FINDING EMOTIONS WHERE THEY BELONG: RETHINKING AUDIO EMOTION RECOGNITION THROUGH MASKED TEMPORAL AFFECTIVE GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio emotion recognition (AER) typically assigns a single label to an entire recording, leaving the temporal scope of that label ambiguous when multiple speakers and affective events are present. We address this limitation by reformulating AER as a Temporal Affective Grounding (TAG) task that associates emotions with temporally bounded speech spans and vocal tone descriptions. To support this formulation, we curate temporally annotated versions of existing emotion recognition datasets and construct recordings containing two to four affective speech spans, including overlapping speech. Training in this longer format with a standard language-modeling objective can degrade both emotion recognition and temporal grounding performance, while tone descriptions can provide shortcuts for emotion prediction. To address these challenges, we introduce Masked Temporal Affective Grounding (M-TAG), a supervised training objective that combines full-sequence language modeling with emotion and timestamp cross-entropy losses under attention masking. The masking varies the context visible to emotion-label tokens to reduce reliance on shortcuts and improve generalization, while the timestamp loss incorporates a distance-aware weight to penalize larger temporal errors. We evaluate EMO-TAG, a model fine-tuned using our dataset and objective, on emotion recognition and affective temporal-grounding against three AER and audio-language baselines: Flamingo-Next, Audio-Reasoner, and AffectGPT. Our results show that existing models achieve limited affective temporal-grounding despite competitive emotion recognition performance. Across MELD and IEMOCAP, our model improves over our strongest baseline by an average of 10 percentage points in emotion accuracy on both the original and our refined test sets and by over 20 percentage points in Emotion-Duration F1 on our affective temporal-grounding benchmark. These improvements generalize to the MME-Emotion benchmark, where EMO-TAG achieves an average gain of five percentage points relative to the best baseline. Training and dataset-creation code and model checkpoints will be released upon publication.

1 INTRODUCTION

Audio emotion recognition (AER) is commonly formulated as utterance-level classification, where a model assigns a single emotion label to an entire recording. Recent audio-language models (ALMs) extend this with natural-language affective descriptions and rationales (Zhao et al., 2025; Lian et al., 2025a), but the scope of the predicted emotion still covers the whole sample. The generated reasoning may describe broad acoustic or semantic cues without identifying where the relevant cues occur, and the reasoning often does not align with the predicted emotion (Rha et al., 2026). Therefore, a correct predicted label does not necessarily indicate that the model has correctly localized or interpreted the correct emotional cues in the input. Additionally, the utterance-level assumption is poorly matched to the content of commonly used emotion recognition benchmarks (Busso et al., 2008; Poria et al., 2019). Inspection of these benchmarks revealed that many recordings that are treated as a single sample contain more than one speaker, overlapping speech, background laughter, noise, or simply empty audio. Treating such recordings as ordinary single-label examples leads to ambiguous evaluation and makes it unclear whether a model recognizes affect from the intended speech, relies on unrelated cues, or predicts the correct emotion despite insufficient audio evidence.

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.

To address this limitation, we reformulate the AER problem as *Temporal Affective Grounding* (TAG). Given an audio recording, a model generates a structured output of speech spans, where each span is a temporally bounded region of speech associated with a transcript, tone description, and emotion label. Here, temporal grounding refers to assigning an emotion label to a speech span. The boundaries indicate where the speech span of the emotion starts and ends, rather than the precise onset and offset of emotional expression (Wang et al., 2023).

To support this formulation, we curate a temporally annotated dataset from popular public datasets (Poria et al., 2019; Busso et al., 2008; Adigwe et al., 2018) and introduce a novel training objective. We filter and temporally annotate the datasets using voice activity detection (VAD) (Tan et al., 2020) and then generate tone descriptions of the filtered samples. We create refined single-label training and testing datasets, multi-span training and testing datasets containing two to four speech spans with known temporal and affective composition, and a challenging subset composed primarily of noisy samples, such as those containing multiple speakers or overlapping speech, from the original benchmarks.

Learning this structured task introduces a challenge beyond generating the required format. Under a standard language-modeling objective, emotion labels and timestamps account for only a small fraction of the target sequence, so their contribution to the direction of the gradient can be overshadowed by the more dominant transcription and tone tokens. Furthermore, during teacher forcing (Williams & Zipser, 1989), the reference tone description precedes the emotion label and can provide a shortcut for predicting it, reducing the exposure to the audio. Timestamp prediction presents an additional problem: categorical cross-entropy (CE) does not explicitly distinguish a small temporal error from a distant prediction.

