end{cases} \quad (5)$$

This ensures that selected neurons remain at or above a fixed activation floor during generation.

Deactivation. Finally, we test necessity by zeroing the selected activated gate dimensions:

$$[\tilde{g}_{l,t}]_n = \begin{cases} 0, & n \in \mathcal{I}_l^{(m,e)}, \\ [g_{l,t}]_n, & \text{otherwise.} \end{cases} \quad (6)$$

This removes the contribution of the selected neurons while leaving the rest of the computation unchanged.

4 Experiment Setup

4.1 Models

As outlined, we evaluate three LALMs: Qwen2.5-Omni-7B (Xu et al., 2025a), MiniCPM-o 4.5 (Yao et al., 2024), and Kimi-Audio (KimiTeam, 2025), all of which are multimodal models with audio-generation capability. We use three automatic emotion evaluators. emotion2vec+large (Ma et al., 2024) serves as a high-accuracy acoustic SER judge. Qwen3-Omni-30B (Xu et al., 2025b) provides an LALM-based judge (Odoom and Koehn, 2026) and helps test whether the observed trends are specific to conventional SER classifiers. Our ESN masks are also identified from Qwen3-filtered successful EVC instances. We additionally evaluate with WavLM (Chen et al., 2022), a non-ESD-exposed acoustic SER judge. Since Qwen3 thus also serves as the identification filter, emotion2vec and WavLM are the evaluators independent of neuron selection.

On a 500-sample held-out subset of original ESD recordings (100 per emotion), emotion2vec+large, WavLM, and Qwen3-Omni-30B achieve 97.6%, 50.6%, and 47.4% overall SER accuracy, respectively. Because emotion2vec+large may have been exposed to ESD during its own training, we avoid treating its absolute emotion match rates as dataset-independent SER accuracy. Our main claims are based on matched relative comparisons between intervened and unintervened LALM-generated outputs under each evaluator, rather than on absolute emotion-match rates or on original ESD recordings. Key findings are further validated by human listening experiments (§5.4), including human-filtered ESN identification (Appendix B).

4.2 Dataset

Our primary test bed is the **Emotional Speech Database (ESD)** (Zhou et al., 2022), which contains 20 speakers, each with 350 parallel utterances per emotion across five categories ("neutral", "happy", "angry", "sad", "surprise"). To avoid language-dependent confounds, we restrict the dataset to the 10 English speakers. We further split the data by utterance index (parallel across emotions) to prevent lexical leakage across emotions. As summarized in Table 1, we use a hybrid split: a 6/2/2 speaker partition, with utterance-index splits within the seen-speaker portion. Unless noted otherwise, all reported results are computed on TEST-SEEN and TEST-UNSEEN. Neither test split shares any (speaker, utterance-index) pair with the identification or development splits, and TEST-UNSEEN additionally contains no identification or development speakers.

Speakers	S1–S6	S7–S8	S9–S10	**Total**
IDENTIFICATION	001–200 (200)	–	–	1200
DEVELOPMENT	201–250 (50)	001–250 (250)	–	800
TEST-SEEN	251–350 (100)	251–350 (100)	–	800
TEST-UNSEEN	–	–	001–350 (350)	700

Table 1: **Speaker and utterance-ID split of ESD.** Utterance IDs are parallel across the five emotions for each speaker. "Seen" speakers (S1–S8) are those used during identification or development. TEST-SEEN evaluates generalization to unseen utterances of seen speakers, while TEST-UNSEEN evaluates generalization to held-out speakers.

4.3 Task Implementation and Metrics

While EVC is inherently bidirectional, we focus on converting neutral speech to a target emotion $e_{\text{tgt}} \in \{\text{happy, angry, sad, surprise}\}$. For LALM-based conversion, we use a fixed prompt containing the input audio and target emotion instruction (Appendix A.3). We employ deterministic decoding when available (greedy, temperature=0). All generations are resampled to 16 kHz.

We evaluate three aspects of EVC: (1) **emotion shift**: emotion match rate, the fraction of samples whose predicted emotion matches e_{tgt}. Automatic SER judges are provided with $(x, \hat{x}_{e_{\text{tgt}}})$ and predict its emotion label, while human evaluation uses pairwise preference between baseline and intervened outputs (§5.4). (2) **content preservation**: WER between the reference transcript and the decoded text generated by the LALMs, normalized with whisper-normalizer.[2] (3) **naturalness**: perceptual quality estimated using UTokyo-SaruLab Mean Opinion Score (UTMOS) (Saeki et al., 2022).

	Qwen2.5-Omni-7B	MiniCPM-o 4.5	Kimi-Audio
Angry	0.13 / 10.93 / 0.13	3.73 / 12.93 / 3.27	**10.60 / 16.00 / 18.80**
Happy	**45.40 / 10.13 / 8.02**	2.13 / 5.33 / 1.67	39.60 / 7.87 / 2.41
Sad	**52.20** / 21.60 / 18.20	2.33 / 9.80 / 12.60	42.89 / **31.93** / **40.82**
Surprise	2.67 / 5.07 / 0.00	0.07 / 3.47 / 0.00	**7.13** / **6.67** / 0.00
Avg.	**25.10** / 11.93 / 6.59	2.07 / 7.88 / 4.39	25.06 / **15.62** / **15.51**
WER	19.00	2.95	**2.82**
UTMOS	4.00	**4.26**	3.18

Table 2: **Baseline EVC performance** of the three evaluated LALMs on ESD, before applying any neuron-level intervention. We report the three EVC axes used throughout the paper: target-emotion match rate in % (emotion2vec / Qwen3 / WavLM), content preservation (WER in %), and naturalness (UTMOS) on its 1–5 scale (higher is better).

