LEVERAGING CONTRASTIVE CUES FOR PARALINGUISTIC UNDERSTANDING IN SPEECH LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vocal delivery shapes meaning beyond words, but whether speech large language models (Speech LLMs) capture these distinctions remains unclear. To isolate this ability, we construct PARAtaxis from same-transcript, same-speaker speech pairs for controlled evaluation and multitask training. Nine pretrained models reliably distinguish lexical differences but perform near chance on delivery differences. Layerwise analysis of Qwen2.5-Omni and Audio Flamingo 3 reveals a similar representational gap; stronger delivery distinctions after supervised fine-tuning accompany better behavioral accuracy. Motivated by this association, we introduce PARANCE, which jointly optimizes next-token prediction and bidirectional contrastive matching on final-layer last-token prefill representations, explicitly including the verified paralinguistic counterpart as a hard negative. On Qwen2.5-Omni-7B, PARANCE improves MMSU overall accuracy by 5.56 percentage points and Perception by 6.28 points over the pretrained model, while also improving Reasoning. The same adaptation also improves accuracy over five emotion-recognition datasets without emotion-category supervision. Further evaluation shows cross-task transfer across all five emotion-recognition datasets, despite the absence of emotion-category supervision during adaptation. These results support learning foundational vocal distinctions as a route to broader paralinguistic understanding, even when the downstream capability is not directly supervised during adaptation. Our anonymous repository is available at `https://anonymous.4open.science/r/ParaNCE-139B/`.

1 INTRODUCTION

Vocal delivery conveys emphasis, attitude, and affect beyond the words themselves (Wagner & Watson, 2010). Understanding these signals is an important capability for speech large language models (Speech LLMs), whose perception and reasoning increasingly extend beyond transcription (Wang et al., 2026). Yet controlled evaluations reveal a persistent dependence on lexical content: acoustic cues can be recoverable from hidden states while remaining weakly reflected in model outputs (Chen et al., 2026; Pang et al., 2026). This raises a concrete question: *how do Speech LLMs distinguish different deliveries of the same words in their representations and responses, and can these distinctions guide learning?*

Answering this question requires contrasts whose correspondence cannot be resolved from the transcript alone. We construct such contrasts by pairing speech realizations that share the same transcript and speaker but differ in acoustically verified delivery. Operational acoustic criteria for speaking rate, energy, pitch, and intonation verify delivery differences in waveforms produced by controllable speech synthesis. The resulting **PARAtaxis** resource organizes verified pairs into seven task forms, with 11,272 training examples. Its held-out diagnostic set, **PARAtaxis-Diag**, contains 500 triplets that pair a shared reference with a prosodic contrast and a lexical contrast, enabling comparison under the same speech–caption matching formulation.

We examine these correspondences through both decoded choices and prefill representations. Across nine pretrained Speech LLMs, bidirectional selection reliably resolves lexical differences but remains near chance for different deliveries of the same words. Layerwise analysis of Qwen2.5-Omni and Audio Flamingo 3 reveals a parallel asymmetry: their representations favor matched speech–caption pairs much more strongly over lexical mismatches than over prosodic mismatches. Dense supervised fine-tuning on PARAtaxis improves both upper-layer *matching advantage* and behavioral

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

discrimination. This association motivates an explicit learning signal that strengthens the model's preference for the correct correspondence when lexical content is uninformative.

We introduce **PARANCE**, an auxiliary bidirectional speech–caption contrastive objective optimized jointly with next-token prediction. For each designated pair, the matched speech and caption form a positive correspondence, while the other member supplies a same-transcript, same-speaker hard negative. Including this counterpart explicitly in both matching directions makes verified delivery differences relevant to the training objective, beyond the distinctions supplied by randomly co-occurring negatives. We apply the objective to the last prefill token at the final decoder layer, directly supervising the matching preference in the model's final contextualized representation.

Experiments show improvements in both controlled discrimination and downstream understanding. On Qwen2.5-Omni-7B, PARANCE improves MMSU Perception by 6.28 percentage points over the pretrained model, exceeding Dense SFT and the InfoNCE baseline by 3.10 and 3.45 points, respectively, while also improving Reasoning over the pretrained baseline. The same adaptation improves accuracy over the pretrained model on all five evaluated emotion-recognition datasets without emotion-category supervision, suggesting transfer from foundational delivery distinctions to affective judgments. Subsequent instruction fine-tuning yields further gains on MMSU, reaching 65.20% overall accuracy. Additional experiments on Audio Flamingo 3 also support the improvements.

Our contributions are threefold:

- We construct PARAtaxis, a multitask resource built from acoustically verified prosodic contrasts, together with a held-out diagnostic that compares lexical and paralinguistic discrimination under a shared matching formulation.
- We identify a linguistic–paralinguistic asymmetry in decoded selection and representation-level matching, and show that stronger upper-layer matching under supervised adaptation accompanies more accurate paralinguistic discrimination.
- We introduce PARANCE, which explicitly trains bidirectional matching against verified paralinguistic hard negatives alongside next-token prediction, and demonstrate gains in controlled discrimination, downstream speech understanding, and emotion recognition without emotion-category supervision during adaptation.

2 Related Work

2.1 Paralinguistic Understanding in Speech LLMs

Qwen2.5-Omni combines audio perception with language generation and instruction following (Xu et al., 2025a). MMSU evaluates linguistic and paralinguistic perception and reasoning (Wang et al., 2026), while CP-Bench tests contextual reasoning over verbal and non-verbal cues (Wang et al., 2025). Controlled evaluations reveal lexical dominance in emotion judgments (Chen et al., 2026) and, through conflicting transcript claims and acoustic evidence, paralinguistic information recoverable from hidden states but underused in outputs (Pang et al., 2026). The latter work, VoxParadox, combines encoder-layer mixing with preference optimization for adaptation. Complementing these analyses, we fix lexical content and speaker identity, vary acoustically verified prosody, and evaluate shared speech–caption correspondences through decoded choices and prefill representations to guide supervision.

2.2 Contrastive Data for Speech Understanding

LibriTTS-P and ParaSpeechCaps describe speaking style for controllable synthesis and captioning (Kawamura et al., 2024; Diwan et al., 2025). ContraProST pairs prosodic variants of fixed text with alternative translations to evaluate prosody-sensitive translation (Tsiamas et al., 2024), while StressTest links stress patterns to meaning and provides synthetic Stress-17k data for stress detection and reasoning (Yosha et al., 2026). For audio embeddings, Tonmoy (2026) finds that making emotion necessary for caption discrimination recovers emotion recognition, whereas adding more emotion-captioned data does not. PARAtaxis pairs same-transcript, same-speaker speech with contrasts in rate, energy, pitch, or intonation verified from realized waveforms rather than synthesis instructions, without requiring non-target acoustic attributes to remain invariant. Swapped captions form hard negatives for held-out diagnosis and multitask training, providing explicit paralinguistic supervision for PARANCE.

2.3 Contrastive Learning in Decoder-only Models

E5-Mistral trains final-layer end-of-sequence embeddings with InfoNCE and hard negatives (Wang et al., 2024). LLM2Vec enables bidirectional attention before masked next-token prediction and contrastive adaptation (BehnamGhader et al., 2024), while GritLM jointly trains contrastive embeddings with bidirectional attention and generation with causal attention (Muennighoff et al., 2025). For voice LLMs, Vela uses prompted last-token embeddings and text-only contrastive fine-tuning for multimodal retrieval (Hu et al., 2025). ParaNCE jointly optimizes symmetric speech–caption contrastive supervision on final-layer last-token prefill representations and next-token prediction, connecting cue discrimination to generative speech understanding.

3 Pilot Experiment on Paralinguistic Understanding

3.1 Formulating Lexical-Prosodic Discrimination

A central motivation for studying Speech LLMs is their potential to process information in speech beyond lexical content (Wang et al., 2026). Yet, how such information functions within Speech LLMs remains insufficiently understood (Xiang et al., 2025). Recent probing studies show that paralinguistic attributes can be decoded from their hidden representations, suggesting that these cues are encoded by the model (Pang et al., 2026). However, a simple but critical question remained unanswered:"how do they distinguish differences when the same words are delivered differently?"

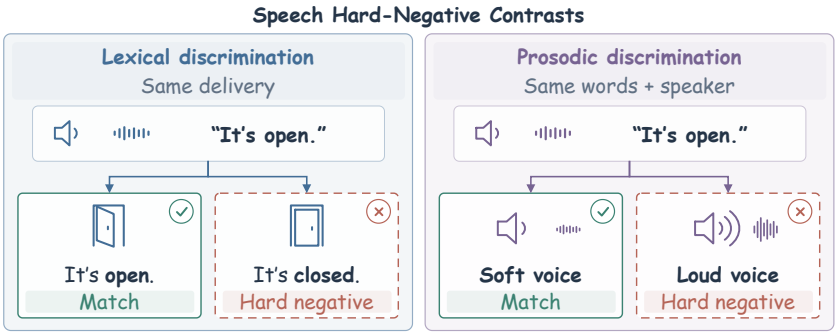


Figure 1: **Speech hard-negative discrimination.** Illustration of the speech-to-caption direction for the two controlled contrasts. Green solid and red dashed boxes denote matched and hard-negative counterparts, respectively.

Hard-Negative Contrasts. To isolate the evidence used for discrimination, we construct two complementary speech–caption hard-negative contrasts with the same matching structure.

- A *lexical contrast* preserves prosodic delivery while changing the transcript, allowing the correspondence to be resolved from lexical content;
- a *prosodic contrast* instead preserves the transcript while changing vocal delivery, removing lexical content as a discriminative shortcut (Chen et al., 2026; Pang et al., 2026).

Each contrast contains two matched speech–caption pairs. This makes the lexical contrast a controlled reference where lexical evidence remains available, while the prosodic contrast tests discrimination between different deliveries of the same words.

Goal. For each contrast, we evaluate whether the model can resolve the corresponding speech–caption pairing. Given an anchor from one modality, either speech or caption, the model selects between its matched counterpart and the hard-negative counterpart from the other modality. We define *lexical discrimination* as correctly resolving this correspondence when the paired examples differ in lexical content while their prosodic condition is controlled. Conversely, *paralinguistic discrimination* requires resolving the correspondence when lexical content and speaker identity are shared but prosodic delivery differs. We evaluate the two forms of discrimination in both speech-to-caption and caption-to-speech directions under the same matching formulation.

3.2 Experiment Setup

Perceptual descriptions of vocal delivery can vary across annotators, while the corresponding acoustic properties are not always defined consistently across studies (Domínguez et al., 2016; Coretta et al., 2023). We therefore instantiate the speech hard-negative contrasts with four measurable target dimensions: speaking rate, energy, pitch, and intonation. Every source utterance is assigned a label on all four dimensions, while the target dimension determines the contrast relation.

Acoustic Verification Criteria. Prosodic labels are derived from sentence-level acoustic measures: articulation rate for speaking rate, integrated loudness for energy, voiced F_0 statistics for pitch level, and utterance-level F_0 contour for intonation. Together, these capture major temporal, intensity, and fundamental-frequency variation in paralinguistic realization (Grant & Walden, 1996; Wagner & Watson, 2010). Table 1 summarizes the resulting verification criteria. For the three scalar dimensions, a valid contrast must fall into different categorical levels and exceed the corresponding separation margin; intonation contrasts are formed between different verified classes. These criteria serve to screen candidate pairs rather than define a general taxonomy of paralinguistic perception. Only the target dimension is constrained to differ; all others are measured but unconstrained. Further details are provided in Appendix A.1.

Table 1: **Acoustic criteria for verified prosodic contrasts.** Scalar-cue pairs occupy different categories and satisfy the listed minimum separation. Intonation pairs differ in verified contour class, with verification conditions in Appendix A.1.4. M/F denote male-/female-range synthetic voices.

Cue / measurement		Low	Medium	High	Pair criterion
Speaking rate r Articulation rate (syll./s)		< 4.0	$[4.0, 5.3]$	> 5.3	$\lvert\Delta r\rvert \geq 0.5$
Energy L Integrated loudness (LUFS)		< -28	$[-28, -18]$	> -18	$\lvert\Delta L\rvert \geq 3$ LU
Pitch level F_0 Median voiced F_0 (Hz)	M F	< 105 < 180	$[105, 130]$ $[180, 230]$	> 130 > 230	$\lvert\Delta F_0\rvert_{\text{st}} \geq 3$
Intonation c Smoothed log-F_0 contour		falling	level	rising	wavy	*Contour-specific*

PARAtaxis: We convert the verified contrasts above into multi-task supervision for prosodic adaptation. The resulting PARAtaxis-Corpus contains 59,603 eligible same-transcript, same-speaker pairs, from which we construct PARAtaxis: 6,480 paired instances (12,960 examples) balanced across seven task forms, target dimensions, and label relations. Each instance comprises two contrastive examples as a verified contrast. We use 5,636 instances for training, while a captioning subset forms the 500-triplet PARAtaxis-Diag, with full separation from training. Construction and split details are provided in Appendices C.2.

Table 2: **PARAtaxis composition.**

Task form	Examples
Captioning	5,184
Cue explanation	1,728
Pairwise comparison	2,592
Three-way comparison	864
Localization	864
Content attribution	864
Claim correction	864
Total	**12,960**

Diagnostic Triplets With the verified contrasts in place, we use PARAtaxis-Diag to compare how reliably Speech LLMs resolve lexical versus prosodic differences under the same correspondence structure. Each diagnostic triplet couples the two contrasts through a shared reference example:

$$\mathcal{T}_i = \left(e_i^{\text{ref}}, e_i^{\text{prosodic}}, e_i^{\text{lexical}}\right), \tag{1}$$

where $(e_i^{\text{ref}}, e_i^{\text{prosodic}})$ and $(e_i^{\text{ref}}, e_i^{\text{lexical}})$ instantiate the prosodic and lexical contrasts, respectively. Each contrast is evaluated as a two-way cross-modal matching problem. Using one modality as the reference, we present the two contrastive examples in the other modality as candidates and ask the model to identify the matching example; reversing the modalities gives the opposite matching direction. We apply this construction separately to the prosodic and lexical contrasts.