Therefore, we introduce *Masked Temporal Affective Grounding* (M-TAG), a supervised training objective that combines full-sequence language modeling with masked emotion and timestamp losses. The emotion loss provides additional supervision on the first token of each emotion label, while the timestamp loss weights the CE loss of timestamp tokens by the temporal distance between the predicted and ground-truth time boundaries. When computing these auxiliary losses, we selectively mask access to the preceding speech and tone tokens to reduce the model's reliance on textual context. The language-modeling loss continues to supervise the complete structured output using the full causal context. Our training follows a curriculum that progresses from single-span format learning to masked grounding and ultimately to multi-span prediction.

We evaluate emotion recognition and affective temporal grounding against audio-language and emotion-recognition baselines. On the constructed benchmarks derived from MELD and IEMOCAP, our fine-tuned model, EMO-TAG, outperforms the strongest compared baseline on Emotion-Duration F1, a temporally-aware emotion recognition metric, by 23.12 and 35.17 percentage points, respectively. It further improves emotion recognition on the original and refined single-span benchmarks, while also achieving the best emotion-recognition accuracy on MME-Emotion benchmark (Zhang et al., 2026). In summary, our main contributions are as follows:

- We formulate AER as TAG, a structured prediction task that associates speech transcription, temporal boundaries, vocal tone, and emotion within individual speech spans.
- We introduce M-TAG, a supervised objective combining language modeling with masked emotion and distance-weighted timestamp supervision for temporal affective grounding.
- We construct temporally annotated single-span and multi-span data from existing AER datasets covering clean recordings, multiple affective events, and noisy subsets.

2 Related Work

Audio-language emotion recognition: General-purpose ALMs, such as Audio-Reasoner (Xie et al., 2025), SALMONN (Tang et al., 2024) and Qwen2-Audio-Instruct (Chu et al., 2024), support emotion recognition alongside speech transcription, audio captioning, and other audio-understanding tasks. More specialized models extend emotion recognition beyond categorical prediction. SECap (Xu et al., 2024) generates natural-language descriptions of emotions expressed in speech, while AffectGPT (Lian et al., 2025a) uses descriptive annotations to train multimodal models for both closed- and open-vocabulary emotion understanding. Other work integrates the acoustic

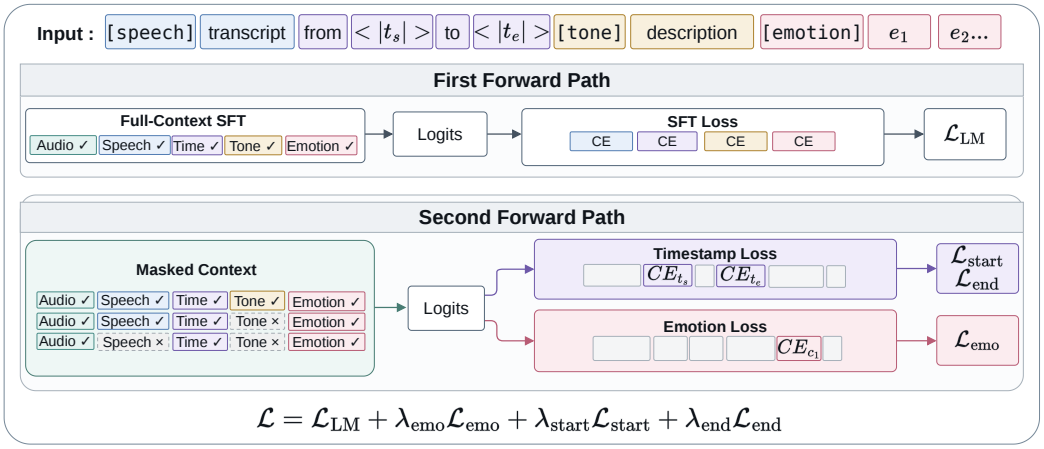


Figure 1: Overview of our objective. The first forward path calculates the language modeling loss using the full sequence. The second forward pass calculate our auxiliary losses using varying context through attention masking. All logits are discarded in the second forward path except for the ones of timestamp tokens and the first token of the emotion label.

cues as an input to the LLM. SpeechCueLLM (Wu et al., 2025) converts utterance-level vocal characteristics into textual descriptions, which are passed with the input prompt, whereas VowelPrompt (Wang et al., 2026) focuses on the vowel letters and describes pitch and energy associated in the vowel segments, and use them to prompt the LLM. These methods aim to provide richer input descriptions and cues to the language model in order to improve the emotion recognition performance, while our model aims to train the LLM to be able to generate these additional cues.