5 Results

5.1 Baseline EVC Performance

Before applying neuron-level intervention, we establish the unintervened EVC baseline of the evaluated LALMs on ESD. As shown in Table 2, baseline emotion match rates are modest across all three models, confirming that instruction-following EVC remains challenging in our setting. Under the Qwen3 judge, Kimi-Audio achieves the highest average emotion match rate (15.62%), followed by Qwen2.5-Omni-7B (11.93%) and MiniCPM-o 4.5 (7.88%); the ranking differs under emotion2vec and WavLM, reflecting judge-dependent calibration.

Content preservation shows a contrasting trend: MiniCPM-o 4.5 and Kimi-Audio yield low WER (2.95% and 2.82%), while Qwen2.5-Omni-7B shows substantially higher WER (19.00%). Thus, emotion realization and content preservation are not aligned across models. This mismatch motivates our two-axis success filtering (§3.2): ESN identification must be based on conversions satisfying both emotion correctness and linguistic fidelity.

5.2 Sensitivity to Identification Parameters

5.2.1 Identification Methods

Building on the baseline EVC results in Table 2, we examine which selector yields the most effective ESN masks across the three evaluation axes. We compare the four selectors introduced in §3.3.2 under shared settings (c=50, r=0.5%, α=1.0). Table 3 shows a clear ranking in emotion specificity. For Qwen2.5-Omni-7B, ConAct and MAD produce the strongest positive self-effects and largest self–cross separation, whereas LAP and LAPE are weaker and less selective. In particular, LAPE exhibits diffuse behavior in which off-target effects can rival or exceed on-target gains. MiniCPM-o 4.5 shows changes within ± 0.5 pp under all selectors, too small to separate them, and for Kimi-Audio the ranking is judge-dependent: LAP yields the largest self-effect under emotion2vec, whereas ConAct does under Qwen3. Emotion gains often come at the cost of content preservation, especially for Qwen2.5-Omni-7B, where the largest WER increases occur for LAPE, ConAct, and MAD. By contrast, UTMOS remains largely stable across settings, suggesting that interventions primarily induce semantic drift rather than acoustic degradation: outputs remain natural-sounding, but may deviate from the instructed content. Overall, selector comparison reflects a trade-off between target-emotion control and unintended content changes, with ConAct achieving the best balance on Qwen2.5-Omni-7B. We therefore adopt ConAct as the default selector in subsequent experiments. Throughout, α=1.0 is used as a probing strength that makes effects clearly visible rather than as a recommended operating point; §5.3.2 shows that α=0.3–0.5 is preferable when content preservation matters. We refer to this configuration (ConAct, c=50, r=0.5%, steering, α=1.0) as the default setting.

5.2.2 Selection Rate

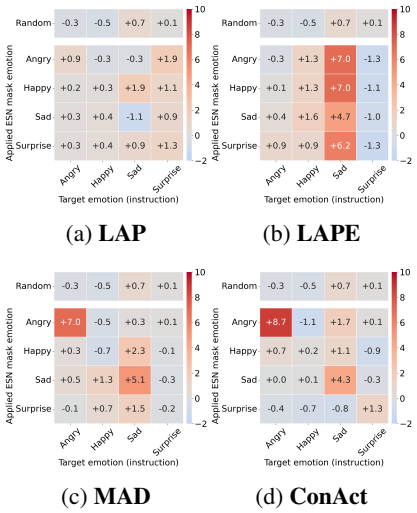
We next examine the selection rate r, which controls mask sparsity and thus the balance between specificity (small r) and coverage (large r).

Figure 4 suggests a largely monotonic trend within the tested range: with r=0.1% the effect is present but weak and sparse (consistent with under-selection), while increasing r to a moderate range, 0.3% and 0.5% (in Figure 3(d)) progressively strengthens the diagonal/self-effect and clarifies the specificity pattern. However, when r reaches 1.0%, off-diagonal changes become more visible (e.g., +4.4 pp for Sad with a Happy-mask). This behavior supports that ESNs in LALMs are distributed rather than ultra-sparse, hence selecting too few neurons leaves part of the relevant signal unused. Since our sweep only extends to r=1.0%, we interpret this monotonicity as holding within the explored range and retain r=0.5% as a practical default balancing compactness and effect strength.