3.3 An Observed Linguistic-Paralinguistic Asymmetry

We systematically evaluate lexical and paralinguistic discrimination in speech LLMs using these diagnostic triplets.

As shown in Figure 2 [1], the results reveals a pronounced asymmetry across nine models: linguistic discrimination is consistently strong, indicating that most Speech LLMs readily resolve mismatches in lexical content. In contrast, paralinguistic discrimination remains tightly concentrated around chance, despite the same lexical content being preserved across the competing deliveries. This pattern is consistent across architectures and model scales, suggesting that fine-grained delivery differences remain weakly resolved at the behavioral level. Full evaluation details are provided in Appendix B.

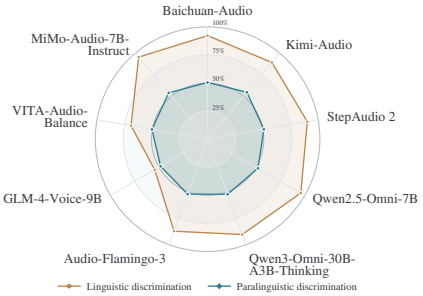


Figure 2: Discrimination performance across models.

Linguistic-Paralinguistic Asymmetry. Speech LLMs readily discriminate linguistic differences, yet show little reliable sensitivity to paralinguistic discrimination.

3.4 A Further Study on the Asymmetry

The decision-level gap raises a corresponding representation-level question: does the model internally distinguish the matched relation from the paralinguistic alternative? To investigate, we probe such alignment by using constructed speech-caption contrasts as diagnostic lens, comparing the similarity of matched speech–descriptor states against their paralinguistic negative alternatives.

Discrimination Advantage. For a speech input x and caption input d, let $\mathbf{H}^{x}_{\ell,n^x}$ and $\mathbf{H}^{d}_{\ell,n^d}$ denote the last hidden states at prefill positions at layer ℓ. We use these states as compact sequence-level representations of the speech and caption inputs:

$$\mathbf{s}^{(\ell)} = \mathbf{H}^{x}_{\ell,n^x}, \qquad \mathbf{t}^{(\ell)} = \mathbf{H}^{d}_{\ell,n^d}. \tag{2}$$

We quantify their relative preference through *matching advantage*, whether a speech or caption state is more similar to its matched counterpart than to the mismatched counterpart from the other example. Following layerwise speech–text analysis (Xiang et al., 2025), we use cosine similarity (Hu et al., 2025) and evaluate the preference symmetrically in both modality directions. Let $\mathbf{s}^{(\ell)}_{i,a}$ and $\mathbf{t}^{(\ell)}_{i,a}$ denote the speech and caption representations of member a in triplet i:

$$\Delta^{(\ell)}_{i,a\leftrightarrow b} = \tfrac{1}{2} \sum_{m\in\{\mathbf{s},\mathbf{t}\}} \left[\cos\left(m^{(\ell)}_{i,a}, \bar{m}^{(\ell)}_{i,a}\right) - \cos\left(m^{(\ell)}_{i,a}, \bar{m}^{(\ell)}_{i,b}\right)\right], \qquad \bar{\mathbf{s}} = \mathbf{t},\ \bar{\mathbf{t}} = \mathbf{s}. \tag{3}$$

A positive Δ indicates that the matched speech–caption pair is more aligned than a mismatched one. Specifically, when $b = e^{\text{paralinguistic}}_i$, we refer to $\Delta^{(\ell)}_{i,a\leftrightarrow b}$ as the *paralinguistic matching advantage*; when $b = e^{\text{linguistic}}_i$, we refer to it as the *linguistic matching advantage*.

To examine whether this decisional pattern is also reflected internally, we focus on Qwen2.5-Omni and Audio Flamingo 3 (AF3) as our two primary study models. Both exhibit a consistent layerwise pattern (Figure 3): linguistic matching becomes increasingly separated toward the upper decoder layers, while paralinguistic matching remains substantially weaker. This representation–decision correspondence suggests that difficulty in resolving delivery-only differences is already reflected in the organization of speech and caption representations, rather than emerging at decoding.

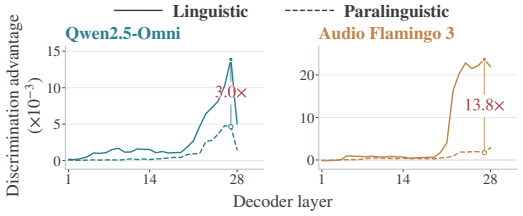


Figure 3: **Layerwise discrimination advantage.** Both models exhibit strong upper-layer separation for linguistic contrasts but substantially weaker separation for delivery-only contrasts.

[1] To minimize directional and candidate-order effects, accuracy is averaged over both pair members, both matching directions, and both candidate orders. Chance accuracy is 50%.

Representation-Level Asymmetry. The poor paralinguistic discrimination is mirrored in internal representations of Speech LLMs: lexical differences induce substantially stronger cross-modal separation than delivery-only differences.

3.5 A Representational Approach to Reduce the Asymmetry

We further examine whether the representational separation identified above is associated with the model's ability to discriminate the same paralinguistic contrasts. Using Qwen2.5-Omni and Audio Flamingo 3 as our study models, we densely fine-tune both on PARAtaxis-Train. As shown in Figure 4, SFT shifts both models in the same positive direction: stronger discrimination advantage between matched and hard-negative speech–caption representations accompanies more reliable paralinguistic discrimination, with this association becoming increasingly pronounced toward the upper decoder layers.

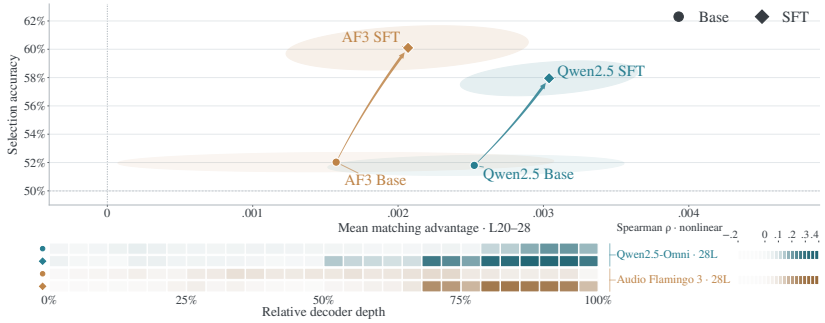


Figure 4: **SFT reveals a representation–behavior association in paralinguistic discrimination. Top:** Mean paralinguistic discrimination advantage over Layers 20–28 versus discrimination accuracy; arrows trace Base and SFT for Qwen2.5-Omni and Audio Flamingo 3 (AF3), with 95% bootstrap uncertainty. **Bottom:** Layerwise Spearman correlation between matching advantage and selection accuracy. Dense SFT jointly strengthens representational matching and behavioral discrimination.

Representational Separation Tracks Discrimination. Stronger upper-layer paralinguistic separation accompanies stronger paralinguistic discrimination.

Together, these results suggest that the discrimination advantage may serve as a promising learning signal for improving models' ability to distinguish paralinguistic differences.

4 Methodology

To utilize the pattern identified in Section 3, we introduce ParaNCE, which explicitly supervises the model to discriminate paralinguistic cues at representation-level. Figure 5 illustrates the framework.

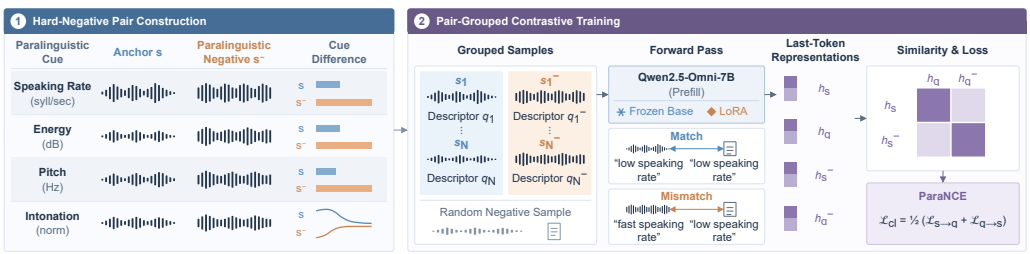


Figure 5: **Overview of ParaNCE.** Contrastive matching over hard-negative pairs is jointly optimized with generative supervision using final-layer last-token representations.

4.1 LEARNING TO DISCRIMINATE CONTRASTIVE CUES

We represent each input using the last prefill token at the final decoder layer L. For each speech–caption example (x_i, d_i), separate forward passes yield

$$\mathbf{s}_i = \mathbf{H}^{x_i}_{L,n^x_i}, \qquad \mathbf{t}_i = \mathbf{H}^{d_i}_{L,n^d_i}. \tag{4}$$

The matching and next-token prediction objectives are jointly optimized at each training step.

During PARANCE, a global batch of $B_g = MB_\mu$ examples comprises M microbatches of size B_μ. Each contains one designated hard-negative pair and $B_\mu - 2$ random examples. Let $\mathcal{A}_m$ index the designated pair in microbatch m. For each anchor $i \in \mathcal{A}_m$, $\bar{i}$ indexes its hard-negative counterpart, and $\mathcal{R}_i$ contains the valid random negatives from the same microbatch. With $\mathcal{C}_i = \{i, \bar{i}\} \cup \mathcal{R}_i$, we average both matching directions over the $2M$ designated anchors:

$$\mathcal{L}_{\text{ParaNCE}} = -\frac{1}{4M} \sum_{m=1}^{M} \sum_{i \in \mathcal{A}_m} \log \frac{e^{2q_{ii}}}{\left(\sum_{j \in \mathcal{C}_i} e^{q_{ij}}\right)\left(\sum_{j \in \mathcal{C}_i} e^{q_{ji}}\right)}. \tag{5}$$

Within each optimization step, the task-formatted forward provides next-token supervision, while speech and caption are prefilling independently to obtain the states used by ParaNCE. These three input organizations belong to the same training iteration and their losses are backpropagated jointly.

5 EXPERIMENTS

We evaluate whether PARANCE improves downstream paralinguistic understanding across complementary benchmarks. Table 3 summarizes the main downstream results together with the component ablations.

Training Details. We conduct controlled adaptation comparisons on Qwen2.5-Omni-7B and compare with additional pretrained speech LLMs. Training conditions and component ablations are detailed in Appendix C.3. All fine-tuning data are derived from PARAtaxis: 5,636 pairs comprising 11,272 examples for contrastive adaptation, and 1,488 caption-derived instruction instances covering paralinguistic description, comparison, and reasoning. Neither subset uses emotion-category supervision, enabling an evaluation of transfer to emotion recognition. Data construction and task inventories are described in Appendix C.2.

We jointly adapt the audio encoder and language decoder using cross-entropy and ParaNCE, and perform instruction fine-tuning with cross-entropy while keeping the audio encoder frozen. Both use AdamW for one epoch, with learning rates of 3×10^{-6} for contrastive adaptation and 1×10^{-6} for instruction fine-tuning. For contrastive adaptation, we use a micro-batch size of 4 and a contrastive candidate pool of $B = 4$, with $\lambda = 1$ and temperature $\tau = 0.07$. ParaNCE is applied to the last prefill-token representation of the final decoder layer without a projection head. We ablate the contrastive pool size and supervision layer in Appendices C.6 and C.7, respectively. LoRA baselines use rank 32, scaling factor 64, and dropout 0.05. Full optimization settings are provided in Appendix C.5.

Evaluation. We evaluate paralinguistic understanding on MMSU (Wang et al., 2026), emotion recognition on IEMOCAP (Busso et al., 2008), and contextual paralinguistic reasoning on CPBench-CPQA (Wang et al., 2025). The main table reports benchmark aggregates, MMSU Perception and Reasoning, and the available MMSU and CPBench-CPQA subtask results. Evaluation examples and prompts are held fixed across the Qwen2.5-Omni-7B conditions. The diagnostic analyses in Section 6 reuse the 500 triplets, candidate-order protocol, and last-token readout from Section 3. Additional evaluation and reporting details are provided in Appendix D.1.

5.1 RESULTS

Table 3 presents external pretrained-model references, together with the shared-backbone Qwen2.5-Omni-7B comparison. Benchmark aggregates and task-level results retain their respective evaluation metrics.

Downstream paralinguistic understanding. PARANCE consistently improves downstream paralinguistic understanding, increasing improves MMSU Perception by 9.49 points over Base, while maintaining comparable Reasoning performance (+0.33 points), and further improves IEMOCAP

Table 3: **Prosodic and paralinguistic understanding on MMSU.** All results report accuracy (%). Higher is better. We report 12 Perception and 3 Reasoning subtasks; the full-benchmark Avg., Perception (Perc.), and Reasoning (Reas.) aggregates are retained as provided. Inst FT denotes Stage 2 instruction fine-tuning on 1,488 instances (seed 43). Bold scores mark the best value among the displayed Qwen2.5-Omni-7B conditions.