Emotion-aware training: Recent work considered extending token-level supervision to allow better emotion prediction and understanding. AffectGPT-R1 (Lian et al., 2025b) applies GRPO with emotion-wheel-based rewards (Plutchik, 1980), directly aligning training with open-vocabulary emotion evaluation. VowelPrompt Wang et al. (2026) similarly combines SFT with RL to improve structured prosody-based reasoning, while Emotional Rationale Verifier (Rha et al., 2026) further targets consistency between generated explanations and emotion predictions through an explanation-based reward. While these models use RL in order to improve performance, our approach focuses on leveraging the full capabilities of supervised learning with a carefully designed loss function.

Temporal audio and emotion grounding: Audio Temporal grounding have been rapidly improving over the year. Audio Flamingo Next (Ghosh et al., 2026) grounds intermediate reasoning steps in timestamps, while Timestamped Audio Captioning (Kumar et al., 2026) generates temporally localized descriptions of events in complex audio scenes. The closest work in emotion temporal grounding is Speech Emotion Diarization, introduced together with the Zaion Emotion Dataset (ZED) by Wang et al. (2023). ZED contains utterances with manually annotated precise emotion frames and aims to identify which emotion appears at each point in time. Our formulation differs in what the predicted boundaries represent. Creating a dataset large enough for ALLM fine-tuning with this level of annotation would require substantial human effort. We therefore adopt a simpler formulation that remains relevant to real-world settings. Rather than estimating the precise onset and offset of emotional expression within an utterance, our model aligns an emotion label with a temporally bounded speech span to which it represents.

3 Dataset Creation

Our dataset build on existing public datasets, namely IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and Emov-DB (Adigwe et al., 2018). We filter these datasets from empty audio and ambiguous speech. We use a pipeline to check the audio, refines temporal boundaries, and reject inconsistent samples. Figure 4 in Appendix D shows a complete overview of our pipeline.

3.1 AUDIO PREPROCESSING

We use the voice activity detection (VAD) tool rVAD (Tan et al., 2020) to identify empty audio samples and segment valid ones into speech and non-speech regions. We then address dataset-specific issues. The first major problem was in MELD. As the dataset is derived from the sitcom *Friends*, many samples contain background laughter tracks. This can give models misleading cues and affect the interpretation of these samples. We inspect the non-speech regions identified by VAD and remove only those with high energy, aiming to remove laughter tracks while preserving natural pauses and sighs. The second prominent issue was the presence of multiple speakers and overlapping speech in the same recording in both IEMOCAP and MELD. To address this, we transcribe each detected speech region with WhisperX (Bain et al., 2023) and use forced alignment with the ground-truth transcript. We then check whether the words in each segment are part of the original utterance. We consider a segment unrelated if it contains no words matching the ground-truth transcript. We then concatenate the remaining speech segments and run WhisperX with forced alignment again to trim any remaining irrelevant segments. If, after this final step, the transcript contains two or more consecutive insertions (extra words that are not part of the utterance) or three or more replacements (different words that may have been misrecognized by WhisperX), we reject the sample. Further details can be found in Appendix A.2

3.2 TONE ANNOTATION

We use FlamingoNext (Ghosh et al., 2026) to transcribe each recording before describing the speaker's vocal tone. We retain samples only when the generated transcription has at least 60% text similarity to the reference transcript. Qwen3 extracts the speech and tone descriptions from FlamingoNext's full output, converts them into our target format, and checks consistency between the tone and ground-truth emotion. Flagged samples are re-annotated by FlamingoNext with the ground-truth emotion supplied in the prompt, samples that remain inconsistent are discarded.
The initial tone descriptions vary in length and structure, which degraded performance in our experiments. Inspired by Wu et al. (2025); Wang et al. (2026), we refine and normalize these descriptions using low-level acoustic features. We divide each clip into five segments and extract features using openSMILE (Eyben et al., 2010). We also compute speaker-specific reference statistics for features such as pitch, intensity, and speaking pace. ChatGPT receives the initial tone description, segment-level features, and speaker references to produce a concise, normalized tone description. Tone examples are provided in Appendix D.2 and prompts are in Appendix G.3.

3.3 MULTI-SPAN CONSTRUCTION

We construct two-, three-, and four-span examples by concatenating cleaned single-span recordings without replacement. For each dataset, the gap between consecutive spans is sampled uniformly between 0.5 and 2.0 seconds from silence regions extracted from recordings in the same dataset. This avoids inserting digital zeros and preserves dataset-specific background characteristics. We similarly construct overlapping samples by uniformly sampling an overlap between 0.5 and 2.0 seconds between segments. We select these segments from a pool of audio clips with a minimum duration of three seconds. More details can be found in Appendix A.3.