[2] https://pypi.org/project/whisper-normalizer/

LALM	Selector	ΔEmotion Match vs. Base (emotion2vec / Qwen3)			WER ↓ (Δ)	UTMOS ↑ (Δ)
		Self-Effect ↑	Cross-Effect Avg. (→ 0)	Self–Cross Gap ↑		
Qwen2.5-Omni-7B	Random	+0.52 / +0.01	–	–	22.54 (+3.54)	3.99 (-0.01)
	LAP	-2.15 / +0.04	-1.89 / +0.07	-0.25 / -0.03	**18.94 (-0.06)**	3.92 (-0.08)
	LAPE	+1.82 / +1.07	+2.21 / +1.83	-0.39 / -0.77	34.70 (+15.70)	3.94 (-0.06)
	MAD	+2.62 / +2.82	+0.81 / **+0.05**	+1.81 / +2.30	24.22 (+5.22)	**3.99 (-0.01)**
	ConAct	**+4.27 / +3.60**	**+0.34 / -0.05**	**+3.93 / +3.65**	29.65 (+10.65)	**3.99 (-0.01)**
	ConAct-H	+3.15 / +2.92	+0.62 / +0.11	+2.53 / +2.81	24.56 (+5.56)	**3.99 (-0.01)**
MiniCPM-o 4.5	Random	-0.05 / +0.14	–	–	3.14 (+0.19)	4.25 (-0.01)
	LAP	-0.33 / -0.38	-0.42 / -0.16	**+0.09** / -0.23	8.98 (+6.03)	4.25 (-0.01)
	LAPE	-0.28 / +0.33	-0.18 / +0.34	-0.10 / -0.01	**2.81 (-0.14)**	4.25 (-0.01)
	MAD	-0.22 / +0.33	-0.17 / +0.16	-0.05 / -0.18	2.84 (-0.11)	**4.26 (0.00)**
	ConAct	**-0.20 / +0.40**	**-0.08 / +0.09**	-0.12 / **+0.31**	2.95 (0.00)	4.25 (-0.01)
Kimi-Audio	Random	-0.76 / -0.12	–	–	3.72 (+0.90)	**3.17 (-0.01)**
	LAP	**+1.52** / +2.71	+0.35 / +1.49	+1.17 / +1.22	4.13 (+1.31)	3.15 (-0.03)
	LAPE	-0.72 / -0.48	-0.89 / **+0.46**	+0.16 / -0.94	**2.90 (+0.08)**	3.16 (-0.02)
	MAD	-0.90 / +1.17	**-0.02** / +1.15	-0.88 / +0.02	4.69 (+1.87)	3.15 (-0.03)
	ConAct	+1.19 / **+7.65**	-0.19 / +0.96	**+1.38 / +6.70**	4.74 (+1.92)	3.16 (-0.02)

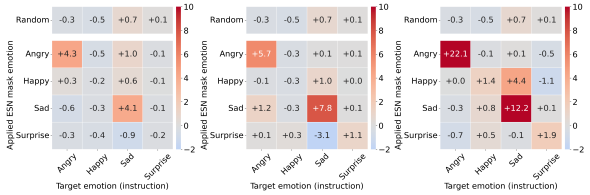
Table 3: **Comparison of ESN selectors under steering** (c=50, r=0.5%, α=1.0). For each selector, we report changes in target-emotion match relative to the unintervened baseline in Table 2, together with post-intervention WER and UTMOS (changes in parentheses). ConAct-H uses human-filtered instances for identification (Appendix B). Bold marks the best value per model and judge; for cross-effect, the value closest to zero.



(a) **LAP** (b) **LAPE**

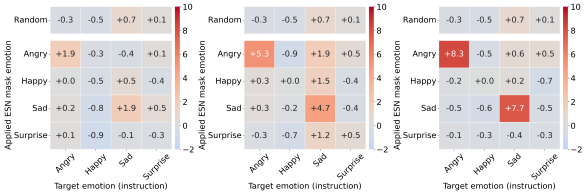
(c) **MAD** (d) **ConAct**

Figure 3: **Per-emotion emotion match rate changes by ESN selector for Qwen2.5-Omni-7B**, under the setting of Table 3 and evaluated by the Qwen3 judge (as are all heatmaps in this paper). Rows denote the applied ESN mask emotion (e_{mask}) and columns the instructed target emotion (e_{tgt}); entries are signed changes (pp) in emotion match rate relative to the unintervened baseline. Diagonal cells are **self-effects** (e_{mask}=e_{tgt}), and each off-diagonal cell is a single **cross-effect** pair ($e_{\text{mask}}\neq e_{\text{tgt}}$); the Cross-Effect Avg. column in Table 3 averages these off-diagonal cells.



(a) r = 0.1% (b) r = 0.3% (c) r = 1%

Figure 4: **Sensitivity to ESN selection rate** r (Qwen2.5-Omni-7B, ConAct, c=50, steering, α=1.0).



(a) c=10 (b) c=20 (c) c=100

Figure 5: **Sensitivity to # success instance** c (Qwen2.5-Omni-7B, ConAct, r=0.5%, steering, α=1.0). This sweep uses human-filtered success instances (Appendix B), of which 100 per emotion are available, so the cap c is binding in every panel.

5.2.3 Success Set Size

We then vary the success instance size c, the maximum number of filtered EVC successes retained per target emotion for activation aggregation. c changes the amount of evidence used to estimate the neuron statistics in §3.3.1, so it is expected to affect the stability of ESN ranking.

