Benchmarking Sequential Speech Emotion Recognition in Multi-turn Conversations with Audio Language Models

Melissa Lieffers and Shuju Shi and Charlotte Zhu and Thai Le
Indiana University Bloomington
{mealieff,shi16,zz74,tle}@iu.edu

Abstract

Speech emotion recognition (SER) benchmarks for AI speech models typically evaluate isolated clips, providing limited evidence of whether models can track emotion over continuous speech like in multi-speaker conversations. Humans, however, detect emotion not only at a single point in time but also track its dynamics throughout natural conversation, which is critical in enabling us to respond naturally to others. Recent advances in audio-language models (ALMs) show great potential, yet benchmarking of this crucial capability remains lacking. In this paper, we therefore aim to understand how well ALMs perform on sequential SER on natural human conversational speech. We compare pretrained and LoRA fine-tuned ALMs with supervised non-ALM baselines, evaluating categorical emotion recognition and temporal localization. Pretrained ALMs perform poorly on sequential SER task while fine-tuning substantially improves classification and temporal overlap, although performance, particularly localization, declines for longer sequences. Analyses of emotion transitions and valence–arousal–dominance (VAD) contours show that fine-tuned ALMs follow the main emotional trajectories but do not recall nuanced contours like human labels. Our findings call for a shift in how ALMs are evaluated: toward protocols that jointly assess what emotion is expressed and when it occurs over variable-length conversational audio.

1 Introduction

Speech emotion recognition (SER) is an important component of spoken dialogue systems, enabling systems to interpret a speaker's affective state and respond appropriately (Zhou et al., 2023; Kosgi et al., 2022). However, emotion is not a fixed property of an utterance: it emerges, persists, and changes over time. Reliable automated approaches for emotion tracking therefore require identifying not only *what* emotion a speaker expresses, but also *when* it begins, persists, and transitions to others. Tracking both *what* and *when* across a conversational speech is even more important, as human speakers typically match or change speech emotions overtime during a conversation. Therefore, this capability is critical to numerous real-life applications, especially those involving extended conversational audio, including customer-service interactions (Feng and Devillers, 2023), clinical or therapeutic settings (Denecke and Gabarron, 2024; Mak et al., 2026), and potential direct human-AI interactions via speech channel in physical world.

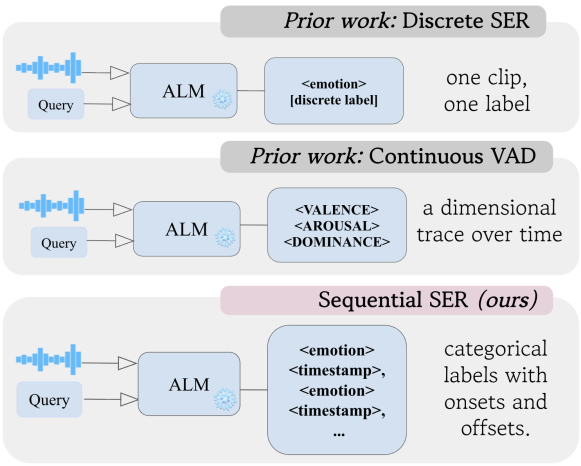


Figure 1: Our framework bridges traditional approaches to discrete and continuous SER by jointly predicting categorical emotions and their temporal boundaries directly from unsegmented audio.

Despite this temporal requirement, SER is commonly evaluated on isolated, predefined speech segments, and this also applies to advanced AI speech models such as audio-language models (ALMs). In particular, existing emotion benchmarks for ALMs largely emphasize clip-level recognition (Hu et al., 2025; Zhang et al., 2025) or continuous VAD predictions (Naini et al., 2025) (Fig. 1). However, such evaluations provide limited evidence of whether ALMs can maintain accurate, human-

understandable emotion predictions (e.g., *happy*, *sad*, *angry*) together with the temporal alignment that is important for understanding long conversational audio. Although prior work has identified limitations in audio models' temporal grounding and localization (Wang et al., 2025; Sun et al., 2026; Bhattacharya et al., 2025; Sridhar et al., 2024), the effects of increasing sequence length on categorical emotion recognition and temporal localization, especially when evaluated together, remain insufficiently characterized. This makes sequential SER evaluation very timely especially to objectively assess the technological maturity and readiness of advanced ALMs for real-life applications when they are increasingly adopted in the AI speech community in both academic and industry settings. Yet, few, if any, works have addressed this problem.

Therefore, this work investigates this gap by evaluating ALMs on continuous audio constructed from natural human conversational speech, using podcast audio as the focused settings, with sequence lengths ranging from 3-5 seconds up to nearly 3 minutes. We evaluate competitive ALMs including Qwen2-Audio (Chu et al., 2023; Bai et al., 2023), Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025), Audio-Flamingo-3 (Goel et al., 2025), and Kimi-Audio (KimiTeam et al., 2025) in zero-shot and LoRA fine-tuned settings, alongside a non-ALM baseline incorporating techniques from leading systems from the *MSP-Podcast challenge* (Chatzichristodoulou et al., 2025; Feng et al., 2025; Dutta et al., 2025) that further improved for our task. Models are tasked to generate ordered emotion labels with onset and offset timestamps, allowing us to evaluate both *temporal categorical emotion recognition* and *temporal localization* separately using macro-F1 and temporal intersection-over-union (tIoU), respectively. **Our contributions are summarized as follows.**

1. We are the first work that investigates evaluation of ALMs on sequential SER with timed output. The novelty of this approach also considers naturalistic multi-turn sequential SER evaluation with many emotion changes per recording.
2. We present the first fine-tuning of ALMs to generate emotion spans as interleaved timestamps, facilitating a span-level error analysis of timed emotion output from ALMs.
3. Typical classification metrics alone cannot reveal whether a model tracks how emotion changes over time. We therefore evaluate temporal localization and contour-level behavior (emotion transitions, n-gram patterns and VAD contours) alongside categorical metrics, and show that fine-tuned ALMs preserve classification performance while producing over-smoothed and less accurately localized contours than human perception.

2 Related Work

2.1 Naturalistic Speech Emotion Recognition

MSP-Podcast is a major benchmark for categorical SER in spontaneous speech, providing naturalistic speaking turns typically evaluated as independent segments (Busso et al., 2025; Lotfian and Busso, 2019; Martinez-Lucas et al., 2024). Although naturalistic corpora capture the variability of spontaneous emotional expression, segment-level evaluation does not establish whether models can track emotion over continuous audio. MSP-Conversation provides complementary continuous-time dimensional annotations, including valence and arousal, but does not provide the same categorical emotion labels (Martinez-Lucas et al., 2020). Recent shared tasks have established competitive MSP-Podcast baselines using multimodal representations, soft-label learning, and class-imbalance handling (Goncalves et al., 2024; Naini et al., 2025; Chatzichristodoulou et al., 2025; Feng et al., 2025); other work addresses annotator disagreement and label uncertainty (Chou et al., 2022; Omidi and Hansen, 2026). These methods advance categorical SER but leave open how recognition and temporal localization behave when models process increasingly long sequences without predefined boundaries.