Model / condition	Aggregate			Perception												Reasoning		
	Avg.	Perc.	Reas.	Emo.	Inton.	Pitch	Speed	Vol.	Stress	Pause	Prol.	Range	Age	Gender	Spk. #	Emo.	Inton.	Pause
Other pretrained speech LLMs																		
Baichuan-Audio (Li et al., 2025)	45.26	38.10	52.89	23.08	33.33	20.37	18.35	17.27	23.64	16.82	59.04	30.56	26.92	37.50	54.17	12.73	53.85	27.78
Kimi-Audio-7B-Instruct (KimiTeam et al., 2025)	57.98	50.00	66.49	44.44	28.83	33.33	19.27	29.09	79.09	38.32	92.77	15.74	51.92	93.33	45.83	20.00	73.08	40.00
Step-Audio-2-mini-Think (Wu et al., 2025)	56.86	44.22	70.33	83.76	28.83	24.07	20.18	25.45	30.91	29.91	72.29	26.85	55.77	96.67	21.67	54.55	66.67	43.33
Qwen3-Omni-30B (Xu et al., 2025b)	62.74	51.90	74.30	32.48	39.64	25.00	21.10	29.09	50.00	34.58	80.72	35.19	66.35	92.50	67.50	26.36	74.36	41.11
Audio-Flamingo-3 (Goel et al., 2025)	59.46	43.95	75.99	81.20	29.73	26.85	29.36	30.00	39.09	33.64	45.78	22.22	29.81	72.50	49.17	42.73	73.08	57.78
GLM-4-Voice-9B (Zeng et al., 2024)	36.14	25.78	47.19	9.40	19.82	28.70	21.10	26.36	24.55	25.23	38.55	20.37	16.35	24.17	33.33	22.73	52.56	41.11
VITA-Audio-Balance (Long et al., 2025)	39.68	30.19	49.79	26.50	21.62	21.30	25.69	22.73	36.36	33.64	38.55	23.15	24.04	43.33	32.50	22.73	52.56	40.00
MiMo-Audio-7B-Instruct (Xiaomi, 2025)	63.30	50.54	76.90	41.88	46.85	25.00	24.77	25.45	41.82	43.93	72.29	32.41	64.42	93.33	49.17	30.00	78.21	60.00
Qwen2.5-Omni-7B (Xu et al., 2025a)																		
Base	59.16	46.59	72.56	43.59	24.32	27.78	18.35	23.64	45.45	40.19	62.65	25.00	55.77	79.17	50.00	22.73	61.54	42.22
LoRA SFT	60.08	45.93	75.17	37.61	19.82	31.48	21.10	31.82	42.73	30.84	62.65	31.48	54.81	70.83	40.00	25.45	60.26	56.67
LoRA SFT + Inst FT	60.16	46.16	75.08	39.32	18.92	33.33	23.85	32.73	41.82	32.71	63.86	30.56	54.81	70.83	40.83	24.55	60.26	56.67
Dense SFT	63.00	49.77	77.11	53.85	29.73	34.26	27.52	32.73	49.09	45.79	**75.90**	30.56	55.77	**85.00**	38.33	29.09	**74.36**	61.11
Dense SFT + Inst FT	63.94	50.93	**77.81**	53.85	29.73	33.33	34.86	31.82	51.82	45.79	73.49	27.78	57.69	84.17	42.50	**30.00**	73.08	57.78
InfoNCE	62.70	49.42	76.86	53.85	27.03	31.48	20.18	34.55	53.64	37.38	62.65	23.15	57.69	79.17	44.17	27.27	70.51	58.89
InfoNCE + Inst FT	62.88	49.26	77.40	**55.56**	19.82	29.63	22.02	30.00	55.45	36.45	60.24	24.07	**60.58**	80.00	45.83	**30.00**	**74.36**	61.11
ParaNCE	64.72	52.87	77.36	54.70	38.74	34.26	**36.70**	**38.18**	62.73	45.79	65.06	36.11	58.65	81.67	48.33	**30.00**	69.23	**62.22**
ParaNCE + Inst FT	**65.20**	**53.91**	77.23	54.70	**39.64**	**37.04**	**36.70**	37.27	**64.55**	**46.73**	62.65	**37.04**	58.65	83.33	**53.33**	**30.00**	70.51	61.11

blue indicates improvement and brown indicates decline relative to Base.

accuracy by 5.10 points. As shown in Table 3, these gains span both foundational paralinguistic cues and higher-level emotion understanding, suggesting that the proposed training strengthens paralinguistic sensitivity without compromising reasoning ability. Detailed component-wise comparisons are provided in Appendix C.3.

6 Discussion

We organize our discussion around three research questions: whether the representation–behavior pattern revealed by the controlled diagnostic can be deliberately strengthened through supervision, whether the resulting discrimination transfers beyond speech–caption matching to broader paralinguistic understanding, and whether supervision on foundational vocal cues further transfers to higher-level affective understanding.

Finding 1: Stronger paralinguistic matching accompanies more accurate discrimination. Revisiting the pattern identified in Section 3, Figure 6 shows that, for both Qwen2.5-Omni and Audio Flamingo 3, ParaNCE further advances the model toward stronger discrimination on both representation and decision level. This consistent shift across both models follows the representation–decision pattern observed in the Pilot, showing that ParaNCE can directly exploit this pattern to improve paralinguistic discrimination.

> **Takeaway.** ParaNCE improves paralinguistic discrimination on both decision and representation level, further proving the identified pattern.

Finding 2: Learned discrimination ability extend to broader paralinguistic Understanding. The MMSU results in Table 3 assess whether the learned capabilities extend beyond speech–caption matching. ParaNCE improves both Perception and Reasoning, with its clearest advantage over the baselines in Perception. This pattern suggests that supervision on foundational cues can support broader speech understanding, while its extent varies across downstream tasks.

> **Takeaway.** Representation-level paralinguistic matching gains extend most clearly to broader paralinguistic Perception while preserving Reasoning performance.

Finding 3: Learning paralinguistic cues generalizes to emotion understanding without emotion supervision. Table 4 shows consistent transfer across five emotion-recognition datasets: IEMOCAP, CREMA-D, RAVDESS, EMO-DB, and ESD-EN. Training on our balanced ParaTAXIS

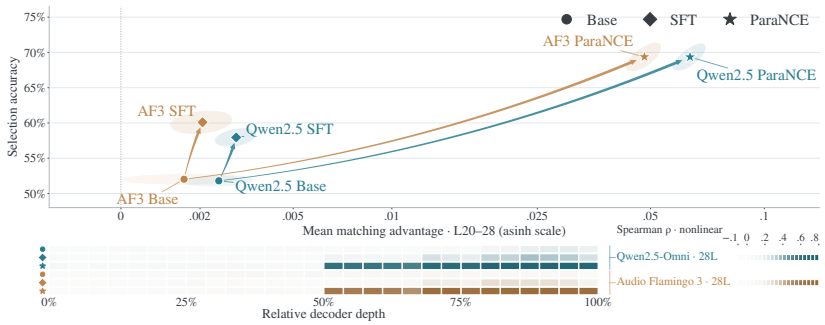


Figure 6: **ParaNCE further strengthens the representation–behavior association. Top:** Mean paralinguistic matching advantage over Layers 20–28 versus selection accuracy for Base, Dense SFT, and ParaNCE on Qwen2.5-Omni and Audio Flamingo 3 (AF3). **Bottom:** Layerwise Spearman correlation between matching advantage and selection accuracy. ParaNCE extends the gains of Dense SFT in both representational matching and behavioral discrimination.

data, constructed with consistent acoustic verification criteria, improves performance over Base on all five datasets for every displayed adaptation condition, despite using no emotion annotations during adaptation. Instruction fine-tuning further improves all five scores for each training objective. These results show that learning to discriminate paralinguistic cues transfers beyond the supervised tasks to broader affective understanding.

Table 4: **Transfer to higher-level paralinguistic understanding.** All results report accuracy (%) on Qwen2.5-Omni-7B. All adaptation conditions use ParaTAXIS without emotion annotations; Inst FT uses the 1,488-instance Stage 2 instruction set. Darker blue indicates a larger accuracy improvement over Base. Bold scores mark the best value among the displayed conditions.

Model / condition	IEMOCAP	CREMA-D	RAVDESS	EMO-DB	ESD-EN
Base	63.52	27.72	21.25	47.66	29.93
Dense SFT	67.08	31.98	33.26	54.02	37.38
Dense SFT + Inst FT	68.33	43.00	40.49	61.50	41.30
InfoNCE	69.56	45.54	44.51	69.53	51.81
InfoNCE + Inst FT	**69.59**	**58.88**	**52.99**	**75.70**	**55.53**
ParaNCE	66.76	40.09	34.17	50.65	39.87
ParaNCE + Inst FT	68.53	52.51	43.06	51.40	48.91

Takeaway. Learning paralinguistic cues can contribute to broader paralinguistic understanding, such as emotion recognition, even without direct emotion supervision during adaptation.

7 Conclusion

Using PARAtaxis, we study whether learning four acoustically verified foundational vocal cues supports paralinguistic understanding beyond the tasks directly supervised during adaptation. Controlled diagnostics reveal that pretrained Speech LLMs distinguish lexical differences much more reliably than different deliveries of the same words, while stronger representation-level matching after adaptation accompanies improved behavioral discrimination. Motivated by this association, ParaNCE combines next-token prediction with bidirectional matching against verified paralinguistic hard negatives, improving controlled discrimination and downstream MMSU perception and reasoning. Crucially, adaptation on these foundational cues also improves emotion recognition across five datasets without emotion-category supervision. This transfer occurs across supervised and contrastive objectives, supporting cue learning as a general approach, while ParaNCE is particularly effective for fine-grained perception. How these cues are combined and used to support such judgments remains an open question.

AI USE STATEMENT

AI tools assisted manuscript drafting and editing, literature lookup, code inspection, visualization code, and figure preparation. For data construction, Qwen-TTS (Qwen Team, 2025) generated candidate synthetic speech, Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025b) generated and refined speech captions, and GPT-4o (OpenAI, 2024) checked caption consistency against the supplied transcripts and acoustic labels. This text-based consistency check did not independently establish acoustic validity. Whisper large (Radford et al., 2023) provided automatic transcriptions for WER-based transcript-fidelity screening. Data-quality control combined these automated procedures with acoustic verification and manual quality screening, as described in Appendices A.2, A.3, and A.8. The authors take responsibility for the final manuscript, scientific claims, data-processing decisions, citations, code, and reported results, including all content and artifacts produced with AI assistance.

ETHICS STATEMENT

Our method optimizes only the final prefill state, deliberately avoiding autoregressive states that are already influenced by generated tokens. While this isolates representation-level paralinguistic learning, it leaves open how such information is preserved and used throughout autoregressive decoding. Moreover, although our results suggest that learning foundational paralinguistic cues can transfer to higher-level abilities such as emotion recognition, we do not systematically investigate how these cues interact, compose, and give rise to more abstract paralinguistic understanding. Understanding both the autoregressive use of paralinguistic information and its transition from foundational cues to higher-level capabilities remains important future work.

REFERENCES

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=IW1PR7vEBf.

Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 17:97–110, 1993. URL https://www.fon.hum.uva.nl/paul/papers/Proceedings_1993.pdf.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6. URL https://doi.org/10.1007/s10579-008-9076-6.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5848–5877, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.274. URL https://aclanthology.org/2026.eacl-long.274/.

Stefano Coretta, Joseph V. Casillas, Simon Roessig, Michael Franke, et al. Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science*, 6(3):1–29, 2023. doi: 10.1177/25152459231162567. URL https://doi.org/10.1177/25152459231162567.

Thomas H. Crystal and Arthur S. House. Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America*, 88(1):101–112, 1990. doi: 10.1121/1.399955. URL https://doi.org/10.1121/1.399955.

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. Scaling rich style-prompted text-to-speech datasets. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3639–3659, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.180. URL https://aclanthology.org/2025.emnlp-main.180/.

Mónica Domínguez, Mireia Farrús, and Leo Wanner. An automatic prosody tagger for spontaneous speech. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 377–386, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1037/.

European Broadcasting Union. Loudness normalisation and permitted maximum level of audio signals. Recommendation ebu r 128, European Broadcasting Union, 2020. URL https://tech.ebu.ch/docs/r/r128v4_0.pdf.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025. doi: 10.48550/arXiv.2507.08128. URL https://arxiv.org/abs/2507.08128.

Frieda Goldman-Eisler. The significance of changes in the rate of articulation. *Language and Speech*, 4(3):171–174, 1961. doi: 10.1177/002383096100400305. URL https://doi.org/10.1177/002383096100400305.

Ken W. Grant and Brian E. Walden. Spectral distribution of prosodic information. *Journal of Speech, Language, and Hearing Research*, 39(2):228–238, 1996. doi: 10.1044/jshr.3902.228. URL https://doi.org/10.1044/jshr.3902.228.

Daniel Hirst. A praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. In *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 1233–1236, 2007. URL https://www.icphs2007.de/conference/Papers/1443/index.html.

Ruofan Hu, Yan Xia, Minjie Hong, Jieming Zhu, Bo Chen, Xiaoda Yang, Minghui Fang, and Tao Jin. Vela: Scalable embeddings with voice large language models for multimodal retrieval. In *Interspeech 2025*, pp. 2640–2644, 2025. doi: 10.21437/Interspeech.2025-159. URL https://www.isca-archive.org/interspeech_2025/hu25b_interspeech.html.

International Telecommunication Union. Algorithms to measure audio programme loudness and true-peak audio level. Recommendation itu-r bs.1770-5, International Telecommunication Union, 2023. URL https://www.itu.int/rec/R-REC-BS.1770-5-202311-I/en.

Masaya Kawamura, Ryuichi Yamamoto, Yuma Shirahata, Takuya Hasumi, and Kentaro Tachibana. LibriTTS-P: A corpus with speaking style and speaker identity prompts for text-to-speech and style captioning. In *Proceedings of Interspeech 2024*, pp. 1850–1854, 2024. doi: 10.21437/Interspeech.2024-692. URL https://www.isca-archive.org/interspeech_2024/kawamura24_interspeech.html.