4 METHOD

4.1 TASK FORMULATION

Given an audio recording x, TAG predicts a sequence of speech spans $\mathcal{S} = (S_1, \ldots, S_K)$, where $S_k = (\tau_k^{\text{start}}, \tau_k^{\text{end}}, u_k, c_k, z_k)$. Here, τ_k^{start} and τ_k^{end} denote the temporal boundaries of span k, u_k is its transcription, c_k is a natural-language description of the vocal tone, and z_k is the emotion label. The boundaries identify the speech span to which the emotion label applies. Our training samples contain $K \in \{1, 2, 3, 4\}$ spans. Each span is formatted as follows:

```
<speakerX>:  [speech] "transcript" from <|start|>s to
<|end|>s, [tone] tone description, [emotion] emotion.
```

Let $y = (y_1, \ldots, y_T)$ denote the complete serialized target sequence. The model generates this sequence autoregressively as $p_\theta(y \mid x) = \prod_{t=1}^{T} p_\theta(y_t \mid x, y_{<t})$.
Following Kumar et al. (2026), we represent temporal boundaries using learnable timestamp tokens added to the model vocabulary. We quantize timestamps with a step size of $\Delta = 0.25$ seconds, covering recordings up to $\tau_{\max} = 30$ seconds. A continuous timestamp τ is mapped to $Q(\tau) = \min\{\tau_{\max}, \Delta \text{ round}(\tau/\Delta)\}$.

4.2 Masked Temporal Affective Grounding

Full-context Supervision: We compute the language-modeling loss over the complete target sequence: $\mathcal{L}_{\text{LM}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_t \mid x, y_{<t})$.
As in vanilla language modeling, only target tokens contribute to the loss, while prompt, audio, and padding tokens are ignored. This objective provides the supervision for transcription, timestamps, tone descriptions, and emotion labels in the target output format.

Masked Emotion Supervision. The auxiliary emotion loss places additional emphasis on the emotion label of each span. There are two challenges when implementing this loss. First, some emotion labels are represented by multiple tokens, so once the first token is predicted, the subsequent tokens often have very small CE losses. Averaging the CE loss over the complete label can dilute the supervision signal for multi-token emotions compared with single-token ones. Since the first token of each emotion label in our vocabulary is unique, we compute the auxiliary emotion loss using only the first token of each emotion label.

The second challenge is that the tone description immediately precedes the emotion label. This introduces a shortcut where the model can directly predict the emotion label by focusing only on the tone description. This issue becomes more prominent in our setting when we introduce multiple spans. Inspecting attention weights during training revealed that attention from emotion label positions to audio tokens weakens over time, especially as the distance between these positions and the audio tokens increases. To reduce this dependency, we vary the context available when predicting the emotion label. With probability p_{mask}, we sample one of two strategies: *tone masking*, which blocks attention to tone tokens, or *tone-and-speech masking*, which blocks attention to both speech and tone tokens. This encourages greater attention to more distant tokens, including the audio tokens. Let q_k^{emo} denote the position of the first emotion-label token in span k, and let M denote the sampled attention mask. The per-span emotion loss is $\ell_k^{\text{emo}} = -\log p_\theta\big(y_{q_k^{\text{emo}}} \mid x, y_{<q_k^{\text{emo}}}; M\big)$.

Context masking is applied in a second forward pass through the attention mask, similar to causal attention (Vaswani et al., 2017). The selected tokens remain in the input sequence, but attention from the emotion label to these tokens is blocked. As shown in Figure 1, the full-context forward pass used for $\mathcal{L}_{\text{LM}}$ is unaffected and provides the full-sequence supervision. This formulation does not only provide a regularization effect but also allow datasets with only emotion annotations to be included in training. This allows using emotion recognition datasets without requiring additional speech or tone annotations. We demonstrate this use case in Appendix C.

Masked Temporal Supervision: For our temporal loss, we supervise the start and end timestamp tokens of each span. Timestamp prediction should depend only on the audio recording and previous timestamp tokens. However, masking the tone and speech in every training step would frequently cost an expensive extra forward pass. To avoid this, we use the same second forward pass and the attention mask M of the emotion loss. We additionally weight timestamp tokens CE relative to the distance between the predicted and ground-truth boundaries. Let $\hat{\tau}_k^b$ and τ_k^b denote the predicted and ground-truth timestamps for boundary b of span k. We define the absolute temporal error d_k^b and its distance-dependent weight as

$$d_k^b = \left|\hat{\tau}_k^b - \tau_k^b\right|, \qquad w_k^b = \text{clip}\left(\frac{d_k^b}{\delta}, c_{\min}, c_{\max}\right). \tag{1}$$