2.2 Temporal Understanding in Audio Models

Prior work investigates temporal grounding and acoustic event localization in audio models, highlighting challenges in associating predictions with specific time intervals (Wang et al., 2025; Sun et al., 2026; Bhattacharya et al., 2025; Sridhar et al., 2024). Separately, research on long-form speech processing documents input-length constraints and performance degradation as audio sequences grow (Feng et al., 2025; Dutta et al., 2025; Lin and Busso, 2021). However, processing extended audio and localizing emotional states are distinct challenges: emotional expressions may persist across utterances or transition without clear acoustic boundaries. These findings motivate evaluating whether ALMs can maintain both categori-



Figure 2: Example of four consecutive podcast segments concatenated into one 11.52 s recording. The model must recover both the emotion sequence (*Disgust*, *Happy*, *Neutral*, *Sad*) and the time of each change.

cal recognition and temporal precision as sequence length increases.

2.3 Emotion Evaluation for ALMs

Existing ALM emotion benchmarks primarily evaluate emotional understanding at the clip level, providing limited evidence of performance on temporally localized emotion tracking over continuous speech (Hu et al., 2025; Zhang et al., 2025). Timestamped long-form speech systems, including VibeVoice-ASR (Peng et al., 2026), provide a precedent for generating temporally structured predictions, although linguistic alignment differs from estimating the temporal extent of emotional states. We extend this evaluation perspective by assessing ALMs on consecutive MSP-Podcast segments at increasing sequence lengths, measuring categorical recognition and temporal localization separately to characterize how these capabilities change with sequence length and fine-tuning.

3 Problem Formulation

Sequential SER Task. We formulate the sequential SER task under conditions that match how a deployed listener receives speech. Consider a natural human-human conversation consisting of N consecutive speaking turns or segments $x_1, \ldots, x_N$ where each segment x_i carries a ground-truth emotion label $e_i \in \mathcal{E}$, and $\mathcal{E}$ is the set of eight emotions: Angry, Sad, Happy, Surprised, Fear, Disgust, Contempt, and Neutral. Concatenating the segments back to back yields a single recording X and the ground truth for X is the ordered list of timed emotion spans:

$$Y = \{(t_i^s, t_i^e, e_i)\}_{i=1}^{N} \quad (1)$$

where t_i^s and t_i^e are the onset and offset of the i-th span. In the sequential SER task, a model hears X as one recording, with no boundaries marked, and must recover Y. It returns an ordered list of *timed* emotion spans:

$$\hat{Y} = \{(\hat{t}_i^s, \hat{t}_i^e, \hat{e}_i)\}_{i=1}^{\hat{N}}, \quad (2)$$

where $\hat{t}_i^s$ and $\hat{t}_i^e$ are the onset and offset of the i-th predicted span and $\hat{e}_i \in \mathcal{E}$. The model predicts both the number of spans $\hat{N}$ and their timing, and decodes in a single pass. This output follows the timestamped transcripts of long-form speech recognition (Peng et al., 2026).

Natural Conversational Speech Data. Natural human conversational speech data with discrete, high-quality emotion labels are lacking in the current literature. We opt to use MSP-Podcast (Busso et al., 2025) (v1.12) data, which is built from real-life podcast recordings by an automatic pipeline that detects speech, discards regions with music, overlapping speakers or low signal-to-noise ratio (SNR), and splits the remaining speech into speaking turns of 3–11 s. Formally, each segment $x_{p,k}$ is the k-th speaking turn of podcast p, with duration $d_{p,k} \in [3, 11]$ s, and is annotated in isolation by human annotators whose give its label $e_{p,k} \in \mathcal{E}$. The corpus therefore contains no emotion labels above the segment level. However, segment identifiers encode the source podcast p and the position k within it, which lets us recover the original, natural temporal order, which is necessary for our task. Fig. 2 shows an example of four consecutive segments.

4 Evaluation Setup

4.1 Audio-language Models (ALMs)

We focus on evaluating open-source models due to the ability of fine-tuning them: Audio-Flamingo-3 (AF3) (Goel et al., 2025), which is trained on long audio inputs, Qwen2-Audio-7B-Instruct (Chu et al., 2023), which shares the Qwen2 family of LLM backbones, Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025), a recent state-of-the-art open-weight ALM, and Kimi-Audio-7B-Instruct (KimiTeam et al., 2025), which use audio-encoding design that is different than the others. All models receive the same fixed instruction, asking for one line per emotion span with its start time, end time and label, and use greedy decoding.

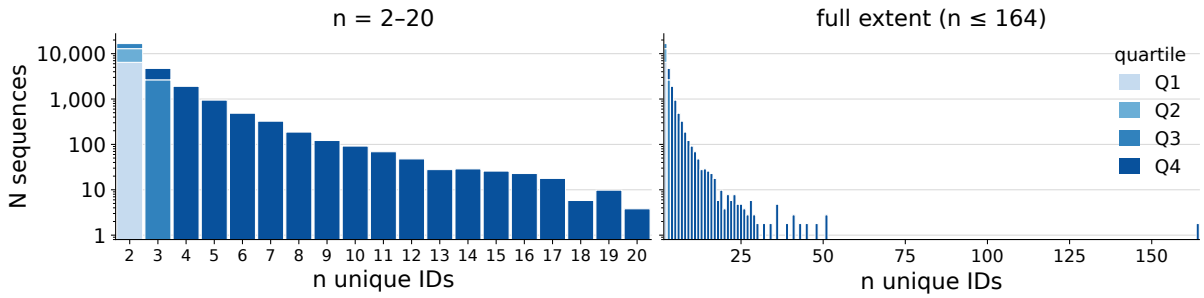


Figure 3: Near-contiguous sequence length distribution across MSP-Podcast (log-scale frequency), split into a zoomed view (n=2–20) and the full extent. Bars are shaded by which of the four equal-count quartiles that length falls into. Q1 and Q2 both fall entirely within n=2–3, and Q4 spans the entire remaining tail.

We test these models in both pre-trained (PT) and supervised fine-tuned (SFT) setting. In SFT setting, we train LoRA adapters (Hu et al., 2021) on the language model for three epochs on sequences from Train split. The target lists the gold spans, with times to two decimals. Because short sequences dominate Train split, each sequence's loss is weighted by the inverse frequency of its length bin. We keep the epoch with the lowest cross-entropy on Development.

4.2 Dataset

We use the official partitions: Train for training, Development for model selection and Test1 from the MSP-Podcast dataset which collects podcast conversations in English. Within each partition, we chain segments with strictly consecutive indices from the same podcast into a sequence. Segments labeled *Other* or *No agreement* lack a usable consensus label. We treat them as breakpoints that end the current sequence, so every segment in a sequence carries a plurality-vote emotion label. Listening checks confirmed that consecutive segments are contiguous in the source recording, so we concatenate them with no inserted silence, and gold boundaries are their cumulative durations.

This gives 9,495 Train, 2,557 Development and 3,648 Test1 sequences, with up to 28 segments (Table 1).[1] Length distributions are nearly identical across splits, and about 71% of sequences have two segments, so we report results in four length bins: L=2, L=3, L=4–6 and $L\geq$7 (Fig. 3). All systems are trained and evaluated on these same sequences. Because MSP-Podcast partitions are speaker-independent, most sequences contain a single speaker (85.0% of Test1 and 93.2% of Development); 14.7% of Test1 sequences contain two speakers, and only 10 contain three.