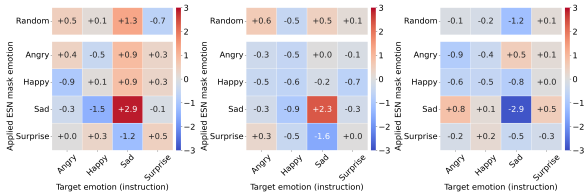
To keep the number of available instances fixed across caps, this sweep identifies ESNs from human-filtered success instances (Appendix B). Figure 5 shows a clear trend. With a very small cap (c=10), self-effects are weak or absent (at most +1.9 pp), indicating unstable ESN estimation. At c=20, the angry and sad self-effects rise to +5.3 and +4.7 pp, but off-diagonal changes of up to +1.9 pp also appear. At c=100, the self-effects strengthen further (+8.3 and +7.7 pp) while all off-diagonal changes fall below 1 pp, suggesting that more qualified instances yield both stronger and cleaner emotion-specific ESN sets within the tested range. Happy and surprise remain near zero at every cap, consistent with their weaker signals elsewhere in the paper.

5.3 Inference-Time Intervention Design

Having established that identification parameters can affect ESNs' emotion control effects, we next analyze how inference-time intervention design affects the EVC outcome.

5.3.1 Intervention Methods

We compare alternative ESN interventions beyond the steering used in our main experiments. Figure 6 shows that additive and clamping interventions



(a) Add, α**=0.30** (b) Clamp, α**=0.10** (c) Deactivation

Figure 6: **Comparison of ESN intervention methods** (Qwen2.5-Omni-7B, ConAct, c=50, r=0.5%). As explained in §3.4, α is method-specific, so we use different values to keep intervention strength roughly comparable across methods. Deactivation is parameter-free and hence reported without α.

can induce partial target-emotion effects, but fail to produce consistent positive self-effects across emotions. Deactivation provides two additional insights. First, merely suppressing non-target-emotion activations is insufficient to reliably induce the desired expressive changes. Second, the consistently negative diagonal under deactivation further supports the functional relevance of ESNs: silencing identified neurons impairs rather than improves target-emotion realization, suggesting that these neurons participate in emotion-related generation under the tested intervention setting. Based on these results, we retain steering as the default intervention in subsequent analyses.

5.3.2 Intervention Strength

Taking steering as a probe method, we report the sensitivity of intervention effects to the intervention strength α in Table 4; per-emotion heatmaps are given in Appendix C. Within the tested range, increasing α generally strengthens the intended self-effect on emotion match, especially for the stronger steering settings. Off-diagonal cross-effect, on the other hand, does not monotonically expand with α but rather fluctuates, leading to widening self–cross gaps.

	ΔEmotion Match vs. Base (emotion2vec / Qwen3)				
α	Self-Effect ↑	Cross-Effect Avg. (→ 0)	Self–Cross Gap ↑	WER ↓ (Δ)	UTMOS ↑ (Δ)
0.3	+0.78 / +0.37	+0.48 / -0.09	+0.30 / +0.45	**19.33 (+0.33)**	**4.00 (0.00)**
0.5	+2.00 / +1.13	**-0.11** / -0.22	+2.11 / +1.36	20.67 (+1.67)	**4.00 (0.00)**
1.0	+4.27 / +3.60	+0.34 / -0.05	+3.93 / +3.65	29.65 (+10.65)	3.99 (-0.01)
2.0	**+8.89 / +20.67**	-0.14 / **+0.02**	**+9.02 / +20.64**	203.65 (+184.65)	3.91 (-0.09)

Table 4: **Emotion match and WER trade-off when increasing the intervention strength** (Qwen2.5-Omni-7B, ConAct, c=50, r=0.5%, steering). Bold marks the best value per column and judge; for cross-effect, the value closest to zero. WER can exceed 100% because insertions from long hallucinated or repetitive outputs count as errors.

The largest α further highlights the trade-off with content preservation: α=2.0 yields the strongest emotion match gains and self–cross gap, but catastrophically harms content preservation and lowers naturalness. In contrast, moderate steering (α=0.3–0.5) keeps WER near the baseline and leaves UTMOS unchanged, while still providing measurable improvements in target emotion match.

5.4 Human Evaluation

To complement automatic SER evaluation, we conduct a human listening study under the default setting (Figure 3(d)). In each trial, participants listen to a pair of speech samples and select the clip that better matches a target emotion, with an additional "indecisive" option. Each pair compares a baseline output against its ESN-intervened counterpart generated from the same source speech and instruction. We collect responses from 20 English-proficient participants on a shared set of 100 pairs (25 per emotion), sampled uniformly at random from the parallel baseline/intervened generations on the test split. The A/B order within each pair and the order of pairs are randomized, clips are volume-normalized, and raters are blind to the condition. As shown in Figure 7, listeners prefer the intervened system across all emotions, with average win/tie/loss rates of 62%/11%/27%. Gains are strongest for happy (69% win) and remain favorable for sad, surprise, and angry, supporting that ESN steering produces perceptible emotion shifts.

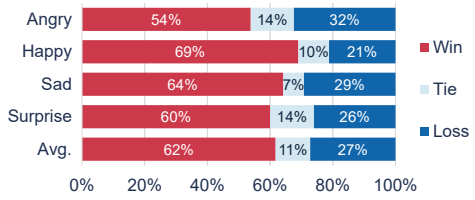


Figure 7: **Human listening evaluation under the default setting.** Win/Tie/Loss rates comparing the intervened system (Qwen2.5-Omni-7B, ConAct, c=50, r=0.5%, steering, α=1.0) against the unintervened baseline. 20 participants.