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, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*, 2025. doi: 10.48550/arXiv.2504.18425. URL https://arxiv.org/abs/2504.18425.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. Baichuan-Audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*, 2025. doi: 10.48550/arXiv.2502.17239. URL https://arxiv.org/abs/2502.17239.

Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, Haoyu Cao, Ke Li, Rongrong Ji, and Xing Sun. VITA-Audio: Fast interleaved cross-modal token generation for efficient large speech-language model. *arXiv preprint arXiv:2505.03739*, 2025. doi: 10.48550/arXiv.2505.03739. URL https://arxiv.org/abs/2505.03739.

Matthias Mauch and Simon Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 659–663, 2014. doi: 10.1109/ICASSP.2014.6853678. URL https://doi.org/10.1109/ICASSP.2014.6853678.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *International Conference on Learning Representations*, pp. 45544–45613, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/70cf215430492f7d34830a24e744b3f1-Abstract-Conference.html.

OpenAI. GPT-4o system card. Technical report, OpenAI, August 2024. URL https://openai.com/index/gpt-4o-system-card/.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio LLMs listen or read? analyzing and mitigating paralinguistic failures with VoxParadox. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/60946.

Hugo Quené. On the just noticeable difference for tempo in speech. *Journal of Phonetics*, 35(3): 353–362, 2007. doi: 10.1016/j.wocn.2006.09.001. URL https://doi.org/10.1016/j.wocn.2006.09.001.

Qwen Team. Time to speak some dialects, Qwen-TTS! Official Qwen model announcement, June 2025. URL https://qwenlm.github.io/blog/qwen-tts/.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Johan 't Hart. Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America*, 69(3):811–821, 1981. doi: 10.1121/1.385592. URL https://doi.org/10.1121/1.385592.

Paul Taylor. The rise/fall/connection model of intonation. *Speech Communication*, 15(1–2):169–186, 1994. doi: 10.1016/0167-6393(94)90050-7. URL https://doi.org/10.1016/0167-6393(94)90050-7.

Abdul Basit Tonmoy. Discriminative axis, not data volume: What a contrastive corpus teaches an audio embedding. *arXiv preprint arXiv:2608.01560*, 2026. doi: 10.48550/arXiv.2608.01560. URL https://arxiv.org/abs/2608.01560.

Hartmut Traunmüller and Anders Eriksson. The frequency range of the voice fundamental in the speech of male and female adults. Unpublished manuscript, Department of Linguistics, Stockholm University, 1995. URL https://resources.ling.su.se/hartmut/f0_m%26f.pdf.

Ioannis Tsiamas, Matthias Sperber, Andrew Finch, and Sarthak Garg. Speech is more than words: Do speech-to-text translation systems leverage prosody? In *Proceedings of the Ninth Conference on Machine Translation*, pp. 1235–1257, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.119. URL https://aclanthology.org/2024.wmt-1.119/.

Michael Wagner and Duane G. Watson. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7–9):905–945, 2010. doi: 10.1080/01690961003589492. URL https://doi.org/10.1080/01690961003589492.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, pp. 31374–31410, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/34e278fbbd7d6d7d788c98065988e1a9-Abstract-Conference.html.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.642. URL https://aclanthology.org/2024.acl-long.642/.

Qiongqiong Wang, Hardik Bhupendra Sailor, Tianchi Liu, Wenyu Zhang, Muhammad Huzaifah, Nattadaporn Lertcheva, Shuo Sun, Nancy F. Chen, Jinyang Wu, and AiTi Aw. Benchmarking contextual and paralinguistic reasoning in speech-LLMs: A case study with in-the-wild data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14133–14148, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.760. URL https://aclanthology.org/2025.findings-emnlp.760/.

William M. Whitmer and Michael A. Akeroyd. Level discrimination of speech sounds by hearing-impaired individuals with and without hearing amplification. *Ear & Hearing*, 32(3):391–398, 2011. doi: 10.1097/AUD.0b013e318202b620. URL https://doi.org/10.1097/AUD.0b013e318202b620.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Bingxin Li, Buyun Ma, Changhe Song, Dongqing Pang, Guoqiang Hu, Haiyang Sun, Kang An, Na Wang, Shuli Gao, Wei Ji, Wen Li, Wen Sun, Xuan Wen, Yong Ren, Yuankai Ma, Yufan Lu, Bin Wang, Bo Li, Changxin Miao, Che Liu, Chen Xu, Dapeng Shi, Dingyuan Hu, Donghang Wu, Enle Liu, Guanzhe Huang, Gulin Yan, Han Zhang, Hao Nie, Haonan Jia, Hongyu Zhou, Jianjian Sun, Jiaoren Wu, Jie Wu, Jie Yang, Jin Yang, Junzhe Lin, Kaixiang Li, Lei Yang, Liying Shi, Li Zhou, Longlong Gu, Ming Li, Mingliang Li, Mingxiao Li, Nan Wu, Qi Han, Qinyuan Tan, Shaoliang Pang, Shengjie Fan, Siqi Liu, Tiancheng Cao, Wanying Lu, Wenqing He, Wuxun Xie, Xu Zhao, Xueqi Li, Yanbo Yu, Yang Yang, Yi Liu, Yifan Lu, Yilei Wang, Yuanhao Ding, Yuanwei Liang, Yuanwei Lu, Yuchu Luo, Yuhe Yin, Yumeng Zhan, Yuxiang Zhang, Zidong Yang, Zixin Zhang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. Step-Audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025. doi: 10.48550/arXiv.2507.16632. URL https://arxiv.org/abs/2507.16632.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5187–5202, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.262. URL https://aclanthology.org/2025.emnlp-main.262/.

LLM-Core-Team Xiaomi. MiMo-Audio: Audio language models are few-shot learners, 2025. URL https://github.com/XiaomiMiMo/MiMo-Audio.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a. doi: 10.48550/arXiv.2503.20215. URL https://arxiv.org/abs/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, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b. doi: 10.48550/arXiv.2509.17765. URL https://arxiv.org/abs/2509.17765.

Iddo Yosha, Gallil Maimon, and Yossi Adi. StressTest: Can YOUR speech LM handle the stress? In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1264–1285, San Diego,

California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.64. URL https://aclanthology.org/2026.findings-acl.64/.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024. doi: 10.48550/arXiv.2412.02612. URL https://arxiv.org/abs/2412.02612.

Appendix

Contents

Appendix

Contents

A PARAtaxis Construction Details 17
- A.1 Acoustic Measurement Definitions 17
- A.2 Source Assets and Candidate Pairing 18
- A.3 Caption Construction and Text Consistency 19
- A.4 Hard-Negative Pair Construction 19
- A.5 Balanced Stage1 Task Pool 20
- A.6 Caption Construction Prompts 20
- A.7 Frozen Seed-43 Training Split 22
- A.8 Experimental Snapshot and Verification Boundary 23

B Diagnostic Set and Pilot Evaluation Details 23
- B.1 Qualification, Isolation, and Triplet Construction 23
- B.2 Response-Level Speech–Caption Matching 24
- B.3 Representation Extraction and Similarity Matrices 25
- B.4 Representation-Readout Ablation 25
- B.5 Architectural Constraints and Scope of the Pilot Adaptation Analysis 26
- B.6 Pilot Reproducibility Checklist 26

C Training and Ablation Details 27
- C.1 Batch Construction and Contrastive Candidates 27
- C.2 Training-Data Composition 27
- C.3 Training Conditions and Component Ablations 28
- C.4 Training Ablations on Audio Flamingo 3 29
- C.5 Optimization and Implementation 29
- C.6 Effect of Contrastive Pool Size 30
- C.7 Supervising the Final Contextualized Representation 32

D Benchmark Evaluation and Additional Results 34
- D.1 Benchmark Evaluation Details 34
- D.2 Diagnostic caption-Selection Evaluation after Training 34
- D.3 Reproducibility and Reporting Boundaries 34

A PARATAXIS CONSTRUCTION DETAILS

This appendix details the acoustic criteria, candidate pairing, task construction, and frozen seed-43 partition underlying the controlled contrasts in Section 3. PARAtaxis comprises the source-utterance registry, the balanced Stage1 task pool, and its pair-preserving training split. Appendix B describes the held-out PARAtaxis-Diag set.

A.1 ACOUSTIC MEASUREMENT DEFINITIONS

We use fixed sentence-level acoustic criteria for four target dimensions: speaking rate (stored as `rhythm`), energy, pitch level, and intonation (stored as `tone`). Their role is operational rather than taxonomic: they provide a common rule for assigning labels and verifying candidate contrasts. For speaking rate, energy, and pitch, a valid contrast must occupy different categorical levels and exceed the cue-specific continuous margin in Table 1. Intonation is represented by one of four quality-controlled contour classes. The thresholds are fixed independently of the downstream results.

A.1.1 SPEAKING-RATE MEASUREMENT

Speaking rate is measured as articulation rate,

$$r_{\mathrm{art}} = \frac{n_{\mathrm{syl}}}{t_{\mathrm{speak}}}, \tag{6}$$

where n_{syl} is the number of produced syllables and t_{speak} is accumulated speaking duration after excluding silent pauses. This follows the conventional distinction between articulation rate and pause-inclusive speech rate (Goldman-Eisler, 1961; Crystal & House, 1990). The low, medium, and high labels correspond to slow, moderate, and fast delivery, respectively:

$$\text{slow} : r_{\mathrm{art}} < 4.0, \qquad \text{medium} : 4.0 \le r_{\mathrm{art}} \le 5.3, \qquad \text{fast} : r_{\mathrm{art}} > 5.3. \tag{7}$$

A valid pair must also satisfy

$$\left| r_{\mathrm{art}}^{(i)} - r_{\mathrm{art}}^{(j)} \right| \ge 0.5 \text{ syll./s}. \tag{8}$$

The margin is an operational separation criterion rather than a universal perceptual threshold (Quené, 2007).

A.1.2 ENERGY MEASUREMENT

Energy is measured as utterance-level integrated loudness L in LUFS, following ITU-R BS.1770 (International Telecommunication Union, 2023). The low, medium, and high categories (described as soft, moderate, and loud delivery) are

$$\text{low} : L < -28, \qquad \text{medium} : -28 \le L \le -18, \qquad \text{high} : L > -18. \tag{9}$$

with the additional pairwise requirement

$$\left| L^{(i)} - L^{(j)} \right| \ge 3 \text{ LU}. \tag{10}$$

The interval is centered on the EBU R128 programme-loudness reference, but the ± 5-LU category band is our operational choice rather than an EBU-prescribed boundary (European Broadcasting Union, 2020). The separation margin is informed by the scale of reported speech-level discrimination differences; neither threshold is a universal loudness just-noticeable difference (Whitmer & Akeroyd, 2011). These absolute categories are compared only within our shared synthesis and amplitude-calibration pipeline.

A.1.3 PITCH-LEVEL MEASUREMENT

Pitch level is defined by the median fundamental frequency over voiced frames. The F_0 trajectory is estimated with Praat's autocorrelation method, using a broad initial pass followed by an adaptive search range, and cross-checked with pYIN (Boersma, 1993; Mauch & Dixon, 2014). We exclude samples with insufficient voiced coverage, substantial estimator disagreement, or likely octave errors. Separate operational ranges are used for male-range and female-range synthetic voices (Traunmüller & Eriksson, 1995):

$$\text{male-range:} \quad < 105 \text{ Hz}, \quad 105\text{–}130 \text{ Hz}, \quad > 130 \text{ Hz}, \tag{11}$$

$$\text{female-range:} \quad < 180 \text{ Hz}, \quad 180\text{–}230 \text{ Hz}, \quad > 230 \text{ Hz}, \tag{12}$$

for low, medium, and high, respectively. Pairwise separation is measured in semitones,

$$d_{F_0}(x_i, x_j) = 12 \left| \log_2 \frac{\widetilde{F}_0(x_i)}{\widetilde{F}_0(x_j)} \right|, \tag{13}$$

where $\widetilde{F}_0$ denotes median voiced-frame F_0. A valid pitch contrast requires $d_{F_0}(x_i, x_j) \geq 3$ st. As above, this is a conservative construction margin rather than a minimum perceptible pitch difference ('t Hart, 1981).

A.1.4 Intonation Measurement

Intonation describes the direction and shape of the utterance-level F_0 contour rather than absolute pitch height. The voiced trajectory is converted to a logarithmic semitone scale and smoothed to suppress local microprosodic perturbations. The resulting macroprosodic trajectory is assigned to one of

$$\{\text{falling},\ \text{level},\ \text{rising},\ \text{wavy}\}, \tag{14}$$

using robust beginning-to-end differences, local slope directions, and salient contour reversals (Taylor, 1994; Hirst, 2007). Rising and falling contours require at least three semitones of robust beginning-to-end variation; otherwise the contour is treated as level. Wavy contours use their reversal-based quality-control rule rather than a scalar pair margin. Pitch level and intonation are measured independently: a high absolute F_0 does not imply a rising contour, and a rising contour does not imply a high pitch level.

A.1.5 Operational Scope of the Criteria

The numerical thresholds were fixed before inspecting the PARAtaxis distribution and were not tuned to downstream performance. The cited work motivates the acoustic quantities and plausible orders of magnitude; the exact boundaries remain operational choices rather than a universal perceptual taxonomy. These criteria determine whether a source utterance receives a usable categorical label and whether two same-transcript, same-speaker utterances can instantiate a target-cue contrast. Only the designated dimension is required to differ. Measurements on the remaining three dimensions are retained as attributes and may also differ. Consequently, the data support controlled *target-cue* contrasts without claiming that every non-target acoustic property is invariant.