Split	$L=2$	$L=3$	$L=4$–6	$L\geq 7$	Total
Train	6,729	1,626	922	218	9,495
Dev	1,810	428	267	52	2,557
Test1	2,604	590	371	83	3,648
Total	**11,143**	**2,644**	**1,560**	**353**	**15,700**

Table 1: Sequence statistics for MSP-Podcast 1.12. Sequences are maximal runs of strictly consecutive segment IDs from one podcast within one split, concatenated back to back; segments labeled O or X are dropped and end a run. L is the number of segments per sequence ($2 \leq L \leq 28$).

4.3 Baselines

We compare our evaluation across open-source ALMs and also with a competitive non-ALM baseline: WavLM (IS 2025).

WavLM is the released categorical emotion prediction model from the Interspeech 2025 challenge (Naini et al., 2025), using WavLM-large (Chen et al., 2022) with attentive statistics pooling to classify individual gold segments into one of eight emotion classes. Since WavLM is not designed for sequential SER, we extend it, resulting in *WavLM-TimeAug* that enables both temporal emotion recognition and localization. The model adds two branches: a boundary detector that predicts scores at 100 ms intervals, and a segment-level emotion classifier that applies attentive pooling within predicted segments and a bidirectional LSTM across segments, allowing each prediction to incorporate neighboring context.

Training combines class-reweighted KL divergence against annotator vote distributions with focal loss on smoothed boundary targets. The convo-

[1]Test3 is not used because its segment IDs do not preserve the original order.

Model	Setting	$L=2$ ($n=2{,}604$)			$L=3$ ($n=590$)			$L=4$–6 ($n=371$)			$L\geq 7$ ($n=83$)			**All** ($n=3{,}648$)		
		Acc	F1	tIoU_F	Acc	F1	tIoU_F	Acc	F1	tIoU_F	Acc	F1	tIoU_F	Acc	F1	tIoU_F
Audio-language models																
Qwen2-Audio	PT	32.5	12.8	41.6	25.8	12.9	**43.4**	21.0	10.6	**33.3**	21.2	11.0	**29.1**	28.3	12.5	39.9
	SFT	59.1	30.8	82.5	**53.1**	27.7	70.5	46.3	16.2	26.5	**49.3**	8.1	19.7	54.9	23.6	54.3
Audio-Flamingo-3	PT	**41.0**	3.0	2.2	**40.4**	3.3	2.4	**38.9**	3.5	3.3	**47.8**	6.4	3.5	**41.1**	3.3	2.4
	SFT	60.6	**34.4**	81.5	51.6	**30.4**	66.4	44.2	27.1	56.7	47.5	22.9	50.9	54.9	**32.1**	72.1
Qwen3-Omni	PT	40.8	**23.7**	**44.9**	33.6	**20.2**	40.2	35.1	**17.1**	32.3	36.7	**18.8**	27.0	38.1	**21.5**	**40.2**
	SFT	**62.4**	32.7	**85.0**	52.2	28.4	**73.2**	46.9	26.2	**66.1**	48.2	23.5	**63.0**	**56.5***	30.7	**78.0**
Kimi-Audio	PT	32.5	9.2	17.3	27.5	7.0	15.0	26.3	7.2	14.1	24.1	7.8	17.7	29.7	8.5	16.3
	SFT	61.2	33.0	82.9	51.0	28.3	68.4	**47.0**	**28.6**	60.6	48.0	**24.2**	50.4	55.6	31.0	73.7
Supervised, non-ALM baselines																
WavLM (IS 2025)	Utt.	52.5	34.5	–	50.6	30.3	–	51.6	33.2	–	51.6	29.2	–	51.9	33.5	–
WavLM-TimeAug	Seq.	48.5	28.7	71.5	44.5	25.6	64.5	40.8	24.9	57.9	42.3	23.4	52.4	45.9	27.7	66.4

Table 2: Performance in Acc and F1 measure emotion recognition and tIoU_F temporal overlap of onset and offset timings averaged across all segments. Within each setting (PT or SFT), the best ALM score is shaded dark and in bold, and the second best is shaded light. "–" means not applicable.Qwen2-Audio SFT hears only, and is therefore not grounded in, the first 30,s of each sequence; 78 of the 83 sequences with $L \geq 7$.

lutional encoder and bottom 12 transformer layers remain frozen, while the upper 12 layers and newly introduced components are trained. Full implementation, architectural illustration, and training details are provided in Appendix A.

4.4 Metrics

Figure 4 illustrates scoring on one sequence. Within each sequence, predicted and gold spans are paired one-to-one, greedily by highest temporal intersection-over-union,

$$\text{tIoU}(\hat{y}, y) = \frac{\max\big(0, \min(\hat{t}^e, t^e) - \max(\hat{t}^s, t^s)\big)}{\max(\hat{t}^e, t^e) - \min(\hat{t}^s, t^s)}, \tag{3}$$

where (t^s, t^e) and $(\hat{t}^s, \hat{t}^e)$ are the gold and predicted onsets and offsets (Eq. 1). A pair with the same label is a true positive. Unpaired gold spans are false negatives, unpaired predictions are false positives, and a mislabeled pair counts as one of each. Counts are pooled within each length bin.

Recognition. *Accuracy* is true positives over gold spans. *Macro-F1* over the eight classes also penalizes spurious predictions. Because pairing is one-to-one, segmentation errors also cost recognition: a merged span can match only one of the gold spans it covers, and each extra fragment is a false positive.

Timing. *tIoU$_F$* measures boundary quality regardless of label. It is the harmonic mean of recall-side tIoU (averaged over gold spans) and precision-side tIoU (averaged over predicted spans clipped to the audio), with unpaired spans scored as 0, so both missed and spurious spans lower it. Details are in Appendix A.1.

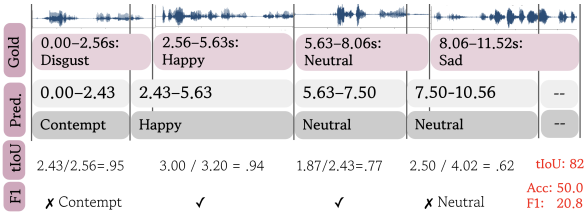


Figure 4: An illustration of how evaluation metrics are calculated on a sequence of four consecutive segments.

5 Results

5.1 Averaged performance across all single segments (n-grams with n=1)

Fine-tuning substantially improves sequential SER performance (Table 2). When averaging all performance across all the segments, Qwen3-Omni achieves the strongest overall accuracy (56.5) and tIoU (78.0) after fine-tuning, whereas Audio-Flamingo-3 achieves the highest macro-F1 (32.1). Across the four ALMs, fine-tuning significantly improves macro-F1 and temporal IoU (tIoU) relative to the pretrained models, although the magnitude of the improvement varies substantially by architecture. For instance, Audio-Flamingo-3 improves from 3.3 to 32.1 macro-F1 and from 2.4 to 72.1 tIoU, while Kimi-Audio improves from 8.5 to 31.0 macro-F1 and from 16.3 to 73.7 tIoU.