Here, δ is the distance pivot and $c_{\min}$ and $c_{\max}$ bound the coefficient. Let q_k^b denote the target-token position of the boundary b. The corresponding per-span temporal loss is

$$\ell_k^b = -w_k^b \log p_\theta\Big(y_{q_k^b} \mid x, y_{<q_k^b}; M\Big), \qquad b \in \{\text{start}, \text{end}\}. \tag{2}$$

Full objective: For a sample containing K spans, our auxiliary losses are aggregated as:

$$\mathcal{L}_{\text{emo}} = \sum_{k=1}^{K} \ell_k^{\text{emo}}, \qquad \mathcal{L}_{\text{start}} = \sum_{k=1}^{K} \ell_k^{\text{start}}, \qquad \mathcal{L}_{\text{end}} = \sum_{k=1}^{K} \ell_k^{\text{end}}. \tag{3}$$

and the final training objective becomes:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_{\text{emo}}\mathcal{L}_{\text{emo}} + \lambda_{\text{start}}\mathcal{L}_{\text{start}} + \lambda_{\text{end}}\mathcal{L}_{\text{end}}. \tag{4}$$

The λ coefficients control the contributions of the full-sequence language modeling loss, emotion recognition loss, and the two temporal boundaries loss.

4.3 Training Curriculum

Stage I: Format learning. We first train on single-span samples using only $\mathcal{L}_{\text{LM}}$. This stage teaches the model to generate our structured sequence of speech, timestamps, tone, and emotion.
Stage II: Masked grounding. Once the model generation follows our format, we introduce the masked emotion and temporal losses as in Equation (4) with $k = 1$. This stage emphasizes emotion and boundary accuracy while varying the context available to the auxiliary predictions. We randomly add noise or silence in the range of 0.5 to 1 seconds to avoid the model simply predicting the start and the end of the audio clip.
Stage III: Multi-span temporal learning. Finally, we introduce recordings containing two, three, and four speech spans. The model learns to distinguish multiple affective events and associate each predicted emotion with its corresponding temporal region. The auxiliary losses are accumulated over the annotated spans as defined in Equation (3).

5 Experiments

Baselines: We compare against FlamingoNext (Ghosh et al., 2026), Audio-Reasoner (Xie et al., 2025), and AffectGPT (Lian et al., 2025a). Both FlamingoNex and Audio-Reasoner can generate timestamped audio descriptions, while AffectGPT is specifically trained for emotion recognition. Since neither of the temporal models natively follows our structured output format, we retain each model's default prompting style and explicitly ask it to predict timestamps and an emotion from our predefined label set. We then pass the generated reasoning and the final response to Qwen3 (Yang et al., 2025), which extracts the predicted emotion and temporal boundaries into our format. To avoid penalizing differences in generation style, as shown in Appendix G.2.1, the extraction prompt is deliberately permissive and extract any mention of the emotion and their timestamp if predicted. Since AffectGPT is not trained to predict temporal spans, we evaluate it only on emotion recognition using the evaluation script and the final training checkpoint provided in its original codebase.
Evaluation Datasets: We use 8 main emotions in our training and evaluation across all datasets, mainly neutral, happiness, sadness, anger, surprise, fear, disgust, and frustration. We use *Raw* when we point to the original benchmark, *Refined* when we point to our cleaned version of the benchmark, *Noisy* when we point to the noisy samples of the original benchmark, and *multi-span* when we point to our curated multiple spans dataset. More details on the datasets composition is in Appendix D.

5.1 Evaluation Metrics

Our task requires evaluating *what* emotion is predicted and *when* it is predicted. Therefore, we report three evaluation metrics: emotion segment accuracy, emotion hit accuracy, and Emotion-Duration F1. Emotion segment accuracy provides the conventional utterance-level comparison, emotion hit accuracy checks if the model predicted the correct emotion in any predicted span, and Emotion-Duration F1 evaluates whether the correct emotion is assigned to the correct temporal region.
Emotion Segment Accuracy: Dedicatedly for single-span samples, we report standard emotion classification accuracy. Let z_i denote the ground-truth emotion label of sample i, and let $\hat{z}_i$ denote the first valid emotion label generated by the model. Emotion segment accuracy is computed as: $\text{Acc}_{\text{emo}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{z}_i = z_i]$, where N is the total number of evaluated samples and $\mathbf{1}[\cdot]$ equals one when the predicted emotion matches the ground-truth emotion and zero otherwise.
Emotion Hit Accuracy: For multi-span and single-span recordings containing overlapped or multiple speakers, a temporally-aware model would generate more than one emotion prediction. Evaluating only the first prediction can therefore penalize the model even when the ground-truth emotion is