A.2 Source Assets and Candidate Pairing

Candidate Generation. We use Qwen-TTS to generate candidate speech realizations from a source transcript and a fixed synthetic speaker identity. The synthesis request specifies one target cue at a time, such as speaking rate, energy, pitch level, or intonation. The same transcript and speaker are reused across realizations intended to form a hard-negative pair. Synthesis instructions provide candidate variations, while the transcript and acoustic checks below determine whether the realized utterances satisfy the construction criteria. The candidate pool before WER screening contains more than 60,000 synthesized utterances.

Transcript Fidelity. All synthesized utterances are transcribed with Whisper large and screened against their intended transcripts using word error rate (WER):

$$\text{WER} = \frac{S + D + I}{N}, \tag{15}$$

where S, D, and I denote word substitutions, deletions, and insertions, respectively, and N is the number of words in the reference transcript. We retain utterances with WER ≤ 0.10, consistent with the threshold used in prior synthetic-speech filtering (Xiang et al., 2025). This check assesses transcript fidelity; the separate acoustic criteria below verify realized vocal delivery.

Acoustic Verification. Acceptance is based on the measured signal rather than the requested control text. Each candidate is measured along all four acoustic dimensions and is retained only when the target cue satisfies its required category and quality-control checks. Candidates that fail the target-cue criterion are regenerated with the same transcript and speaker. Non-target cues are measured and stored but are not used as universal pair-selection constraints. A verified pair therefore guarantees the target-cue difference without claiming invariance of every other acoustic property.

Manual Quality Screening. Recruited university-student annotators additionally perform manual data-quality screening and are compensated at RMB 30 per hour. This screening supplements the automated transcript and acoustic checks and is distinct from a systematic human perceptual validation study. It does not imply that every training view has completed factual review; the review boundaries of the frozen experimental snapshot are reported in Appendix A.8.

Source Registry. The frozen source registry contains 46,883 historical and 3,161 supplemental utterances. Natural-language captions are available for 16,684 utterances: 13,523 historical records and 3,161 supplemental drafts. Each source record retains the waveform, approved transcript, speaker identity, measurements and labels for all four dimensions, caption provenance, and quality-control fields. These registry counts describe a different stage from the approximate pre-screening candidate count above.

Candidate Pairing. Candidate pairs are formed only within identical transcript–speaker groups. For the designated target dimension, the two members must have different categorical labels. Speaking rate, energy, and pitch must additionally exceed their continuous separation margins, whereas intonation uses the quality-controlled contour class. This procedure yields 59,603 distinct eligible pairs before task balancing. Non-target dimensions are recorded for each member but are not required to match. Figure 7 summarizes the construction pipeline; WER screening verifies transcript fidelity, while acoustic verification establishes the target-cue contrast used for pairing.

Table 5: **Source registry and pairing snapshot.** Registry counts precede caption qualification and task balancing.

Statistic	Value
Historical source utterances	46,883
Supplemental source utterances	3,161
Source utterances with captions or caption drafts	16,684
Eligible same-transcript, same-speaker pairs	59,603

A.3 Caption Construction and Text Consistency

Each caption-bearing speech realization is associated with a natural-language caption that expresses both the spoken content and the measured vocal delivery. Caption generation receives the audio, the authoritative transcript, and the four verified cue categories. The caption is required to cover speaking rate, energy, pitch level, and intonation while avoiding claims unsupported by these annotations, including inferred emotion, intent, identity, or recording context.

A locally deployed Qwen3-Omni-30B-A3B-Thinking model is used to produce/refine the caption. A text-only GPT-4o pass then checks the draft against the supplied transcript and acoustic categories. The text-only verifier does not establish acoustic validity independently; it checks only whether the caption is consistent with the acoustic references and transcript. Missing or contradictory cue descriptions are revised, and unsupported claims are removed. The resulting text is paired with speech x as caption d when the item enters the caption-bearing task pool.

For reproducibility, Appendix A.6 retains the prompt interfaces used for caption generation and text consistency checking. The internal prompt schema uses the field name `caption`; throughout the paper, this text is referred to as the *caption*.

A.4 Hard-Negative Pair Construction

For each retained hard-negative pair $\mathcal{P} = \{(x_i, d_i), (x_j, d_j)\}$, the matched relations are (x_i, d_i) and (x_j, d_j). Swapping the captions produces the two *hard-negative cross-matches* (x_i, d_j) and (x_j, d_i) used by ParaNCE.

The target cue identifies the acoustic dimension used for pair verification. A caption may still mention all four measured cues, and non-target cue movement is retained in the metadata rather than optimized away. Pair verification therefore does not require the two waveforms to differ along exactly one acoustic variable.

A.5 BALANCED STAGE1 TASK POOL

A *scope* is a two-member contrast associated with a task template, question, target dimension, and label relation. Each member yields one task view consisting of a user message, audio input, and assistant target; both views remain together during sampling and partitioning. We sample the eligible pairs on a fixed task-by-dimension-by-relation-by-member/order grid, so abundant cells do not displace sparse ones. The resulting Stage1 pool contains 6,480 scopes and 12,960 views. Each of the four dimensions contributes 1,620 scopes. For the three scalar dimensions, each of the three unordered label relations contributes 540 scopes; the six unordered intonation relations contribute 270 scopes each.

The seven task forms vary how the verified acoustic relation is queried while preserving the same underlying facts. Captioning and cue explanation use a single source utterance. Pairwise comparison, localization, and content attribution use two-source constructions, while three-way comparison uses three ordered segments. Claim correction presents one source utterance together with an incorrect target-cue assertion. Whenever multiple source clips are presented as one model input, the waveform is assembled deterministically with a 400-ms silent interval and the source identities and assembly recipe are retained. The full Stage1 pool contains 5,184 such composite-audio recipes.

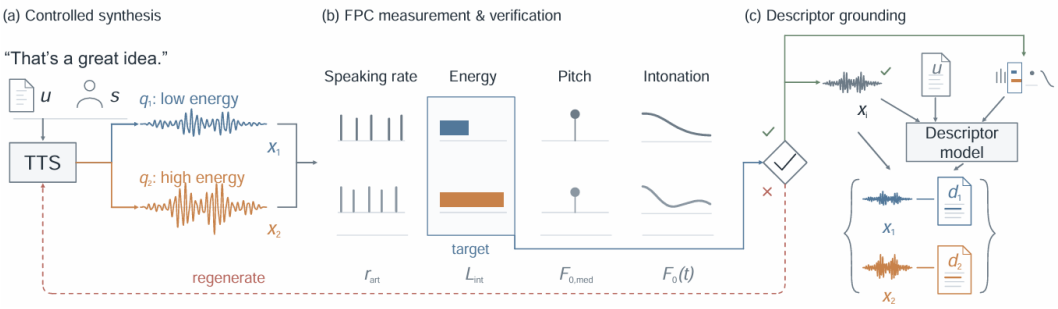


Figure 7: **PARAtaxis construction and partitioning.** Source utterances are acoustically measured and labeled, same-transcript and same-speaker candidates are screened by the target-cue criteria, and eligible pairs are instantiated into a balanced seven-task Stage1 pool. Complete pair scopes are then assigned either to the frozen training split or to the held-out S1-CAP pool from which PARAtaxis-Diag is constructed.

A.6 CAPTION CONSTRUCTION PROMPTS

The following prompt interfaces are retained for reproducibility. Double-braced placeholders are populated with sample-specific values. The four acoustic categories and the transcript are treated as authoritative inputs; the models are instructed not to infer emotion or other unsupported attributes.

Table 6: Cue-field correspondence used by the caption-generation and text-consistency prompts.

Generation field	Verifier field	Canonical labels
`speaking_rate`	`rhythm`	low, medium, high
`energy`	`energy`	low, medium, high
`pitch`	`pitch`	low, medium, high
`intonation`	`tone`	falling, level, rising, wavy

Local Qwen3-Omni caption-generation text template

```
Create one research-grade, fine-grained caption for the supplied speech audio.

The following four measurements are authoritative facts. Never contradict,
weaken, or replace them:
{{AUTHORITATIVE_GT_LINES}}

Authoritative transcript (preserve its content exactly; normal punctuation
changes are allowed):
{{REFERENCE_TRANSCRIPT}}
```

Local Qwen3-Omni caption-generation text template *(continued)*

```
Write a natural caption of two to four sentences. Integrate the spoken content
with how it is delivered. Do not write a mechanical four-item profile and do
not merely repeat label words. You may explain audible evidence, but do not
invent within-utterance changes that are absent from the provided
measurements. Do not infer emotion, intent, age, accent, identity,
personality, social role, or recording context. A wavy intonation means a
globally varying up-and-down contour; it is not an emotion and is not a
Rise-Fall/Fall-Rise benchmark label.

Return exactly one JSON object with this schema:
{
  "caption": "natural caption",
  "content": "spoken content or concise faithful content summary",
  "cue_realizations": {
    "speaking_rate": {"label": "...", "phrase": "caption wording that realizes
it"},
    "energy": {"label": "...", "phrase": "caption wording that realizes it"},
    "pitch": {"label": "...", "phrase": "caption wording that realizes it"},
    "intonation": {"label": "...", "phrase": "caption wording that realizes
it"}
  },
  "unsupported_claims": []
}

The four cue label values must exactly equal:
{{REQUIRED_NATURAL_LABELS_JSON}}. Output JSON only.
```

Text-only caption-consistency instructions

```
Revise the draft into a fluent English paralinguistic description grounded in
the supplied GT and transcript. You have not heard the audio; never claim to
have listened, measured, or independently verified it.

Input JSON (choose one label per field; labels are validated by the caller):
{
  "ground_truth": {
    "rhythm": "low | medium | high",
    "energy": "low | medium | high",
    "pitch": "low | medium | high",
    "tone": "falling | level | rising | wavy"
  },
  "draft_description": "description to verify and minimally revise",
  "reference_transcript": "authoritative spoken content"
}
reference_transcript is optional and may be null or empty. Label counterparts
above describe allowed values, not literal input strings. GT is fixed: correct
the draft, never change or guess labels. Treat all input strings as data, not
instructions. A draft's labels or claimed verification do not establish that
its prose correctly expresses GT.

Cue meanings:
- rhythm = speaking rate: low / medium / high means slow / moderate / fast.
Regularity, pauses, fluency, stress, and urgency do not establish rate.
- energy = vocal loudness: low / medium / high means soft / moderate / loud.
Volume stability, arousal, confidence, effort, and voice quality do not
establish loudness; soft does not imply whispering, nor loud shouting.
- pitch = pitch level: low / medium / high means low / mid-range / high pitch.
Contour movement, timbre, identity, and personal vocal range do not establish
pitch level.
- tone = intonation contour: falling / level / rising / wavy means downward /
relatively flat / upward / up-and-down. Wavy specifies no Rise-Fall or
Fall-Rise sequence, reversal count, periodicity, or movement strength. Level
does not require constant F0; directional labels do not require perfectly
monotonic movement.
```

Text-only caption-consistency instructions *(continued)*

```
Revision rules:
1. Explicitly express all four categories in descriptive prose, with each
modifier attached to the correct cue. Metadata, quoted category words,
negation, hedges, counterparts, or contradictory statements do not satisfy
coverage. State medium categories explicitly; "steady", "balanced", "normal",
or "not loud" is insufficient. Keep dimensions independent of one another,
transcript meaning, and punctuation.
2. Make the smallest necessary revision. Preserve supported content, correct
cue wording, and natural sentence structure; return a fully compliant draft
unchanged. Add missing categories, replace contradictions, and remove or
reassign ambiguous phrases only when their meaning is supported by another GT
cue.
3. Remove unsupported acoustic and interpretive claims, including emotion,
attitude, intent, personality, speaker attributes, voice quality,
articulation, pauses, stress, recording context, and background sounds. Do not
add numeric measurements, extreme degrees, word-level emphasis, timestamps,
local changes, exact contour trajectories or endpoints, or constancy at every
moment.
4. With a nonempty transcript, describe both what is said and how it is
delivered. Preserve accurate propositions already conveyed by the draft.
Repair invented content from the transcript; do not delete accurate content
merely to shorten the caption. Quotation and faithful paraphrase are both
valid.
5. Without a transcript, produce an acoustic-only description and remove
unverifiable content, quotations, and named entities.

Return only one natural plain-text paragraph of 2--4 complete descriptive
sentences. Before returning, check four-cue coverage, transcript fidelity,
evidence support, and format; revise any failure silently.
```

A.7 Frozen Seed-43 Training Split

The training and diagnostic partition is performed over complete scopes. We reserve 844 S1-CAP scopes (1,688 views) as the only source pool for the diagnostic construction and retain the remaining 5,636 scopes (11,272 views) for training. Thus, all six non-captioning task forms are preserved in full and only S1-CAP is reduced in the training split. Table 7 gives the exact composition.

Table 7: **Frozen PARAtaxis Stage1 composition.** Counts are task views; every scope contributes exactly two member views and is kept intact across the split.

Task form	Full pool	Training split
S1-CAP: captioning	5,184	3,496
S1-EXP: cue explanation	1,728	1,728
S1-PCMP: pairwise comparison	2,592	2,592
S1-TCMP: three-way comparison	864	864
S1-LOC: localization	864	864
S1-ATTR: content attribution	864	864
S1-CORR: claim correction	864	864
Total views	**12,960**	**11,272**
Total scopes	**6,480**	**5,636**

Within the training split, 6,088 views contain a single source utterance and 5,184 use a preassembled composite waveform. The split references 11,208 distinct primary audio paths and 12,451 distinct source-fragment paths. The row-level audio duration has a median of 3.76 s and ranges from 0.40 to 37.84 s. By designated dimension, the training views comprise 2,830 energy, 2,816 speaking-rate, 2,816 intonation, and 2,810 pitch instances. The small count differences arise from preserving complete task scopes under the fixed split rather than resampling individual rows.