Fine-tuned ALMs outperform the supervised sequence baseline overall. All four fine-tuned

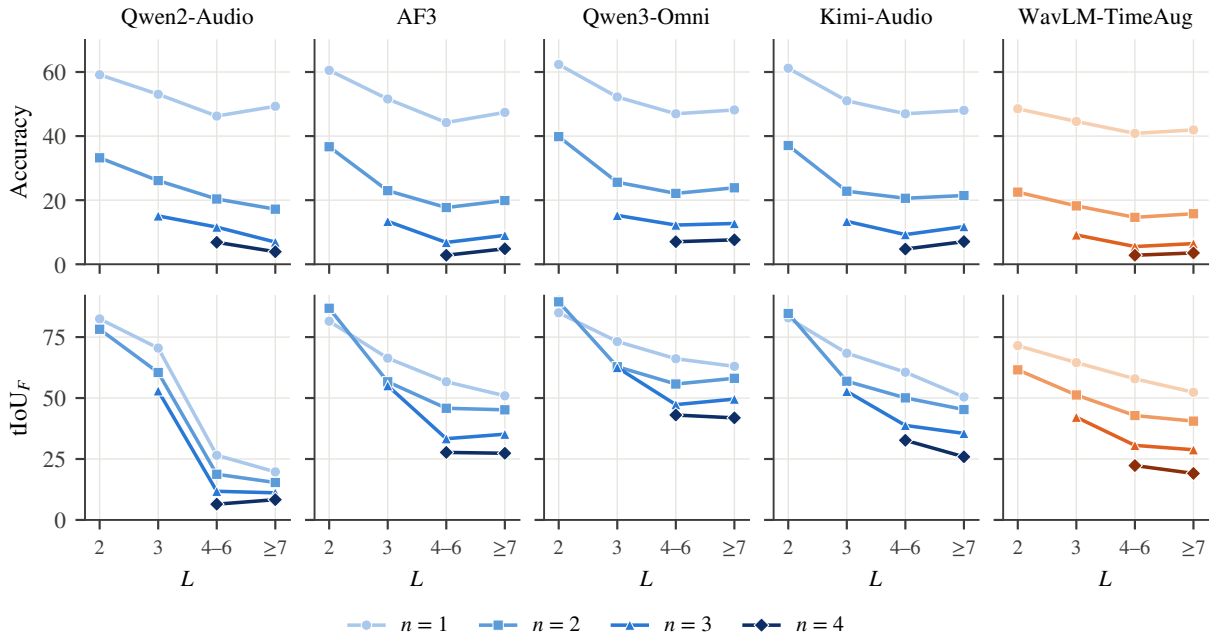


Figure 5: Accuracy and tIoU of n-gram units of discrete emotions (n consecutive spans) as sequence length L increases, for each fine-tuned ALM and WavLM-TimeAug. Windows exist only for $L \geq n$, so longer n-grams start at larger L. Please refer to Table 3 for n=1 results.

ALMs outperform WavLM-TimeAug on overall macro-F1 and tIoU (Table 3). Thus, generating emotion spans with an ALM can outperform a dedicated supervised sequence model on these metrics, but this advantage is not uniform across models or metrics. These comparisons should also be interpreted in light of the systems' different input-handling strategies and training objectives.

Temporal localization generally deteriorates as sequence length increases. For the fine-tuned systems, tIoU decreases across successive length bins. Audio-Flamingo-3 declines from 81.5 for two-segment sequences to 50.9 for sequences with at least seven segments. Qwen3-Omni declines from 85.0 to 63.0, Kimi-Audio from 82.9 to 50.4, and WavLM-TimeAug from 71.5 to 52.4. The corresponding declines are 30.6, 22.0, 32.5, and 19.1 points, respectively. This consistent pattern across generative and supervised sequence models indicates that temporal localization becomes less reliable on longer recordings under the present evaluation protocol.

However, the results do not isolate the number of boundaries as the sole cause: longer sequences may also contain more difficult emotion transitions, and the longest bin contains only 83 Test1 sequences. The pretrained models show more varied patterns. Consequently, the monotonic length-related decline is clearest among the fine-tuned systems and WavLM-TimeAug. For reference, we also report performance on discrete SER (Table 3, Appendix), which establishes a baseline for each ALM's emotion recognition ability independent of temporal effects.

5.2 Averaged performance across spans of multiple segments (n-grams with n>1)

Figure 5 extends the evaluation from single spans (n=1, Table 3) to windows of n consecutive spans, where a window counts as correct only if its n spans are paired with n consecutive predicted spans that carry the same labels. This is a much more challenging evaluation since it evaluates how accurate emotion recognition is for different consecutive emotions (e.g., "Happy→Sad", "Angry→Happy") under long conversational horizon.

Accuracy decays quickly with n. We observe that recognition accuracy falls sharply as n grows over all lengths L. If span-level successes were independent, accuracy would fall as the n-th power of the n=1 accuracy in Table 3. Although bi-gram accuracy is close to that level (e.g., 31.0 vs. 32.0 for Qwen3-Omni or closely 56.5^2 for 56.6 is the accuracy for n=1 marked by * in Table 3), but tri-gram and 4-gram accuracy fall well below it for most models. Overall, a model that labels a single span correctly more than half of the time therefore recovers only about 13% of three-span contours.

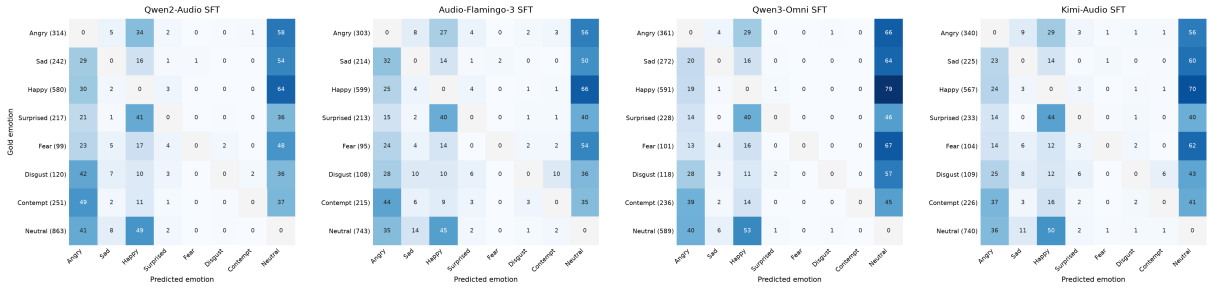


Figure 6: Label errors per model after fine-tuning: gold spans covered by a single prediction with IoU ≥ 0.5 that covers no other gold span but names the wrong emotion. Errors collapse onto Neutral, Happy and Angry (87–96% of mislabeled predictions), while Fear, Disgust and Contempt together account for under 4%.