5.5 Localization of ESNs

We also examine whether ESNs reside in downstream synthesis modules; this analysis is restricted to Qwen2.5-Omni-7B, whose speech synthesis module (the Talker) is a separate autoregressive decoder with its own MLP layers. As shown in Table 5, applying the same identification-intervention pipeline to the Talker MLPs largely removes the

positive self-effects observed in decoder-side LM backbone MLPs. Across selectors, self–cross gaps become near-zero or negative under both SER judges, indicating that synthesis-side interventions are mostly inert rather than emotion-specific. Note that WER is computed on the text emitted by the LM backbone (Thinker), which Talker interventions cannot change; the WER and UTMOS columns of Table 5 are therefore reported for completeness only, and this conclusion rests on the emotion columns. These results suggest that the most actionable ESNs reside in LM decoder MLPs rather than downstream synthesis modules, making LM-side feed-forward blocks a more effective target for controllable affect steering. We further provide a layer-wise distribution analysis of ESNs in Appendix D.

	ΔEmotion Match vs. Base (emotion2vec / Qwen3)				
Selector	Self-Effect ↑	Cross-Effect Avg. (→ 0)	Self–Cross Gap ↑	WER ↓ (Δ)	UTMOS ↑ (Δ)
Random	+0.52 / +0.02	–	–	19.21 (+0.21)	4.01 (+0.01)
LAP	**+0.32** / -0.08	**+0.12** / +0.23	**+0.20** / -0.32	**19.00 (0.00)**	3.96 (-0.04)
LAPE	-0.87 / -0.22	-0.13 / -0.27	-0.73 / **+0.06**	**19.00 (0.00)**	**4.01 (+0.01)**
MAD	-0.68 / **+0.08**	+0.51 / **+0.13**	-1.19 / -0.04	**19.00 (0.00)**	4.00 (0.00)
ConAct	+0.12 / -0.42	+0.45 / -0.19	-0.33 / -0.22	19.19 (+0.19)	3.99 (-0.01)

Table 5: **Steering in the speech synthesis module (Talker) of Qwen2.5-Omni-7B.** Results obtained by applying the same settings (c=50, r=0.5%, α=1.0) as in Table 3, but with ESNs identified in, and interventions applied to, the synthesis-module MLP layers. Bold as in Table 3.

5.6 Do Intervention Effects Generalize to Unseen Speakers?

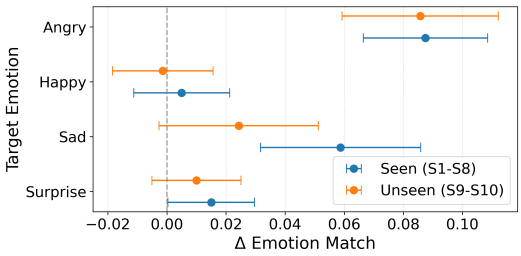


Figure 8: **Intervention effect on seen and unseen speakers.** Changes in emotion match rate (as a fraction; 0.02 corresponds to 2 pp) under intervention (Qwen2.5-Omni-7B, ConAct, c=50, r=0.5%, steering, α=1.0) relative to the unintervened baseline, evaluated by the Qwen3 judge, for seen speakers (TEST-SEEN) and unseen speakers (TEST-UNSEEN), with 95% CI.

Finally, we evaluate whether ESN-based control generalizes beyond speakers used during neuron identification. Figure 8 shows that the intervention effect partially transfers to unseen speakers, with strongly emotion-dependent reliability. Angry is the only emotion whose gain is clearly positive on both splits, indicating a relatively speaker-robust signal. Sad is clearly positive on seen speakers but drops on unseen speakers, where its interval includes zero, suggesting partial speaker-specificity. Surprise remains positive in point estimate on both splits, but its intervals include zero. Happy shows no reliable gain on either split. A practical implication is that speaker-robust ESN control is feasible, but likely benefits from more diverse identification data.

6 Conclusion

We presented a systematic neuron-level study of emotion control in speech-generative LALMs for emotional voice conversion. Our results show that compact ESNs are interventionally actionable and support inference-time emotion steering for the evaluated models. Meaningful ESN identification requires success-filtered activation aggregation that satisfy the multi-objective nature of EVC, while contrastive-margin- and mean-deviation-based selectors yield the most emotion-specific control. We further identify a trade-off between emotion realization and content preservation: moderate interventions improve target-emotion match, whereas stronger steering increases WER through semantic drift despite stable naturalness. Identified ESNs concentrate in intermediate-to-late decoder MLP layers of the LM backbone, and interventions in the downstream synthesis module remain largely ineffective. Together, these findings provide an intervention-based analysis framework for training-free emotion control and suggest broader potential for inference-time manipulation of paralinguistic attributes in speech-generative systems.

Limitations

A limitation of our evaluation is that emotion2vec+large, one of our automatic SER judges, may have been exposed to ESD during its own training. Therefore, although it provides a strong acoustic emotion recognition signal, its absolute emotion match rates in ESD-related experiments should be interpreted with caution. To reduce dependence on a potentially overlapping evaluator, we additionally use WavLM, an acoustic SER judge not exposed to ESD, and Qwen3-Omni-30B, an LALM-based judge. These two evaluators have lower accuracy on original ESD recordings than

emotion2vec+large, and are therefore noisier, but they provide complementary validation signals.