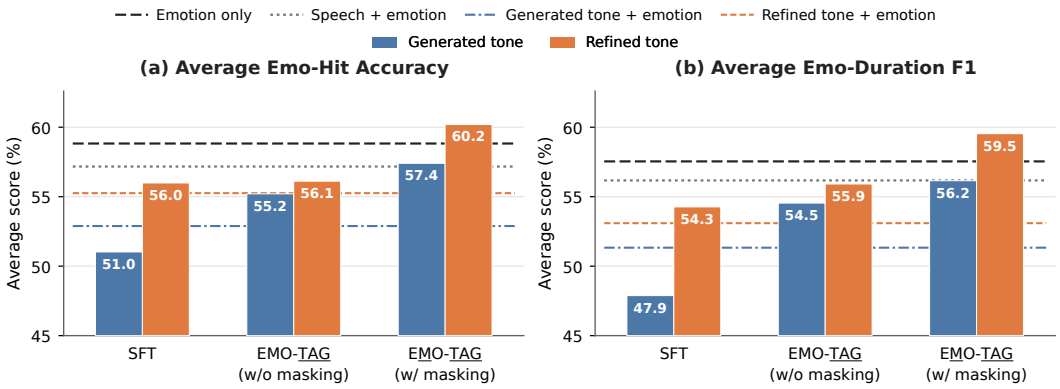


Figure 2: Average performance across context settings. Bars compare generated tone (before low-level features) and refined tone (after low-level features) with full-context, while dashed lines show the partial-context SFT settings. All runs share the same random seed and hyperparameters.

correctly identified in the appropriate temporal region. We define an emotion hit for a ground-truth span when its emotion label appears among the predicted emotion labels associated with that span. Emotion hit accuracy is formulated as:

$$\text{Hit}_{\text{emo}} = \frac{1}{\sum_{i=1}^{N} K_i} \sum_{i=1}^{N} \sum_{k=1}^{K_i} \mathbf{1}\left[z_{i,k} \in \widehat{\mathcal{Z}}_{i,k}\right], \tag{5}$$

where K_i is the number of ground-truth spans in sample i, $z_{i,k}$ is the emotion of the k-th span, and $\widehat{\mathcal{Z}}_{i,k}$ denotes the set of predicted emotions associated with that span. For single-span samples, $K_i = 1$, the metric checks whether the ground-truth emotion appears anywhere among the model's predicted emotion spans. For multi-span samples, the comparison is performed independently for each ground-truth span, preventing an emotion predicted for one span from being credited to another.

Emotion-Duration F1: To evaluate temporal grounding, i.e., whether an emotion is assigned to the correct temporal speech region, we divide each recording into non-overlapping 0.25-second frames, matching the resolution of our timestamp vocabulary. For each emotion class c, let $\mathcal{G}_c$ and $\mathcal{P}_c$ denote the sets of ground-truth and predicted frames assigned to that emotion, respectively. We compute duration precision and recall as

$$P_{\text{dur}} = \frac{\sum_c |\mathcal{G}_c \cap \mathcal{P}_c|}{\sum_c |\mathcal{P}_c|}, \quad R_{\text{dur}} = \frac{\sum_c |\mathcal{G}_c \cap \mathcal{P}_c|}{\sum_c |\mathcal{G}_c|}. \tag{6}$$

The intersection $\mathcal{G}_c \cap \mathcal{P}_c$ represents frames which both the predicted emotion label and its temporal location match the ground truth. Duration precision measures the proportion of predicted emotion frames that follows the ground-truth span, while duration recall measures the proportion of ground-truth emotion frames that are correctly recovered. Predictions with the correct temporal location but an incorrect emotion label do not contribute to the intersection and are therefore penalized. For example, if anger is annotated from 2.0 to 4.0 seconds but the model predicts anger from 2.5 to 3.5 seconds, the overlapping interval is correctly recovered, while the missed portions from 2.0–2.5 and 3.5–4.0 seconds reduce recall. Predicting anger beyond the annotated interval would reduce precision. The reported Emo-Duration F1 is calculated as the harmonic mean of P_{dur} and R_{dur}.