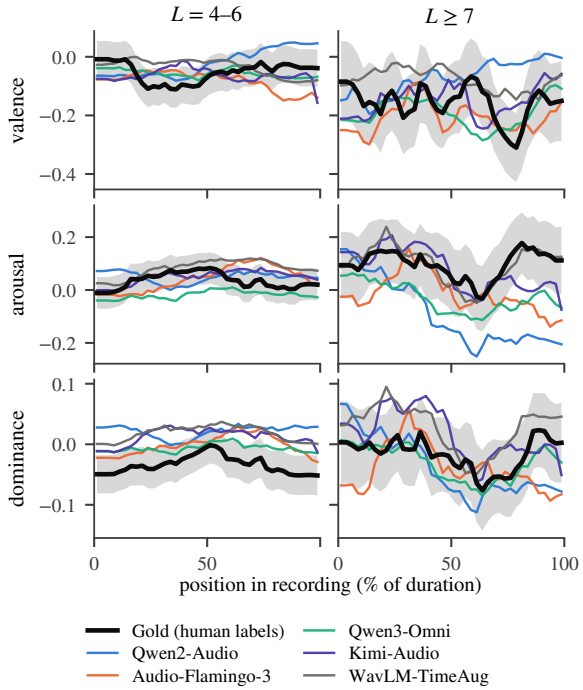


Figure 7: Average VAD contours for sequences with L=4–6 (n=371) and $L \geq 7$ (n=83). Each recording is rescaled to 0–100% of its duration before averaging. Emotion labels are mapped to words extracted from NRC-VAD dictionary (Mohammad, 2018).

This provides clear evidence for our benchmark design, which exposes failures in tracking consecutive emotional states that span-level accuracy alone cannot reveal.

Architectural advance in temporal emotion recognition and localization for $n > 1$. Consistent with n=1 (Table 3), Qwen3-Omni maintains the best model or tied for best on both metrics at every n and length where windows exist. Its tIoU_F shows to be more stable as n increases especially when compared with Qwen2-Audio, an earlier version, where we observe a significantly drop in localization capability when n=3 $\rightarrow$ 4. Contrastingly, non-ALM WavLM-TimeAug model has the lowest accuracy at every n, and its inferiority in tIoU_F grows when n=1 $\rightarrow$ 4 (orange lines, Fig. 5). These results show that span-level scores hide the main difficulty of sequential SER: keeping several consecutive spans simultaneously correct in label and in time will require architectural advance.

6 Discussion

6.1 Errors converge to frequent emotions

Figure 6 exhibits similar label confusion patterns across all four SFT models. *Neutral*, *Happy*, and *Angry* account for 87–96% of predictions on misclassified spans, while the minority gold classes, *Surprised*, *Fear*, *Disgust*, and *Contempt*, account for 25–27% of errors, with predominant absorption in the more frequent classes. This reflects an imbalance in class distribution that fine-tuning does not address.

In terms of confusion errors, *Neutral* is the primary destination absorbing 70% of *Happy* errors, 59% of *Angry* errors, 58% of *Fear* errors, and 57% of *Sad* errors across models. Qwen3-Omni exhibits the highest concentration of such label redistribution, as 79% of *Happy* and 45-66% of labels in other classes are misattribute to *Neutral*.

6.2 ALMs' correlations with VAD contours

Figure 7 compares human and predicted emotion contours after mapping each label to valence, arousal and dominance across an audio's duration.

Here we map each emotion label to the VAD of the English word that names it (Mohammad, 2018). Thus, the contours reflect how emotion words are used in language rather than acoustic labels. This is intriguing when relating to how ALMs work, as they are built on language-model backbones and report emotions as words. We hypothesize that their moderate correlation with word-based VAD

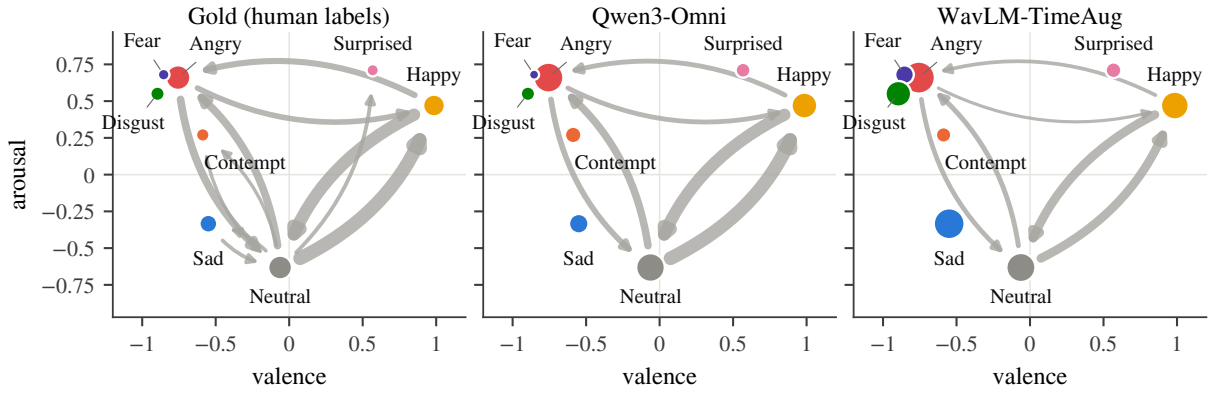


Figure 8: Emotion transitions between consecutive spans on the valence–arousal coordinates extracted from NRC-VAD dictionary (Mohammad, 2018) using English words mapping our eight emotions. Predictions from Qwen3-Omni and WavLM-TimeAug are ones having localization with tIoU ≥ 0.5. Node size indicates conditional transition probability or $P(\text{next} = e \mid \text{current} = e)$ where e is an emotion label. Arrow width indicates the share of all adjacent pairs making that change of at least 1.2%.

contours could partly reflect emotion knowledge acquired from natural-language contexts, although further work is needed for verification.

Overall, we observe that the correlation between human and predicted contours is moderate: 0.41–0.47 (valence), 0.34–0.39 (arousal) and 0.31–0.37 (dominance) for L=4–6, and 0.36–0.49, 0.24–0.40 and 0.21–0.30 for $L\geq$7 with Qwen3-Omni correlates best with the VAD contours in most settings.

6.3 Conditional emotion transition flow

Figure 8 shows that human labels and model predictions share very similar transitions but at different rates among three main clusters of emotions with most frequent transitions: (1) *Angry*, *Fear*, *Disgust*, (2) *Happy*, and (3) *Neutral*. For instance, Neutral→Happy and Happy→Neutral account for 7.9% and 7.2% of human pairs and for 8.0% and 7.1% of Qwen3-Omni pairs.

Especially, we observe that some transition flows only appear in human annotators and very rare in models' predictions. For instance, transitions involving *Sad, Surprised,* and *Contempt* appear only in the human annotations, because the ALMs make them less than 1.2% of the time (hence omitted in Fig. 8). ALMs therefore preserve the dominant emotional routes but could not capture fine-grained emotion contours like the annotators, who recognizes more nuanced emotion transitions and hence assign rare emotions briefly between neutral states.

6.4 Implications and future directions

These findings demonstrate the value of evaluating sequential SER through complementary categorical and temporal metrics. Although fine-tuning substantially improves performance, localization remains challenging as sequence length increases, and span-level accuracy alone overstates how well a system follows the course of emotion in a conversation. This matters for applications such as customer service and clinical settings, where emotional changes over time, rather than isolated labels, determine appropriate responses.