This possible evaluator overlap does not change our central conclusions. All intervention results are computed on LALM-generated speech rather than on original emotional ESD recordings, and we do not train or fine-tune the evaluated LALMs. Our claims rely primarily on matched relative comparisons between intervened and unintervened generations under identical source utterances, prompts, target emotions, and evaluation protocols, and the main patterns are further supported by human listening and human-filtered ESN identification.

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

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*.

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*, 1:1–19.

Yuexin Cao, Zhengchen Liu, Minchuan Chen, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Nonparallel emotional speech conversion using vae-gan. In *Interspeech 2020*, pages 3406–3410.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee. 2024. EmoSphere-TTS: Emotional Style and Intensity Modeling via Spherical Emotion Vector for Controllable Emotional Text-to-Speech. In *Interspeech 2024*, pages 1810–1814.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. 2025. Steering off course: Reliability challenges in steering language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19856–19882, Vienna, Austria. Association for Computational Linguistics.

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.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *Preprint*, arXiv:2410.00037.

Junfeng Fang, Zongze Bi, Ruipeng Wang, Houcheng Jiang, Yuan Gao, Kun Wang, An Zhang, Jie Shi, Xiang 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.

Changzeng Fu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2022. An improved cyclegan-based emotional voice conversion model by augmenting temporal dependency with a transformer. *Speech Commun.*, 144(C):110–121.

Lucas Goncalves, Ali N. Salman, Abinay Reddy Naini, Laureano Moro-Velázquez, Thomas Thebaud, Paola Garcia, Najim Dehak, Berrak Sisman, and Carlos Busso. 2024. Odyssey 2024 - Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 247–254.

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*.

Xiangheng He, Junjie Chen, Georgios Rizos, and Björn W. Schuller. 2021. An improved stargan for emotional voice conversion: Enhancing voice quality and data augmentation. In *Interspeech 2021*, pages 821–825.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pages 7827–7845.

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.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. 2024. Improving activation steering in language models with mean-centring. In *Responsible Language Models Workshop (ReLM) at AAAI-24*. Responsible Language Models Workshop at AAAI-24, ReLM@AAAI-24 ; Conference date: 26-02-2024 Through 26-02-2024.

KimiTeam. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu Anh Nguyen, Morgan Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2022. Textless speech emotion conversion using discrete & decomposed representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11200–11214, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.

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.

Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie. 2022. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:853–864.

Yi Lei, Shan Yang, and Lei Xie. 2021. Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 423–430. IEEE.

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.

Jaime Lorenzo-Trueba, Gustav Eje Henter, Shinji Takaki, Junichi Yamagishi, Yosuke Morino, and Yuta Ochiai. 2018. Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis. *Speech Communication*, 99:135–143.

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.

Danial Namazifard and Lukas Galke Poech. 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.

Bismarck Bamfo Odoom and Philipp Koehn. 2026. Leveraging large speech language models as evaluators for expressive speech. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16437–16441.

Navin Raj Prabhu, Bunlong Lay, Simon Welker, Nale Lehmann-Willenbrock, and Timo Gerkmann. 2024. Emoconv-diff: Diffusion-based speech emotion conversion for non-parallel and in-the-wild data. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11651–11655.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech 2022*, pages 4521–4525.

Bastian Schnell and Philip N. Garner. 2021. Improving emotional tts with an emotion intensity input from unsupervised extraction. In *11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 60–65.

Noam Shazeer. 2020. Glu variants improve transformer. *Preprint*, arXiv:2002.05202.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. 2026. Emotion concepts and their function in a large language model. *Preprint*, arXiv:2604.07729.

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.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Andreas Triantafyllopoulos, Björn W. Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, Ruibo Fu, and Jianhua Tao. 2023. An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE*, 111(10):1355–1381.

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.

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.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. AudioBench: A universal benchmark for audio large language models. 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 4297–4316, Albuquerque, New Mexico. 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. 2026a. Do LLMs "feel"? emotion circuits discovery and control. In *Forty-third International Conference on Machine Learning*.

Siyi Wang, Shihong Tan, Siyi Liu, Hong Jia, Gongping Huang, James Bailey, and Ting Dang. 2026b. Cocoemo: Composable and controllable human-like emotional tts via activation steering. *Preprint*, arXiv:2602.03420.

Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, pages 5180–5189. PMLR.

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.

Pengfei Wu, Zhenhua Ling, Lijuan Liu, Yuan Jiang, Hongchuan Wu, and Lirong Dai. 2019. End-to-end emotional speech synthesis using style tokens and semi-supervised training. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 623–627. IEEE.

Tianxin Xie, Shan Yang, Chenxing Li, Dong Yu, and Li Liu. 2025. Emosteer-tts: Fine-grained and training-free emotion-controllable text-to-speech via activation steering. *Preprint*, arXiv:2508.03543.

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. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Chih-Kai Yang, Neo Ho, Yi-Jyun Lee, and Hung-Yi Lee. 2025a. 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.