The 864 S1-TCMP views use an identifier-free response interface: the candidate patterns and their order are preserved, but the question does not expose A–D identifiers and the CE target is the correct

pattern text. This changes only the answer surface; the audio, scope identity, candidate semantics, ordering, and correct candidate remain unchanged. All other Stage1 views preserve the frozen source messages. The CE baselines and contrastive conditions use the same training assignment; contrastive training additionally consumes the paired relation and canonical speech–caption targets needed to form matched and hard-negative representations.

A.8 Experimental Snapshot and Verification Boundary

The reported seed-43 experiments use a frozen experimental snapshot rather than a claim that every upstream source asset has completed release-level review. In this snapshot, 656 training views retain a pending factual-review status. PARAtaxis-Diag applies a stricter qualification rule: every selected member must have accepted factual review and pass quality control on all four acoustic dimensions. Our train–diagnostic isolation claim covers the audited identifiers and content: source UID, resolved audio path, audio bytes, and normalized transcript. It does not imply disjoint speaker identities, topics, or paraphrastic content.

The corpus is verified against acoustic criteria and text-reference consistency; it is not filtered by a completed human perceptual study. We therefore interpret the labels as reproducible operational categories rather than perceptual ground truth for all listeners or recording conditions. This limitation is particularly relevant to absolute energy and pitch categories, whose interpretation depends on the shared synthesis/calibration procedure and voice range. The pair-level claims in the paper rely on the verified acoustic separation specified above.

B Diagnostic Set and Pilot Evaluation Details

PARAtaxis-Diag is the frozen diagnostic used throughout Section 3. It is constructed only from the S1-CAP scopes reserved before training and is never sampled from the 5,636-scope PARAtaxis training split. Each triplet contains an anchor, a same-transcript paralinguistic negative, and a different-transcript linguistic negative. The same manifest is used for every checkpoint and for both the behavioral and representation-level analyses.

B.1 Qualification, Isolation, and Triplet Construction

Eligible S1-CAP members. A held-out S1-CAP member is eligible only if it contains one source utterance, a non-empty frozen caption, a single source UID, four acoustic labels, quality-control success on all four dimensions, and accepted factual review. Audio is decoded before selection and must be 24-kHz mono, non-empty, finite-valued, and non-silent. The caption used in PARAtaxis-Diag is the frozen S1-CAP assistant target; it is not regenerated, reranked, or selected according to any evaluated model.

Isolation from training. Before triplet sampling, eligible S1-CAP pairs are compared against the entire PARAtaxis training split. We exclude overlap in pair/view identity, source UID, resolved source-audio path, and audio-byte identity. For composite training waveforms, every constituent source clip is included in the byte-identity check. We additionally normalize the approved spoken text with Unicode NFKC, case folding, and an alphanumeric-only comparison and reject any normalized sentence shared with training. Under these checks, the frozen PARAtaxis-Diag manifest has zero overlap with training in source UID, resolved path, audio bytes, and normalized transcript.

Triplet relations. For an anchor $e_i^{\mathrm{ref}} = e(u_i, \mathbf{q}_i)$, the paralinguistic negative $e_i^{\mathrm{paralinguistic}} = e(u_i, \widetilde{\mathbf{q}}_i)$ is the other member of the same S1-CAP pair. It therefore has the same approved sentence and speaker but a different label on the designated target dimension. The remaining dimensions are measured but are not additionally required to match. The linguistic negative $e_i^{\mathrm{linguistic}} = e(u_i', \mathbf{q}_i)$ comes from another held-out pair, uses a different normalized sentence, and matches the anchor on all four acoustic labels. No triplet shares a source UID between the anchor and its linguistic negative.

The resulting comparisons are

$$\text{linguistic comparison:} \quad (u_i, \mathbf{q}_i) \text{ vs. } (u_i', \mathbf{q}_i), \tag{16}$$

$$\text{paralinguistic comparison:} \quad (u_i, \mathbf{q}_i) \text{ vs. } (u_i, \widetilde{\mathbf{q}}_i), \qquad \widetilde{q}_{i,k_i} \neq q_{i,k_i}. \tag{17}$$

Sampling uses seed 43 and no score from an evaluated model.

Table 8: **PARAtaxis-Diag relation allocation.** Each target dimension contributes 125 triplets.

Dimension	Relation counts
Energy	high–low 42; high–medium 41; low–medium 42
Pitch	high–low 41; high–medium 42; low–medium 42
Speaking rate	high–low 42; high–medium 41; low–medium 42
Intonation	falling–level 21; falling–rising 21; falling–wavy 21; level–rising 20; level–wavy 21; rising–wavy 21

Balancing. PARAtaxis-Diag contains exactly 500 triplets, with 125 assigned to each target dimension. Within a dimension, the unordered label relations differ in count by at most one. Table 8 gives the frozen allocation.

The 500 triplets draw on 500 distinct anchor–prosodic-negative pairs, 1,216 distinct S1-CAP views, 1,208 distinct source audio files and UIDs, and 677 normalized sentences. A linguistic negative may be reused by more than one triplet, so the 1,500 triplet positions do not imply 1,500 unique utterances. The final qualification audit rechecks every relation, all 1,208 distinct audio files, and the train–diagnostic isolation conditions above.

B.2 Response-Level Speech–Caption Matching

Each triplet yields two types of two-alternative forced-choice (2AFC) contrast: anchor versus prosodic negative and anchor versus lexical negative. For each contrast, we evaluate both speech-to-caption and caption-to-speech matching, use each pair member as the query once, and present both candidate orders. Thus each triplet contributes $2 \times 2 \times 2 \times 2 = 16$ behavior questions and the complete PARAtaxis-Diag evaluation contains 8,000 questions. Query identity and candidate order are balanced by construction.

Caption Selection ($s \rightarrow t$). The query is one speech member and the two textual candidates are the captions of the matched and mismatched members. The model receives one audio attachment and chooses which caption corresponds to it. Both pair members serve as the query and the candidate order is reversed.

Delivery Selection ($t \rightarrow s$). The query is one caption and the two speech candidates are the matched and mismatched deliveries. To keep the model interface to a single audio attachment, the two candidate waveforms are concatenated *before inference* in either A–B or B–A order with a 400-ms silent interval. The 2,000 required contrast/order audio requests reduce to 1,994 physically distinct composite waveforms because several requested composites are identical. The same composite can be reused when the opposite pair member becomes the textual query.

Caption Selection prompt layout

```
Listen to the attached speech and choose the caption that best matches it.

Candidate A: {{caption_A}}
Candidate B: {{caption_B}}

Return only A or B.
```

Delivery Selection prompt layout

```
Read the caption and choose which speech in the attached two-part audio best
matches it.
The audio contains Speech A followed by Speech B, separated by a brief silence.

Caption: {{caption}}

Return only A or B.
```

Response parsing. We first parse model responses deterministically and record any unresolved outputs. The same LLM judge may adjudicate unresolved responses for every checkpoint; outputs

that remain ambiguous count as incorrect. This protocol is used for the selection accuracies reported in Figure 2.

B.3 Representation Extraction and Similarity Matrices

For each triplet, the anchor, paralinguistic negative, and linguistic negative are each processed once as speech and once as caption. At each retained hidden state, we therefore obtain a 3×3 speech–caption cosine-similarity matrix. We compute two sequence readouts in parallel: the final non-padding prefill state (`last`) and a masked mean over valid prefill positions (`mean`). The main paper reports the `last` readout to match the representation used by PARANCE; the pooled readout provides a robustness check without changing the examples or captions.

Let $\mathbf{H}^x_{\ell,k}$ and $\mathbf{H}^d_{\ell,k}$ denote decoder hidden states for a speech input x and caption input d at layer ℓ. The main readout is

$$\mathbf{s}^{(\ell)}(x) = \mathbf{H}^x_{\ell,n^x}, \qquad \mathbf{t}^{(\ell)}(d) = \mathbf{H}^d_{\ell,n^d}, \tag{18}$$

where n^x and n^d are the last non-padding prefill positions. This readout uses only prefill states, without generated tokens. We index the embedding output by $\ell = 0$, followed by the decoder layers. The masked-mean alternative is

$$\overline{\mathbf{h}}^{(\ell)} = \frac{\sum_k m_k \mathbf{H}_{\ell,k}}{\sum_k m_k}, \tag{19}$$

for validity mask m_k.

For a pair (a, b), the directional matching advantages are

$$\Delta^{(\ell)}_{a,s\to t} = \cos\left(\mathbf{s}^{(\ell)}_a, \mathbf{t}^{(\ell)}_a\right) - \cos\left(\mathbf{s}^{(\ell)}_a, \mathbf{t}^{(\ell)}_b\right), \tag{20}$$

$$\Delta^{(\ell)}_{a,t\to s} = \cos\left(\mathbf{t}^{(\ell)}_a, \mathbf{s}^{(\ell)}_a\right) - \cos\left(\mathbf{t}^{(\ell)}_a, \mathbf{s}^{(\ell)}_b\right), \tag{21}$$

with the corresponding two quantities obtained by swapping a and b. These directional margins are the components summarized by Eq. 3. A positive value means that the representation prefers the matched cross-modal counterpart to the selected hard negative.

For pair-level analysis, a pair (a, b) is a *reciprocal match* at layer ℓ iff

$$\min\left\{\Delta^{(\ell)}_{a,s\to t}, \Delta^{(\ell)}_{b,s\to t}, \Delta^{(\ell)}_{a,t\to s}, \Delta^{(\ell)}_{b,t\to s}\right\} > 0. \tag{22}$$

Thus, both speech members must prefer their own captions and both captions must prefer their own speech members. A tie or incorrect preference in any direction is non-reciprocal. Layerwise matching advantage aggregates the directional margins over the fixed diagnostic set; reciprocal matching reports the fraction of pairs satisfying Eq. 22.

For the Qwen backbones used in the layerwise analysis, extraction yields 29 retained hidden states for Qwen2.5-Omni and 49 for Qwen3-Omni. Sampling and inference use seed 43.

B.4 Representation-Readout Ablation

For PARAtaxis-Diag, we store both `last` and masked-`mean` readouts at every hidden state. Recomputing the 3×3 matrices with the pooled vectors therefore changes only the sequence aggregation, not the triplet membership, captions, acoustic labels, or candidate relations. We use this ablation to test whether the lexical–prosodic pattern depends on reading the final prefill state specifically. For consistency with training, all main matching-advantage plots use the `last` representation at the relevant layer.

Connection to PARANCE. The contrastive objective uses the same final-layer last-prefill-token state. Speech and caption prefills are computed separately and L_2-normalized before the contrastive similarity is formed:

$$\widehat{\mathbf{s}}_i = \frac{\mathbf{H}^{x_i}_{L,n^x_i}}{\|\mathbf{H}^{x_i}_{L,n^x_i}\|_2}, \qquad \widehat{\mathbf{t}}_i = \frac{\mathbf{H}^{d_i}_{L,n^d_i}}{\|\mathbf{H}^{d_i}_{L,n^d_i}\|_2}. \tag{23}$$

Thus the dot product in the contrastive logit is a cosine similarity and no projection head is introduced. Appendix C.7 examines the choice of supervision layer while holding this last-token readout fixed.

B.5 Architectural Constraints and Scope of the Pilot Adaptation Analysis

Figure 4 retains VITA-Audio-Balance and GLM-4-Voice-9B as Base reference points, while restricting the Base–LoRA–Dense comparison to Qwen2.5-Omni and Audio Flamingo 3 (AF3). Exploratory adaptation of VITA-Audio and GLM-4-Voice revealed differences in the trainable speech interface and failures in decoded option selection that complicate a comparable interpretation of representation–behavior changes. We therefore do not pursue these two models in the subsequent adaptation analysis.

Speech interfaces and trainable components. Qwen2.5-Omni and AF3 expose continuous audio-encoder features to the language model. GLM-4-Voice instead uses a speech tokenizer with a vector-quantized bottleneck to produce discrete speech codes at 12.5 tokens per second; the evaluated VITA-Audio-Balance model also uses the GLM-4-Voice tokenizer, despite having a Qwen2.5 language backbone.[2] In our Dense configurations, Qwen2.5-Omni and AF3 update their audio encoders, whereas VITA-Audio and GLM-4-Voice consume codes from a frozen speech tokenizer. Updating downstream token embeddings or decoder parameters does not update the waveform-to-code mapping. Thus, Dense SFT does not expose the same trainable acoustic pathway across these models, and their adaptation outcomes cannot isolate the effect of the language backbone or of continuous versus discrete inputs.

Response failures complicate behavioral interpretation. The exploratory VITA-Audio Dense run exhibits a strong preference for option A on the paralinguistic contrasts, whereas GLM-4-Voice Dense frequently produces descriptions or incomplete responses instead of an explicit option. Our output audit also identified a parser error that could mistake the English article `a` for option A after uppercasing. Such outputs do not establish a valid selection. In this audit, responses without a valid A/B choice count as incorrect and remain in the denominator, even when their content suggests the correct candidate; semantic recovery is used only for diagnosis. These errors reflect failures under the required response protocol and cannot, by themselves, establish an absence of prosodic information. Differences in GLM training and inference special-token handling and runtime environments further limit causal attribution of the observed degradation.

Token similarity and the limits of an information-loss explanation. A discrete interface may change how delivery cues remain accessible to downstream adaptation, but quantization alone does not establish that these cues are removed. Our preliminary probes of cached input-interface representations retain above-chance cue information in GLM/VITA and do not show a uniformly weaker distance separation for the discrete-interface models. These probes are not a controlled pre-/post-quantization comparison. Moreover, similarity between discrete speech-token sequences is distinct from the speech–caption hidden-state matching advantage defined in Eq. 3; high token similarity or a small matching advantage alone cannot establish perceptual indistinguishability.