Future work should examine error types across sequence lengths, disentangle sequence duration from emotion-transition complexity, and account for differences in input coverage and boundary uncertainty. These analyses could provide a clearer understanding of the capabilities and limitations of ALMs for temporally structured emotion recognition in real-life applications.

7 Conclusion

We evaluated sequential speech emotion recognition on naturally ordered human-human conversations, comparing pretrained and fine-tuned ALMs with supervised non-ALM baselines. Our findings demonstrate the limitations of clip-level emotion recognition alone in characterizing sequential SER. Future evaluations should distinguish categorical recognition from temporal localization and account for differences in input coverage. Validating segment-based reference spans against emotion annotations collected continuously in conversational context, and extending evaluation to multi-speaker exchanges, would show how far these findings carry over to real conversation.

Limitations

Our evaluation has several limitations. First, MSP-Podcast provides categorical labels for individual speech segments rather than continuous annotations of emotional onset and offset. We therefore use original segment boundaries as reference spans. These measure recovery of segment-level labels, not the exact temporal extent of emotional states, so tIoU_F reflects alignment with the available annotation structure rather than accuracy in locating psychological emotion boundaries.

Models differ in how much audio they hear. In particular, fine-tuned Qwen2-Audio hears only the first 30 s, which affects 78 of the 83 longest Test1 sequences. The Oracle and Joint settings use separately trained adapters, so their gap is not a controlled test of context.

Future work might validate these findings on continuously annotated conversational emotion data with fine-grained boundary agreement, evaluate matched input-coverage conditions and gold-boundary scoring of Joint models, and report uncertainty in length-stratified performance.

Ethics Statement

Sequential SER labels emotions continuously over time, which makes it easier to monitor or profile speakers than clip-level SER does. Deploying such systems in workplaces, classrooms or customer-service settings raises risks of surveillance and of decisions based on unreliable inferences (Stark and Hoey, 2021; Barker et al., 2025); the EU AI Act restricts emotion recognition in workplaces and education (Durovic and Corno, 2024). Our results show that current systems are far from reliable enough for such uses. Fine-tuned models rarely predict minority emotions, never recognize Fear, and name emotions more accurately for male than for female speakers by 4 to 8 points. Categorical labels and segment-level boundaries also simplify how emotion is expressed and perceived, and annotations reflect the perceptions of a specific annotator pool. We use MSP-Podcast under its research license, report only aggregate results, and release no audio or speaker-identifying information.

Computational Resources

All ALM fine-tuning reported here ran on a single NVIDIA A100-SXM4-40GB and took about 120 GPU-hours: about 109 h for utterance-level fine-tuning of Qwen2-Audio, Audio-Flamingo-3 and Kimi-Audio, and about 11 h for sequence-level fine-tuning, including checkpoint selection. Qwen3-Omni required 18 h of fine-tuning, which ran on a H100. To limit cost, we apply LoRA to publicly released checkpoints and initialize WavLM-TimeAug from the released Interspeech 2025 model rather than training any model from scratch.

AI Disclosure

We used an AI agent to schedule, monitor and log experiment runs; it did not write model or evaluation code. We used an agent to help analyze results and to revise paper for clarity. The authors designed the study, wrote the code, ran and verified all experiments, checked every number against the experiment outputs, and take full responsibility for the paper's content.

Software and Artifacts

We use PyTorch, Hugging Face Transformers and PEFT with the publicly released checkpoints of Qwen2-Audio, Audio-Flamingo-3, Qwen3-Omni, and Kimi-Audio, and the released Interspeech 2025 WavLM model and code. The transcript-only classifier uses scikit-learn. MSP-Podcast v1.12 is used under its academic research license, and all artifacts are used in line with their licenses and intended research use. Our code will be released upon publication.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Darlene Barker, Mukesh Kumar Reddy Tippireddy, Ali Farhan, and Bilal Ahmed. 2025. Ethical considerations in emotion recognition research. *Psychology International*, 7(2):43.

Debarpan Bhattacharya, Apoorva Kulkarni, and Sriram Ganapathy. 2025. Benchmarking and Confidence Evaluation of LALMs For Temporal Reasoning. In *Interspeech 2025*, pages 2068–2072.

Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N. Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, Huang-Cheng Chou, and Pravin Mote. 2025. The MSP-Podcast Corpus. *arXiv preprint*. ArXiv:2509.09791 [eess].

Georgios Chatzichristodoulou, Despoina Kosmopoulou, Antonios Kritikos, Anastasia Poulopoulou, Efthymios Georgiou, Athanasios Katsamanis, Vassilis Katsouros, and Alexandros Potamianos. 2025. MEDUSA: A Multimodal Deep Fusion Multi-Stage Training Framework for Speech Emotion Recognition in Naturalistic Conditions. *arXiv preprint*. ArXiv:2506.09556 [cs.CL].

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

Huang-Cheng Chou, Chi-Chun Lee, and Carlos Busso. 2022. Exploiting Co-occurrence Frequency of Emotions in Perceptual Evaluations To Train A Speech Emotion Classifier. In *Interspeech 2022*, pages 161–165. ISCA.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Kerstin Denecke and Elia Gabarron. 2024. The ethical aspects of integrating sentiment and emotion analysis in chatbots for depression intervention. *Frontiers in Psychiatry*, 15:1462083.

Mateja Durovic and Tommaso Corno. 2024. The privacy of emotions: From the GDPR to the AI Act, an overview of emotional AI regulation and the protection of privacy and personal data. *Privacy, Data Protection and Data-driven Technologies*, pages 368–404.

Soumya Dutta, Smruthi Balaji, Varada R, Viveka Salinamakki, and Sriram Ganapathy. 2025. ABHINAYA – A System for Speech Emotion Recognition In Naturalistic Conditions Challenge. *arXiv preprint*. ArXiv:2505.18217 [cs.SD].

Tiantian Feng, Thanathai Lertpetchpun, Dani Byrd, and Shrikanth Narayanan. 2025. Developing a Top-tier Framework in Naturalistic Conditions Challenge for Categorized Emotion Prediction: From Speech Foundation Models and Learning Objective to Data Augmentation and Engineering Choices. *arXiv preprint*. ArXiv:2505.22133 [cs.SD].

Yajing Feng and Laurence Devillers. 2023. End-to-end continuous speech emotion recognition in real-life customer service call center conversations. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE.

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. 2025. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. 2025. EmoBench-M: Benchmarking emotional intelligence for multimodal large language models. *Preprint*, arXiv:2502.04424.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-Audio technical report. *Preprint*, arXiv:2504.18425.

Saiteja Kosgi, Sarath Sivaprasad, Niranjan Pedanekar, Anil Nelakanti, and Vineet Gandhi. 2022. Empathic machines: Using intermediate features as levers to emulate emotions in text-to-speech systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 336–347, Seattle, United States. Association for Computational Linguistics.

Wei-Cheng Lin and Carlos Busso. 2021. Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling. *IEEE Transactions on Affective Computing*, 14(2):1215–1227.