Chih-Kai Yang, Neo S. Ho, and Hung-yi Lee. 2025b. Towards holistic evaluation of large audio-language models: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10144–10170, Suzhou, China. Association for Computational Linguistics.

Zijiang Yang, Xin Jing, Andreas Triantafyllopoulos, Meishu Song, Ilhan Aslan, and Björn W. Schuller. 2022. An Overview & Analysis of Sequence-to-Sequence Emotional Voice Conversion. In *Interspeech 2022*, pages 4915–4919.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

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.

Guangyan Zhang, Ying Qin, Wenjie Zhang, Jialun Wu, Mei Li, Yutao Gai, Feijun Jiang, and Tan Lee. 2023. iemotts: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1693–1705.

Zihan Zhang, Shreeram Suresh Chandra, Zongyang Du, Xiutian Zhao, Aurosweta Mahapatra, Hao Zhang, Philipp Koehn, and Berrak Sisman. 2026. Trace-evc: Text-guided relative affective control for zero-shot emotional voice conversion. *Preprint*, arXiv:2607.03666.

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, Philipp Koehn, Björn Schuller, and Berrak Sisman. 2026b. Multilingual emotion neurons in large audio-language models. *Preprint*, arXiv:2608.08772.

Xiutian Zhao, Björn Schuller, and Berrak Sisman. 2026c. Discovering and causally validating emotion-sensitive neurons in large audio-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15056–15071, San Diego, California, United States. Association for Computational Linguistics.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional voice conversion: Theory, databases and esd. *Speech Commun.*, 137(C):1–18.

Li Zhou, Hao Jiang, Junjie Li, Tianrui Wang, and Haizhou Li. 2026. Emoshift: Lightweight activation steering for enhanced emotion-aware speech synthesis. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17262–17266.

A Reproducibility Details

A.1 Computational Resources

All experiments were conducted on a shared GPU cluster using NVIDIA A100 and H100 GPUs, with access to up to 8 GPUs of each type. The total GPU consumption, including pilot studies, failed runs, and debugging, was approximately 1846 A100-hours and 1146 H100-hours. The main computational cost came from activation logging and inference-time intervention experiments for the three evaluated LALMs under different ESN identification methods, selection rates, success-set sizes, intervention strategies, and intervention strengths. Activation logging required additional storage and memory because decoder-side MLP activations were recorded during speech generation for subsequent neuron selection. Depending on model size and generation configuration, each LALM required approximately 40–80 GB of GPU memory. Reproducing only the final reported experiments is estimated to require approximately 1200 A100-hours.

A.2 Datasets and Models

Models	Sources
Evaluated LALMs	
Qwen2.5-Omni-7B	https://huggingface.co/Qwen/Qwen2.5-Omni-7B
MiniCPM-o 4.5	https://huggingface.co/openbmb/MiniCPM-o-4_5
Kimi-Audio	https://huggingface.co/moonshotai/Kimi-Audio-7B-Instruct
SER judges	
emotion2vec+large	https://huggingface.co/emotion2vec/emotion2vec_plus_large
Qwen3-Omni-30B	https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct
WavLM	https://huggingface.co/3loi/SER-Odyssey-Baseline-WavLM-Categorical

Table 6: Public sources of the employed models. The WavLM (Chen et al., 2022) evaluator adapts the baseline categorical system of the Odyssey 2024 SER Challenge (Goncalves et al., 2024) and is fine-tuned on MSP-Podcast (Busso et al., 2026) rather than ESD.

A.3 Prompt Templates

Listing 1: Prompt template used for EVC generation.

```
Repeat the audio word-for-word exactly.
Speak with {target_emotion} emotion.
Do not add or remove any words. Output speech
    only.
```

B Human Listener Participation

Demographics	
Total Participants	20
Age	[21, 32]
Sex (Female / Male)	7 / 13
English Proficiency	Proficient
Hearing Status	Normal

Table 7: The demographic information of the listening test participants.

In the human listening test, all participants took part voluntarily and provided informed consent before the annotation. Participants were English-proficient listeners. We conducted two human-supervision procedures: an emotion-preference test, used for the pairwise evaluation in §5.4, and a separate success-filtering annotation used for ConAct-H ESN identification. In the emotion-preference test, participants judged only which sample better matched the target emotion, as specified in Listing 2. In the success-filtering annotation, participants judged whether a generated sample both matched the target emotion and preserved the source linguistic content. The resulting human-filtered successful instances were then used to identify ESNs, following the same activation aggregation and ConAct selection procedure as in the automatic-filtering setting.

B.1 Emotion-Preference Evaluation

For the human emotion-preference evaluation, participants were given the following instructions.