5.2 Results

Emotion Recognition: Table 1 reports single-span emotion recognition performance. On IEMOCAP, our model achieves the highest emotion segment accuracy across the Original, Refined, and Noisy sets, with mean accuracies of 65.62%, 65.32%, and 63.51%, respectively. This exceeds AffectGPT by at least 18.22 percentage points, despite AffectGPT receiving both the utterance transcript and audio as input. Our model also achieves the highest results on MELD Original and Refined, with mean segment accuracies of 54.65% and 57.91%, compared with 48.91% and 44.52%

Table 1: Emotion recognition across original, refined, and noisy MELD and IEMOCAP evaluation sets. Scores are reported as Emotion Segment Accuracy / Emo-Hit Accuracy (%). Results for EMO-TAG are averaged over three runs, with sample standard deviations shown beneath the means.

	IEMOCAP			MELD		
Model	**Original**	**Refined**	**Noisy**	**Original**	**Refined**	**Noisy**
Audio-Reasoner	40.76 / 44.65	38.71 / 42.24	35.26 / 36.41	40.15 / 44.14	40.36 / 43.02	39.71 / 49.37
AffectGPT	44.36 / 44.36	47.10 / 47.10	40.98 / 40.98	37.33 / 37.33	40.44 / 40.44	35.82 / 35.82
FlamingoNext	37.58 / 49.22	36.49 / 43.55	39.37 / 47.45	44.52 / 53.45	48.91 / 54.91	33.82 / 45.21
Qwen2-Audio-Instruct	43.52 / 43.52	41.94 / 41.94	43.66 / 43.66	38.88 / 39.08	36.68 / 36.68	30.15 / 30.15
EMO-TAG (ours)	**65.62 / 66.39** ± 1.53 / ± 1.52	**65.32 / 65.78** ± 0.99 / ± 1.06	**63.51 / 63.67** ± 1.70 / ± 1.56	**54.65 / 55.73** ± 0.37 / ± 0.33	**57.91 / 58.12** ± 0.98 / ± 0.91	**44.12 / 51.47** ± 2.16 / ± 1.08

for the strongest baseline, FlamingoNext. On MELD Noisy, it obtains an average segment accuracy of 44.12% and an Emo-Hit Accuracy of 51.47%, compared with 39.71% and 49.37% for Audio-Reasoner. The gap between emotion segment and emotion hit accuracy on the noisy subset of MELD highlights the noisy nature of this subset. We further evaluate emotion recognition on MME-Emotion (Zhang et al., 2026). Table 2 compares our model against the baseline results reported in this benchmark. EMO-TAG achieves the highest mean emotion segment accuracy across all three emotion recognition subsets, reaching 58.20% on Noise, 33.56% on Wild, and 45.27% on Lab. It outperforms FlamingoNext, the strongest baseline across these subsets, by 7.43, 1.06, and 7.64 percentage points, respectively. EMO-TAG also outperforms Audio-Reasoner on all three subsets, despite Audio-Reasoner receiving both the audio and the utterance transcript as input, whereas EMO-TAG uses only the audio.

Emotion Grounding: Table 2 evaluates whether models can pair each ground-truth emotion with the appropriate temporal speech span in recordings containing multiple affective events. On MELD Multi-Span, EMO-TAG reaches an Emo-Duration F1 of 54.55%, exceeding Audio-Reasoner, the strongest baseline at 31.43%, by 23.12 percentage points. The margin increases on IEMOCAP Multi-Span, where EMO-TAG achieves 63.72%, compared with 28.55% for Audio-Reasoner, a gain of 35.17 percentage points. EMO-TAG also obtains the highest Emo-Hit Accuracy on both datasets, reaching 57.09% on MELD and 61.22% on IEMOCAP, compared with the strongest baseline results of 46.03% for Audio-Reasoner and 42.53% for FlamingoNext, respectively.

Effect of Context: Figure 2 shows the effect of varying the context used during SFT training and the impact of applying our masking objective. Surprisingly, training with the emotion label alone is highly effective for emotion recognition and temporal grounding. Emotion-only training achieves Emo-Hit accuracy and Emo-Duration F1 of 58.8% and 57.5% averaged across MELD and IEMOCAP, respectively. Adding just speech leads to an average degradation in emotion recognition and temporal grounding of 1.7 and 1.4 percentage points. The original generated tone has a more pronounced negative effect, using it with the emotion label sharply degrades performance by 5.9 and 6.2 percentage points. However, refining the tone with the low-level features recovers 2.4 and 1.8 points of this degradation. Using SFT training with the full context leads to the worst performance when using the original generated tone. This performance significantly improves by 6.0 and 6.4 points when using the refined tone, achieving slightly higher performance than using refined tone alone. Introducing our auxiliary loses notably improves the original tone SFT performance, with a modest improvement when switching to the refined tone. Nonetheless, the performance remains inferior to training with the emotion label only. With our full M-TAG objective, the performance gain is substantial across the board. Emo-TAG achieves an average improvement in Emo-Hit accuracy and Emo-Duration F1 over SFT training of 7.4 and 8.3 points with the original tone, and 4.2 and 5.2 points with the refined tone. Additionally, it achieves gains of 1.4 and 2.0 points over training with the emotion label only, making it the only setting that outperforms emotion-only training. Detailed ablation and qualitative results are provided in Appendix B and Appendix E.