Reza Lotfian and Carlos Busso. 2019. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Wai Laam Mak, Isibor Kennedy Ihianle, and Pedro Machado. 2026. EmotionAI: A Privacy-Preserving Computational Intelligence Pipeline for Speech-Emotion-Grounded Conversational Analysis. *arXiv preprint*. ArXiv:2606.24941 [cs.SD].

Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. 2020. The MSP-Conversation Corpus. In *Interspeech 2020*, pages 1823–1827. ISCA.

Luz Martinez-Lucas, Wei-Cheng Lin, and Carlos Busso. 2024. Analyzing Continuous-Time and Sentence-Level Annotations for Speech Emotion Recognition. *IEEE Transactions on Affective Computing*, 15(3):1754–1768.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Abinay Reddy Naini, Lucas Goncalves, Ali N Salman, Pravin Mote, Ismail R Ulgen, Thomas Thebaud, Laureano Moro Velazquez, Leibny Paola Garcia, Najim Dehak, Berrak Sisman, et al. 2025. The Interspeech 2025 challenge on speech emotion recognition in naturalistic conditions. In *Proc. Interspeech 2025*, pages 4668–4672.

Zahra Omidi and John H. L. Hansen. 2026. Learning from Annotation Uncertainty: Entropy-Aware Curriculum for Speech Emotion Recognition. *arXiv preprint*. ArXiv:2606.27536 [cs.SD].

Zhiliang Peng, Jianwei Yu, Yaoyao Chang, Zilong Wang, Li Dong, Yingbo Hao, Yujie Tu, Chenyu Yang, Wenhui Wang, Songchen Xu, Yutao Sun, Hangbo Bao, Weijiang Xu, Yi Zhu, Zehua Wang, Ting Song, Yan Xia, Zewen Chi, Shaohan Huang, and 5 others. 2026. VibeVoice-ASR Technical Report. *arXiv preprint*. ArXiv:2601.18184 [cs.SD].

Arvind Krishna Sridhar, Yinyi Guo, and Erik Visser. 2024. Enhancing temporal understanding in audio question answering for large audio language models. *Preprint*, arXiv:2409.06223.

Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 782–793.

Luoyi Sun, Xiao Zhou, Zeqian Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2026. SpotSound: Enhancing Large Audio-Language Models with Fine-Grained Temporal Grounding. *arXiv preprint*. ArXiv:2604.13023 [cs.SD].

Hualei Wang, Yiming Li, Shuo Ma, Hong Liu, and Xiangdong Wang. 2025. Listening Between the Frames: Bridging Temporal Gaps in Large Audio-Language Models. *arXiv preprint*. ArXiv:2511.11039 [cs.SD].

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

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, Ziyu Guo, Zhihong Zhu, Hao Wu, Haixin Wang, Yefeng Zheng, Xiaojiang Peng, Xian Wu, Kun Wang, Xiangang Li, and 2 others. 2025. MME-Emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *Preprint*, arXiv:2508.09210.

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W. Schuller, and Haizhou Li. 2023. Emotion Intensity and its Control for Emotional Voice Conversion. *IEEE Transactions on Affective Computing*, 14(1):31–48.

A Training Details

ALM fine-tuning. We apply LoRA with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05 to the query, key, value, and output projections of the language model and to the query, key, and value projections of the audio encoder; the original encoder and language-model weights stay frozen. For sequence fine-tuning, the multimodal projector is also trained. This gives 25.96M trainable parameters out of 8.42B for Qwen2-Audio (0.31%) and 31.46M out of 8.28B for Audio-Flamingo-3 (0.38%). Utterance fine-tuning uses the same LoRA configuration without training the projector (20.71M and 14.02M trainable parameters). We use AdamW with learning rate 2×10^{-4}, linear decay without warmup, and effective batch size 8. Length-bin loss weights are 0.353 ($L = 2$), 1.460 ($L = 3$), 2.575 ($L = 4$–6), and 10.889 ($L \geq 7$). We select epoch 2 for Qwen2-Audio and epoch 3 for Audio-Flamingo-3. The two Audio-Flamingo-3 training sequences longer than 180 s are excluded due to GPU memory constraints. For pretrained Qwen2-Audio, adjacent same-label spans from overlapping windows are merged only when they meet at a shared window boundary.

At inference, local boundary-score peaks above a Development-tuned threshold define segment boundaries, with segments shorter than 2 s merged and the recording endpoints fixed as boundaries. Emotion predictions are therefore evaluated using model-predicted spans rather than gold boundaries. Training combines class-reweighted KL divergence against annotator vote distributions with focal loss on smoothed boundary targets. The convolutional encoder and bottom 12 transformer layers remain frozen, while the upper 12 layers and newly introduced components are trained. The reported checkpoint is selected using Development macro-F1 and boundary average precision.

Note: Qwen2-Audio's audio encoder accepts at most 30 s. The pretrained model therefore decodes consecutive 30 s windows and joins their spans. The fine-tuned model sees only the first 30 s of each sequence. This affects 203 of the 3,648 Test1 sequences, including 78 of the 83 with $L \geq 7$. Audio-Flamingo-3, Kimi-Audio and Qwen3-Omni process whole sequences.

The model is initialized from a copy of the released checkpoint (seed 7). It processes the same concatenated waveform as the audio-language models and predicts its own emotion spans. Training combines class-reweighted KL divergence against annotator vote distributions with focal loss on smoothed boundary targets. Audio up to 10 s is encoded in one pass. Longer audio is encoded in overlapping 10 s windows with an 8 s hop, discarding frames within 1 s of internal window edges before joining retained frames into a continuous 50 Hz sequence. The convolutional encoder and bottom 12 transformer layers are frozen; the top 12 layers and all newly introduced components are trained. The emotion branch uses the final WavLM layer. The boundary branch uses a learned weighted sum of all 25 layers, pooled to 10 Hz and processed by a two-layer bidirectional LSTM with 256 units per direction.

Boundary targets are Gaussian heatmaps centered on internal gold boundaries, with standard deviation σ linearly annealed from 0.25 to 0.15 s over 20 planned epochs. We use CenterNet's penalty-reduced focal loss ($\alpha = 2$, $\beta = 4$) and initialize the output bias using a boundary prior of 0.02. At inference, boundary probabilities are smoothed with a Gaussian of $\sigma = 0.1$ s; local peaks above the Development-tuned threshold $\theta = 0.25$ define boundaries. The decoder repeatedly removes boundaries until all segments are at least 2 s long, merging the final segment backward and other shortest segments forward. Recording endpoints are fixed boundaries. Emotion predictions are therefore evaluated on model-predicted spans rather than gold boundaries.