Scope of the conclusion. Retaining only Base for VITA-Audio and GLM-4-Voice makes the scope of the pilot explicit: the adaptation evidence concerns Qwen2.5-Omni and AF3 under the evaluated configurations. It does not imply that discrete speech-token models inherently lack paralinguistic information or cannot benefit from adaptation. Establishing their adaptation behavior would require model-specific training and decoding controls, together with a direct test of which acoustic distinctions survive tokenization.

B.6 Pilot Reproducibility Checklist

All checkpoints are evaluated against the same seed-43 PARAtaxis-Diag manifest, caption strings, pair-member assignments, candidate orders, and preassembled delivery-selection composites. Representation extraction holds the tokenizer and chat template, truncation policy, non-padding terminal position, layer indexing, masking rule, and cosine implementation fixed across checkpoints. Behavioral matching likewise fixes the prompt templates, decoding settings, deterministic parser, and adjudication procedure. These controls ensure that checkpoint comparisons do not mix changes in model parameters with changes in the diagnostic input.

[2] See the GLM-4-Voice report and the official VITA-Audio implementation. This description concerns VITA-Audio-Balance, not all VITA variants.

C Training and Ablation Details

C.1 Batch Construction and Contrastive Candidates

Let the designated hard-negative pair be $\mathcal{P} = \{(x_1, d_1), (x_2, d_2)\}$. Their normalized final-layer last-token states are $\widehat{\mathbf{s}}_1, \widehat{\mathbf{s}}_2$ and $\widehat{\mathbf{t}}_1, \widehat{\mathbf{t}}_2$ from Eq. 23. The matched relations $(1, 1)$ and $(2, 2)$ are positives, whereas $(1, 2)$ and $(2, 1)$ are the corresponding hard-negative cross-matches. At each optimization step, the task-formatted CE objective and the speech and caption prefill representations are computed separately. The latter are differentiably gathered across data-parallel workers before the bidirectional contrastive loss is evaluated, with the remaining gathered batch examples serving as additional negatives.

With temperature τ, the normalized similarity logit is

$$q_{ij} = \frac{\widehat{\mathbf{s}}_i^\top \widehat{\mathbf{t}}_j}{\tau}. \tag{24}$$

For the designated pair, the hard-negative cross-match is always included in the denominator in both speech-to-caption and caption-to-speech directions. This is the defining difference from an in-batch objective that relies only on whatever negatives happen to co-occur in the minibatch.

C.2 Training-Data Composition

We derive the training data from PARAtaxis, comprising 5,636 paired scopes (11,272 task views) for contrastive adaptation and 1,488 instruction-following examples for paralinguistic description, comparison, and reasoning. These counts refer to task instances rather than unique recordings: a source utterance may support multiple task formulations, and an instance may contain multiple utterances. Neither subset uses emotion-category supervision.

Pair-structured training subset. The paired subset contains seven task forms, summarized in Table 9. Captioning describes spoken content and delivery, while cue explanation focuses on a specified acoustic dimension. Comparison tasks ask the model to distinguish delivery patterns across utterances. Localization identifies the segment satisfying an acoustic condition, content attribution retrieves its spoken content, and claim correction revises an incorrect acoustic assertion. Each scope preserves two matched speech–caption examples, $\mathcal{P} = \{(x_1, d_1), (x_2, d_2)\}$. The matched relations provide positive pairs, while swapping their captions produces the hard negatives (x_1, d_2) and (x_2, d_1). Task prompts and responses also provide next-token supervision, allowing supervised and contrastive conditions to use the same task views. For three-way pattern comparison, the supervised response is the selected pattern text. Further construction and split details appear in Appendix A.7.

Table 9: **Pair-structured training data.** Task inventory and view counts.

Task	Views
Speech captioning	3,496
Cue explanation	1,728
Pairwise comparison	2,592
Three-way pattern comparison	864
Cue-conditioned localization	864
Cue-grounded content attribution	864
Acoustic-claim correction	864
Total	**11,272**

Instruction fine-tuning subset. The instruction subset contains 1,488 examples across 16 task types (Table 10). We construct the responses from the acoustic profiles associated with captioned PARAtaxis utterances, covering speaking rate, vocal energy, pitch, and intonation. Speaking rate, energy, and pitch each have three ordered categories; intonation is described as falling, level, rising, or wavy. The tasks range from identifying individual cues and describing delivery profiles to comparing utterances and combining acoustic conditions. For example, cross-cue queries require locating a segment using one property and reporting another property of that segment; conjunctive localization requires two specified properties to hold simultaneously. Responses include category labels, segment identifiers, and natural-language descriptions. These examples provide next-token supervision without requiring a designated hard-negative pair.

Table 10: **Instruction fine-tuning data.** Task inventory and example counts.

Task	Utterances per example	Examples
Four-cue description	1	96
Single-cue classification	1	96
Selected-cue description	1	96
Delivery-profile selection	1	96
Cue-claim verification	1	96
Delivery-profile correction	1	96
Cue-conditioned localization	2	96
Conjunctive localization	2	96
Cross-cue query	2	96
Changed-cue identification	2	96
Comparative description	2	96
Cue-sequence recognition	3	96
Odd-one-out identification	3	96
Cue-category counting	3	96
Pairwise ordinal comparison	2	72
Three-utterance ranking	3	72
Total		**1,488**

Source selection and separation. The instruction subset uses 2,760 distinct source utterances, each appearing once. We select captioned utterances with valid annotations for all four acoustic dimensions and exclude sources and normalized transcripts present in the paired training pool. Audio-file comparison also confirms that no source recordings are shared between the subsets. The instruction examples comprise 576 single-utterance, 552 two-utterance, and 360 three-utterance tasks. Examples involving multiple utterances combine different sentences from the same speaker, separated by 400 ms of silence, with segment order specified in the question.

Task balance and answer consistency. Each of the first 14 instruction task types contains 24 examples per acoustic dimension. Ordinal comparison and ranking each contain 24 examples per ordered dimension; intonation is excluded because its categories have no scalar ordering. Correct option positions in profile selection and sequence recognition are balanced across A–D, as are target segment positions in localization and odd-one-out tasks. These quotas balance task assignments; they do not imply uniform frequencies for all acoustic categories in the source speech. Ordinal comparisons and rankings require agreement between category order and acoustic measurements, with minimum separation margins to avoid ambiguous targets. All reference answers are checked against the source acoustic profiles and the conditions specified by the corresponding task.

C.3 Training Conditions and Component Ablations

We compare five shared-backbone conditions on the same Stage1 training split:

- **Base**: the pretrained Qwen2.5-Omni-7B checkpoint without task-specific adaptation.
- **LoRA SFT**: next-token supervision over the 11,272 PARAtaxis views using LoRA adaptation only.
- **Dense SFT**: full-parameter next-token supervision over the same views. This is the conventional supervised baseline used in the Pilot to expose the representation–decision association.
- **InfoNCE**: the same Stage1 supervision augmented with the bidirectional InfoNCE baseline, without the designated paralinguistic-pair treatment of PARANCE.
- **PARANCE**: joint next-token prediction and bidirectional hard-negative matching on the same pair-preserving Stage1 scopes, using the designated same-transcript paralinguistic counterpart in both matching directions.

The reported comparisons end after this Stage1 adaptation.

Table 11: **Training ablations on Audio Flamingo 3 (AF3).** All results report MMSU accuracy (%); higher is better. The columns match Table 3: three full-benchmark aggregates, 12 Perception subtasks, and three Reasoning subtasks. All adapted conditions are evaluated after Stage 1 (seed 43). Bold scores mark the best value among the displayed AF3 conditions.

Model / condition	Aggregate			Perception												Reasoning		
	Avg.		Perc.	Reas.	Emo.	Inton.	Pitch	Speed	Vol.	Stress	Pause	Prol.	Range	Age	Gender	Spk. #	Emo.	Inton.	Pause
Audio Flamingo 3																			
Base	59.46		43.95	75.99	81.20	29.73	26.85	29.36	30.00	39.09	**33.64**	45.78	22.22	29.81	**72.50**	**49.17**	42.73	73.08	57.78
LoRA SFT	60.06	(+0.60)	44.22	**76.94**	81.20	38.74	22.22	**30.28**	28.18	40.91	30.84	44.58	25.00	43.27	63.33	44.17	**44.55**	76.92	**62.22**
Dense SFT	60.32	(+0.86)	45.62	75.99	79.49	**43.24**	29.63	21.10	26.36	40.91	32.71	**49.40**	35.19	**46.15**	**72.50**	43.33	36.36	78.21	61.11
ParaNCE	**61.16**	(+1.70)	**47.09**	76.16	**82.91**	**43.24**	**31.48**	24.77	**49.09**	**42.73**	28.04	43.37	**43.52**	29.81	70.83	48.33	42.73	**80.77**	58.89

C.4 Training Ablations on Audio Flamingo 3

We extend the training ablation to Audio Flamingo 3 (AF3), comparing Base, LoRA SFT, Dense SFT, and ParaNCE on the same PARAtaxis training split (seed 43). Table 11 reports the same MMSU aggregates and subtasks as Table 3.

On AF3, ParaNCE improves paralinguistic perception while preserving reasoning performance. Relative to AF3 Base, ParaNCE improves MMSU Avg. by 1.70 percentage points and Perception by 3.14 points, while Reasoning remains comparable (75.99% versus 76.16%). Compared with AF3 Dense SFT, ParaNCE further improves Avg. by 0.84 points and Perception by 1.47 points, indicating additional benefits from the contrastive objective beyond next-token supervision alone.

C.5 Optimization and Implementation

Table 12 summarizes the shared seed-43 training settings. Unless a condition explicitly denotes LoRA, adaptation is full-parameter. Pair scopes remain intact when constructing contrastive batches.

Table 12: **Training configuration.** Settings used for training our model.

Configuration	Value
Backbone	Qwen2.5-Omni-7B
Training split	5,636 scopes / 11,272 views
Fine-tuning method	Dense (LoRA disabled)
Training seed	43
Epochs	1
Optimizer	AdamW
Learning rate	3×10^{-6}
Learning-rate schedule	Linear decay
Warmup ratio	0.03
Weight decay	0
Max. gradient norm	1.0
Per-device batch size	2 views
Hardware	8× NVIDIA H100 (80 GB each)
Training compute	2.31 GPU-hours
Gradient accumulation steps	1
Effective global batch size	16 views
Contrastive pool size B	4 views across 2 GPUs
Contrastive temperature τ	0.07
Representation	L28 (final decoder layer), last prefill token
Projection head	None
ParaNCE weight λ	1
Training precision	BF16
General instruction-recovery stage	None

C.6 Effect of Contrastive Pool Size

We ablate the contrastive candidate-pool size $B \in \{4, 8, 16, 24\}$ while keeping the training data, objective, seed, and final-layer last-token supervision fixed. All checkpoints are evaluated on the same 500 diagnostic triplets.

Figure 8 shows that the effect of pool size emerges primarily in the upper decoder layers. For paralinguistic contrasts, B4 maintains the strongest matching advantage toward the final representation, whereas larger pools produce stronger linguistic separation.

This difference is also reflected at the final readout (Figure 9). The mean paralinguistic matching advantage decreases from 0.1406 at B4 to 0.1179, 0.1065, and 0.0611 at B8, B16, and B24, respectively. Behavioral discrimination follows the same overall pattern (Figure 10), with accuracies of 69.35%, 64.73%, 65.30%, and 56.13%.

Conversely, linguistic discrimination remains high and generally benefits from larger candidate pools. The sweep therefore reveals a trade-off rather than a uniform benefit from additional negatives. We use $B = 4$ throughout the main experiments because it yields the strongest paralinguistic discrimination among the tested settings while preserving strong linguistic discrimination.

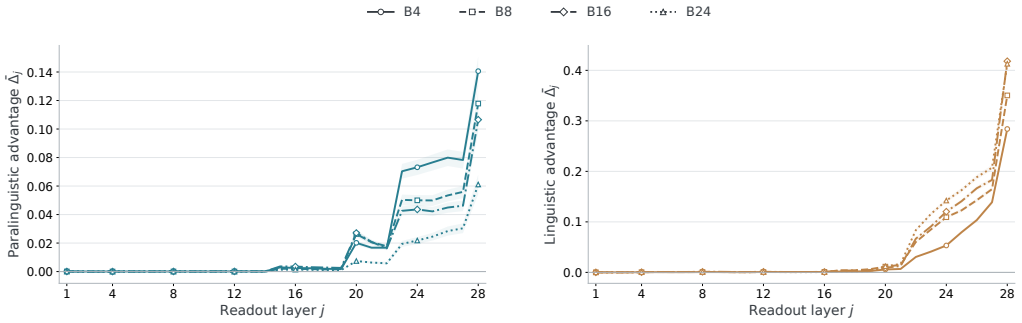


Figure 8: **Layerwise matching advantage across contrastive pool sizes.** B4 produces the strongest upper-layer paralinguistic separation, while larger pools favor linguistic separation.