Listing 2: Emotion Match Preference Test

```
Task:
You will be shown a target emotion and will
    listen to two speech clips, Sample A and
    Sample B.

Please listen to both clips and select the one
    that better matches the target emotion.
    Focus on the emotional expressiveness of the
     speech. You do not need to judge whether
    the linguistic content is correct.

Response options:
- Enter A if Sample A better matches the target
    emotion.
- Enter B if Sample B better matches the target
    emotion.
- Enter N if there is no clear difference
    between the two samples, or if neither
    sample better matches the target emotion.

Notes:
- Some clips may take a few seconds to load. If
    you do not hear anything immediately after
    clicking, please wait briefly and then try
    again.
- After completing the task, please return
    either the completed annotation file or a
    text file containing your responses in the
    provided template.
```

B.2 Human-Supervised ESN Identification

Emotion	Jaccard	Direct Match	JSD
Angry	0.217	35.71%	0.009
Happy	0.075	13.95%	0.040
Sad	0.109	19.57%	0.010
Surprise	0.071	13.27%	0.025
Average	0.118	20.63%	0.021

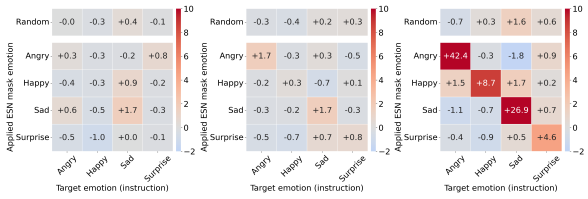
Table 8: **Similarity between ESNs identified from human- vs. SER-filtered instances.** Jaccard overlap is $|A \cap B|/|A \cup B|$ and direct match rate is $|A \cap B|/|A|$, where A and B are the human- and SER-filtered ESN sets; the Jensen–Shannon divergence (JSD, base 2) is computed between the two sets' layer-wise neuron-count distributions.

To assess the consistency between automatic and human supervision during ESN identification, we further compare neuron sets obtained from SER-filtered versus human-filtered success instances. Table 8 shows that overlap at the neuron level is partial, indicating that the two filtering strategies select different specific units. However, the layer-wise distributions between the two are highly similar

(average JSD 0.021), suggesting that both supervision signals concentrate ESNs in similar MLP layers. Agreement is strongest for angry (Jaccard 0.217, JSD 0.009), whereas happy and surprise exhibit lower overlap, consistent with their weaker or more ambiguous emotion signals.

The ConAct-H row in Table 3 replaces SER-filtered successes with human-judged ones for ESN identification, and yields the same qualitative behavior: positive self-effects with limited cross-emotion spillover. Compared to ConAct, ConAct-H slightly reduces self-effect and self–cross gap, but also substantially lowers ΔWER, suggesting that human supervision preferentially retains content-faithful successes and produces a more conservative yet robust ESN set for controllable EVC.

C Intervention Strength Heatmaps



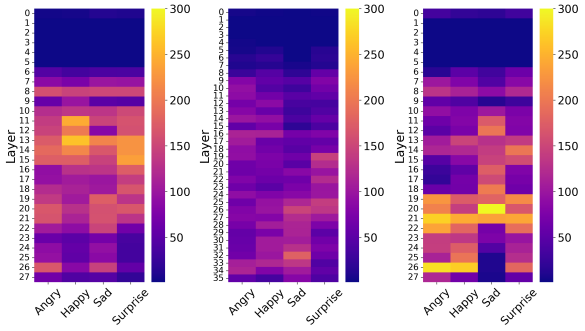
(a) α**=0.30** (b) α**=0.50** (c) α**=2.00**

Figure 9: **Sensitivity to intervention strength** α (Qwen2.5-Omni-7B, ConAct, c=50, r=0.5%, steering). The α=1.0 heatmap is Figure 3(d).

Figure 9 shows the per-emotion emotion match changes underlying Table 4. Increasing α strengthens the diagonal self-effects, most visibly for angry and sad, while off-diagonal changes fluctuate without a consistent monotonic trend.

D Layer-Wise Distribution of ESNs

We visualize the concentrations of the ESNs across the MLP layers in Figure 10, which shows that ESNs are not uniformly distributed. Across all three LALMs, they concentrate primarily in intermediate-to-late layers of the language-model decoder, with relatively few units selected in the earliest layers. For Qwen2.5-Omni-7B, ESNs cluster most strongly in approximately layers 11–15, with pronounced peaks for happy and surprise. MiniCPM-o 4.5 exhibits a more diffuse distribution but still places greater mass in the latter half of the decoder. Kimi-Audio shows the strongest late-layer concentration, particularly in layers 19–22, with marked peaks for happy and angry in layer 26. This consistent pattern suggests that, in the tested models, emotion-relevant control signals are more prominently captured in intermediate-to-late language-model layers than in earlier layers.



(a) Qwen2.5-Omni (b) MiniCPM-o 4.5 (c) Kimi-Audio

Figure 10: **Layer-wise distribution of ConAct-identified ESNs across three LALMs** (r=0.5%). Heatmaps show the number of selected ESNs per decoder layer and target emotion. Qwen2.5-Omni-7B and Kimi-Audio use 28-layer decoders, while MiniCPM-o 4.5 uses a 36-layer decoder. Values indicate where selected neurons are concentrated, not intervention effect size.

E License

All datasets and pre-trained models used in this work are publicly available and were used in accordance with their respective licenses, terms of use, and intended research purposes. The code files are released for academic research use. Users of the artifacts are responsible for complying with the licenses and usage restrictions of the underlying datasets, model checkpoints, and third-party evaluation tools.

F Generative AI Use Disclosure

Generative AI tools were employed solely for language polishing of text written by the authors. These tools were not used to generate scientific content, results, experimental designs, analyses, or conclusions.