For each predicted segment, attentive statistics pooling produces a representation that is classified using a bidirectional LSTM over the ordered segments and an eight-class linear layer. The emotion loss is the mean KL divergence from the inverse-frequency-reweighted, renormalized annotator vote distribution q_k to the predicted distribution p_k:

$$
\begin{aligned}
q_{k,c} &= \frac{w_c v_{k,c}}{\sum_{c'=1}^{8} w_{c'} v_{k,c'}}, \\
\mathcal{L}_{\text{emo}} &= \frac{1}{N} \sum_{k=1}^{N} \sum_{c=1}^{8} q_{k,c} \log \frac{q_{k,c}}{p_{k,c}},
\end{aligned}
\tag{4}
$$

where $v_{k,c}$ is the annotator vote mass, w_c is the inverse-frequency class weight, and N is the number of segments. The total objective is

$$
\begin{aligned}
\mathcal{L} &= \mathcal{L}_{\text{emo}} + \lambda_b \mathcal{L}_{\text{boundary}}, \\
\lambda_b &= 0.40,
\end{aligned}
\tag{5}
$$

with λ_b chosen to balance the loss terms at ini-

tialization. Four sequence-length bins are sampled equally. With probability 0.5, the boundaries used for emotion pooling are jittered by up to ± 0.3 s; boundary targets remain unchanged. Of 9,495 training sequences, 22 longer than 100 s are cropped to random contiguous runs of whole segments no longer than 100 s.

We use Adam with learning rates of 10^{-5} for WavLM parameters and 10^{-4} for other trainable parameters. The reported checkpoint is selected using Development macro-F1 and boundary average precision (AP): we retain the checkpoint that maximizes their mean (epoch 3),

$$S_{\text{dev}} = 0.5\big(\text{Macro-F1} + \text{Boundary AP}\big). \tag{6}$$

Boundary AP ranks predicted peaks by confidence, matching each to at most one gold boundary within ± 0.5 s. AP is 0.759 on Development and 0.656 on Test1.

Discrete SER prompt.

> `Listen to the speech and identify the primary emotion expressed by the speaker. Reply with exactly one word, chosen from this list and nothing else: Angry, Sad, Happy, Surprised, Fear, Disgust, Contempt, Neutral.`

Sequential SER prompt.

> `Listen to the entire audio sequence and identify the emotion expressed by the speaker(s) over time.`
>
> `The sequence may contain multiple speakers and multiple emotional states. Track how emotion changes throughout the audio rather than assigning a single emotion to the entire recording. Use the audio itself to determine the segment boundaries.`
>
> `Segments comprise utterances at the phrasal level. For each distinct segment, provide:`
>
> - `the start time in seconds`
> - `the end time in seconds`
> - `the emotion`
>
> `Use only these emotion labels: Angry, Sad, Happy, Surprised, Fear, Disgust, Contempt, Neutral.`
>
> `Return the result as a chronological list of:`
>
> `(start_time, end_time, emotion)`

A.1 Scoring details

Pairing is computed per sequence and ignores labels. For recall-side tIoU, accuracy and F1, predicted spans are not clipped to the audio, so spans that run past its end lower the score. For precision-side tIoU, predictions are clipped to the audio and spans that start after it ends score 0. Pooling counts within a length bin gives every gold segment equal weight regardless of its sequence. One parser reads every system's raw output; unparseable outputs contribute no predictions. We also report two content-free references (a back-to-back 5 s Neutral grid and one Neutral span covering the audio), which show the tIoU_F available from timing alone.

A.2 Discrete Single-Label SER Baseline

We include the baseline for both pretrained and SFT model performance on MSP-Podcast (v1.12) Test1 split for reference (see Table 3).

Table 3: Performance on MSP-Podcast v1.12 Test1 ($n = 3{,}648$ sequences), stratified by sequence length L. *Utt.* rows are single-label discrete SER baselines that classify each segment independently using gold segment boundaries; *Seq.* rows predict timed emotion spans over the whole concatenated sequence. PT denotes the pretrained model and SFT denotes LoRA fine-tuning on the corresponding task. Audio context differs by row and these are not equivalent full-context evaluations. We abbreviate model names to conserve space: Qwen2-Audio, Audio-Flamingo-3, Qwen3-Omni, Kimi-Audio, and WavLM-TimeAug.

			$L = 2$ ($n = 2{,}604$)			$L = 3$ ($n = 590$)			$L = 4$–6 ($n = 371$)			$L \geq 7$ ($n = 83$)			**All** ($n = 3{,}648$)		
Model	Task	Setting	Acc	F1	tIoU_F	Acc	F1	tIoU_F	Acc	F1	tIoU_F	Acc	F1	tIoU_F	Acc	F1	tIoU_F
Audio-language models																	
Qwen2	Utt.	PT	0.326	0.204	–	0.329	0.177	–	0.310	0.150	–	0.360	0.172	–	0.327	0.189	–
		SFT	0.607	0.374	–	0.608	0.340	–	0.565	0.347	–	0.564	0.297	–	0.596	0.360	–
	Seq.	PT	0.325	0.128	0.416	0.258	0.129	0.434	0.210	0.106	0.333	0.212	0.110	0.291	0.283	0.125	0.399
		SFT	0.591	0.308	0.825	0.531	0.277	0.705	0.463	0.162	0.265	0.493	0.081	0.197	0.549	0.236	0.543
AF3	Utt.	PT	0.343	0.094	–	0.360	0.096	–	0.417	0.113	–	0.400	0.098	–	0.364	0.098	–
		SFT	0.620	0.397	–	0.605	0.382	–	0.570	0.343	–	0.591	0.306	–	0.606	0.381	–
	Seq.	PT	0.410	0.030	0.022	0.404	0.033	0.024	0.389	0.035	0.033	0.478	0.064	0.035	0.411	0.033	0.024
		SFT	0.606	0.344	0.815	0.516	0.304	0.664	0.442	0.271	0.567	0.475	0.229	0.509	0.549	0.321	0.721
Qwen3	Utt.	PT	0.521	0.353	–	0.512	0.303	–	0.526	0.323	–	0.511	0.302	–	0.519	0.339	–
		SFT	0.639	0.425	–	0.630	0.398	–	0.581	0.350	–	0.623	0.341	–	0.626	0.405	–
	Seq.	PT	0.408	0.237	0.449	0.336	0.202	0.402	0.351	0.171	0.323	0.367	0.188	0.270	0.381	0.215	0.402
		SFT	0.624	0.327	0.850	0.522	0.284	0.732	0.469	0.262	0.661	0.482	0.235	0.630	0.565	0.307	0.780
Kimi	Utt.	PT	0.558	0.334	–	0.567	0.320	–	0.549	0.313	–	0.544	0.271	–	0.557	0.326	–
		SFT	0.393	0.179	–	0.438	0.176	–	0.466	0.170	–	0.535	0.180	–	0.426	0.177	–
	Seq.	PT	0.325	0.092	0.173	0.275	0.070	0.150	0.263	0.072	0.141	0.241	0.078	0.177	0.297	0.085	0.163
		SFT	0.612	0.330	0.829	0.510	0.283	0.684	0.470	0.286	0.606	0.480	0.242	0.504	0.556	0.310	0.737
Supervised speech baselines																	
WavLM*	Utt.	–	0.525	0.345	–	0.506	0.303	–	0.516	0.332	–	0.516	0.292	–	0.519	0.335	–
*TimeAug	Seq.	–	0.485	0.287	0.715	0.445	0.256	0.645	0.408	0.249	0.579	0.423	0.234	0.524	0.459	0.277	0.664