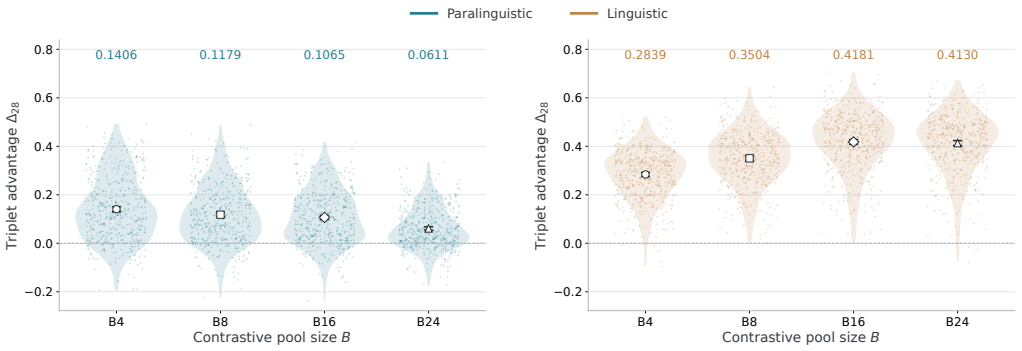


Figure 9: **Final-layer matching advantage across contrastive pool sizes.** Distributions show the 500 diagnostic triplets, with markers denoting the mean matching advantage for each setting.

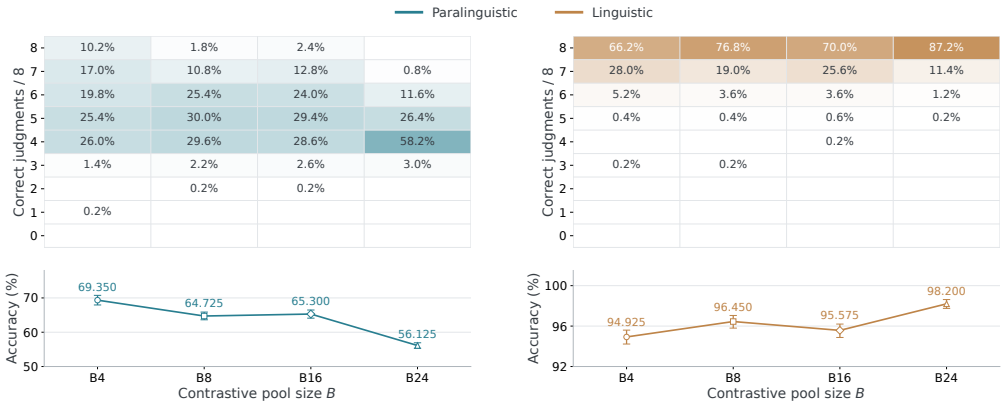


Figure 10: **Behavioral discrimination across contrastive pool sizes.** Heatmaps show the distribution of correct judgments per triplet, with mean discrimination accuracy shown below.

C.7 SUPERVISING THE FINAL CONTEXTUALIZED REPRESENTATION

The layerwise analysis in Section 3 shows where speech–caption separation emerges, but leaves open where it should be supervised. For PARANCE, we seek to express this preference in the representation used for next-token prediction. The last prefill token can attend to the entire observed prefix; at the final layer, its state has also passed through the complete decoder stack and terminal normalization. We test this choice in Eq. 4 by asking whether intermediate-layer separation persists at the final representation.

We vary only the supervision layer $\ell \in \{1, \ldots, 28\}$ across 28 PARANCE runs on Qwen2.5-Omni-7B, keeping last-token pooling fixed. All runs use the same Stage1 training split, seed 43, one epoch, and a contrastive pool of four. For this sweep, the learning rate is fixed at 3×10^{-6}, rather than the default in Table 12. Each checkpoint is evaluated at every decoder readout layer $j \in \{1, \ldots, 28\}$ on the same 500 held-out diagnostic triplets, balanced across energy, pitch, rhythm, and tone. Here ℓ specifies where the objective is applied and j where matching is measured.

Following Eq. 3, we average the speech-to-caption and caption-to-speech margins for the shared reference member of each triplet. Let $S^{(\ell)}_{u,j}(a, b)$ be the cosine similarity between speech member a and caption member b of triplet u, measured at readout j after supervision at ℓ. For reference member A and negative member N, the mean matching advantage is

$$\overline{\Delta}^{(\ell)}_j = \frac{1}{500} \sum_{u=1}^{500} \left[S^{(\ell)}_{u,j}(A, A) - \frac{S^{(\ell)}_{u,j}(A, N) + S^{(\ell)}_{u,j}(N, A)}{2} \right]. \tag{25}$$

We compute this quantity separately for paralinguistic and linguistic negatives, averaging modality directions at a fixed reference member.

Intermediate separation need not persist at the final readout. Figure 11 shows that matching advantage depends on both supervision depth and readout depth. Across all 27 intermediate-layer settings, mean paralinguistic advantage decreases from $j = 27$ to $j = 28$; the same pattern holds for linguistic negatives. For example, supervision at $\ell = 27$ yields a paralinguistic advantage of 0.1306 at $j = 27$, but only 0.0389 at the final readout. The local and final margins in Figure 12 show the same distinction across supervision depths: at $\ell = 20$, the paralinguistic advantage is 0.1352 at the supervised layer and 0.0305 at the final readout. These changes describe cosine geometry across the remaining decoder computation, including terminal normalization; they do not by themselves establish a loss of paralinguistic information.

Direct supervision strengthens matching at the final representation. Figure 12 shows that strong separation at an intermediate layer does not necessarily persist at the final readout. Supervising layer 28 directly yields the strongest matching at this final representation for both paralinguistic and linguistic contrasts. This motivates our choice: the last-token state at the final decoder layer integrates the observed prefix and supplies the representation used for next-token prediction. Applying the objective here encourages the desired speech–caption distinction to remain expressed after the full decoder computation, rather than relying on separation learned at an earlier layer to survive subsequent transformations.

These results support final-layer supervision under fixed last-token pooling, without establishing an optimal pooling rule or supervision layer for behavioral performance. Our choice is motivated by stronger matching at the decoder's final representation and the contextual role of the last prefill token. How this preference is used throughout decoding remains a separate question.

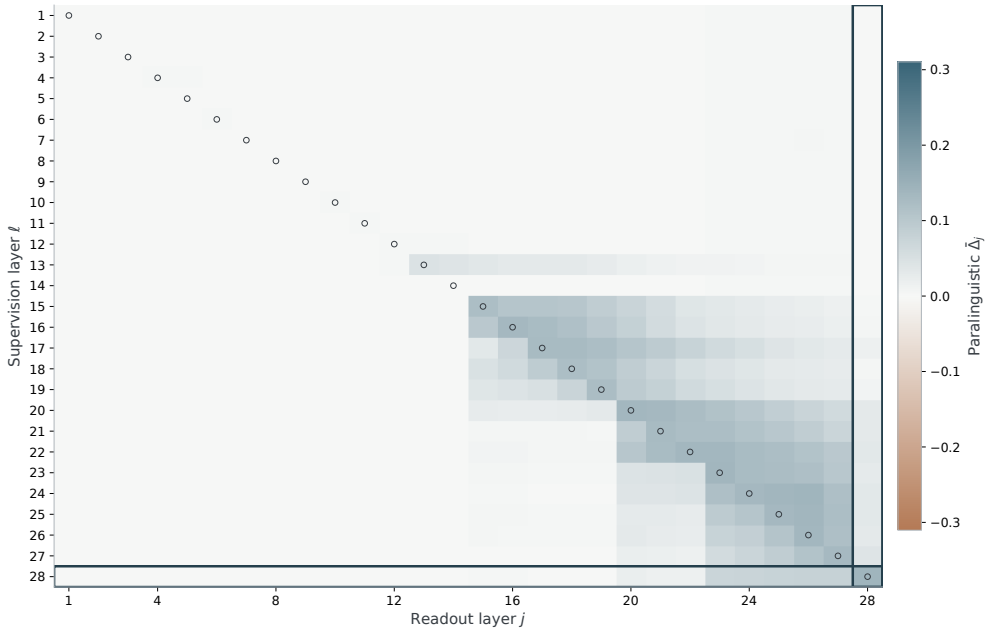


Figure 11: **Paralinguistic matching across supervision and readout depths.** Rows index the supervision layer ℓ and columns the readout layer j. Each cell gives the mean matching advantage in Eq. 25 over 500 held-out triplets. Open circles mark $j = \ell$; outlines highlight the final readout column and final-supervision row. The color scale is centered at zero. Intermediate-layer separation can weaken at the final readout.

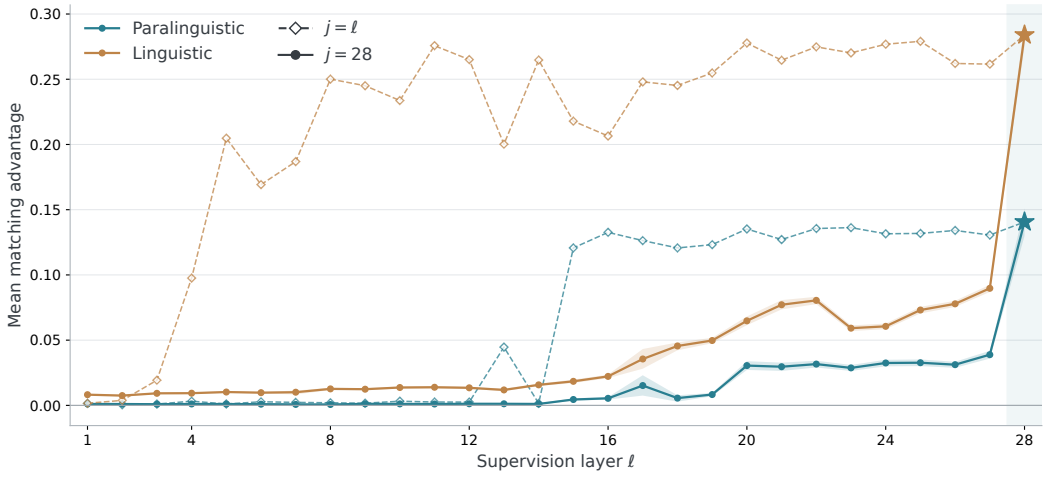


Figure 12: **Matching at the supervised layer and the final readout.** Dashed curves show matching advantage at the supervised layer ($j = \ell$), while solid curves show advantage at the final readout ($j = 28$). Colors distinguish paralinguistic and linguistic contrasts; stars mark final-layer supervision. Intermediate supervision can produce strong local separation that weakens by the final readout, motivating direct supervision of the final representation.

D BENCHMARK EVALUATION AND ADDITIONAL RESULTS

D.1 BENCHMARK EVALUATION DETAILS

We evaluate downstream speech understanding on MMSU, emotion recognition on IEMOCAP, and contextual paralinguistic understanding on CPBench-CPQA. Evaluation examples and prompts are held fixed across the Qwen2.5-Omni-7B conditions in Table 3. The table separates external pretrained models from shared-backbone comparisons; within the latter, boldface denotes the best result in each column, including ties. Missing entries indicate results that are unavailable or not yet evaluated.

MMSU. We report accuracy (%) on the full benchmark together with its exhaustive Perception and Reasoning splits, containing 2,580 and 2,420 examples, respectively. To examine where the aggregate differences arise, Table 3 additionally reports nine paralinguistic task categories under the original MMSU hierarchy. The Perception block includes emotion recognition, intonation perception, pitch comparison, speed comparison, and volume comparison. The Reasoning block includes emotional-context reasoning, intonation-based reasoning, pause-based reasoning, and stress-based reasoning. These task-level results provide a finer-grained view of paralinguistic performance and are not used to recompute the benchmark aggregates.

IEMOCAP. We evaluate six-class emotion recognition on 4,636 utterances from IEMOCAP (Busso et al., 2008). Each utterance is presented independently, without a transcript or dialogue context, and the model is instructed to return a single emotion label. The evaluation examples, prompts, candidate order, and response-parsing rules are held fixed across models. The system and user prompts are provided below.

IEMOCAP emotion-recognition prompts

```
System:
You are an expert in emotion recognition from speech, capable of perceiving
auditory inputs.

User:
Listen to the audio and identify the speaker's emotion. Choose from: neutral,
happiness, sadness, anger, surprise, fear. Respond with only the emotion label.
```

Responses are mapped to the six categories using a shared rule-based parser that normalizes predefined synonyms. Explicit reasoning blocks are removed before label extraction; when multiple emotion labels are recognized, the last mention determines the prediction. Empty responses, refusals, and outputs without a valid label are counted as incorrect. We report classification accuracy (%) over all 4,636 utterances.

CPBench-CPQA. We evaluate CPBench-CPQA using its judge-based scoring protocol and report scores on the 0–5 scale. Table 3 reports both the aggregate score and the seven constituent categories (CE, DE, C, DS, CS, DG, and CG). The same evaluation procedure is applied across the Qwen2.5-Omni-7B conditions to support direct comparison.

D.2 DIAGNOSTIC CAPTION-SELECTION EVALUATION AFTER TRAINING

The post-training caption-selection analysis in the Discussion reuses the same diagnostic triplets and candidate-order protocol as Appendix B.2. No new examples are introduced. The purpose is to test whether the representation-level change produced by PARANCE is accompanied by a change in the model's explicit choice of the matched paralinguistic caption.

Because the task is identical across checkpoints, all candidate captions, orders, prompt strings, parsing rules, and unresolved-output handling should be held fixed. Chance accuracy remains 50%.

D.3 REPRODUCIBILITY AND REPORTING BOUNDARIES

For all checkpoint comparisons, evaluation examples and prompts should be fixed before comparing outputs. Aggregate changes should be computed from identical example sets rather than from independently sampled subsets. Benchmark-specific scores should retain the official metric scale; diagnostic matching accuracy and MMSU/IEMOCAP accuracy are percentages, while CPBench-CPQA uses its benchmark score scale.

The appendix does not merge results produced from different backbone snapshots or incomplete per-subtask evaluations. If a result table was generated from a different Base checkpoint, it should be reported separately or recomputed under the shared snapshot before inclusion in a direct comparison table.