6 Discussion

The multi-span results in Table 2 demonstrate the major gap with existing models between recognizing which emotions are present in a recording and assigning those emotions to the correct speech span. Existing audio-language models frequently recover a relevant emotion while achiev-

Table 2: Comparison on the MELD and IEMOCAP multi-span evaluation sets and the MME-Emotion benchmarks. Avg. denotes the unweighted average across Noise, Wild, and Lab. Results marked with ‡ are taken from Zhang et al. (2026). EMO-TAG Results are averaged over three runs.

Model	MME-Emotion				MELD Multi-Span		IEMOCAP Multi-Span	
	Emo-Segment Accuracy (%)				Emo-Duration F1 (%)	Emo-Hit Acc. (%)	Emo-Duration F1 (%)	Emo-Hit Acc. (%)
	Noise	Wild	Lab	Avg.				
Audio-Reasoner	47.60‡	28.80‡	32.20‡	36.20	31.43	46.03	28.55	34.68
AffectGPT	16.80‡	8.30‡	18.00‡	14.37	–	–	–	–
FlamingoNext	50.77	32.50	37.63	40.30	28.33	42.53	24.33	42.53
Qwen2-Audio-Instruct	38.80‡	22.30‡	20.40‡	27.17	20.00	15.00	22.00	12.30
EMO-TAG (ours)	**58.20** ±1.95	**33.56** ±0.58	**45.27** ±2.69	**45.68** ±1.74	**54.55** ±1.38	**57.09** ±1.19	**63.72** ±2.42	**61.22** ±1.95

ing a substantially lower Emotion-Duration F1, indicating that these models did not attend to the correct speech span. EMO-TAG improves this grounding gap while achieving superior emotion-recognition performance. Our model consistently outperforms the compared models across all evaluated datasets, including the MME-Emotion subsets, despite using only audio, while some baselines additionally receive the utterance transcript. Furthermore, as shown in Table 7 in Appendix D, we use only 12k unique training samples to achieve these results. This validates our TAG task formulation and the possibility of implementing it in a way that does not harm recognition performance.

The context ablation in Figure 2 shows the counter-intuitive observation that providing more context does not necessarily produce a stronger model. Emotion supervision alone achieved the best performance in our ablation, only second to the full objective, while the full-context SFT model gave the worst overall performance. The additional speech and tone description, although more informative to humans, seems to make learning this task harder. Additionally, as shown in Appendix D.2, the initial tone descriptions provided more details about how each utterance was spoken, but varied considerably from one sample to another. Under the language-modeling objective, this inconsistency may have made the target descriptions more complicated to learn, contributing to lower performance. In contrast, the refined tone descriptions were more consistent and concrete, yielding much better performance. We therefore hypothesize a trade-off between the level of detail in the intermediate reasoning and the amount of available training data, more detailed reasoning may require more data to be learned reliably and may otherwise degrade performance.

One more takeaway from our experiments is the untapped potential of supervised learning. We argue that GRPO (Shao et al., 2024) should not be treated as the go-to approach when the goal is to improve a particular metric within a structured output. In our preliminary experiments, applying GRPO after SFT without our objective had modest performance improvements. Unlike hard reasoning tasks such as math or coding, our task requires concise generation, as there are not many ways to transcribe speech and describe a tone. RL finetuning encourages increasing the generation length (Yu et al., 2025; Liu et al., 2025; Mohamed & Kementchedjhieva, 2025), and leads to unnecessary continuation in our case, resulting in less attention to the input audio. More details on our GRPO formulation and limitations are in Appendix C.2 and Appendix F.

7 Conclusion

In this paper, we demonstrated the feasibility of formulating audio emotion recognition (AER) as a temporal affective grounding task. We achieved this by creating temporally annotated datasets containing a variety of span settings and introducing a masked training objective that emphasizes emotion recognition and temporal grounding performance without hindering the language-modeling objective. Our results showed the reliability of our approach by improving both temporal grounding and emotion recognition against widely used baselines. Additionally, we identified and addressed limitations in the current AER training and evaluation datasets. In general, our findings show that explicitly modeling emotion recognition as a temporal task using our objective provides a more practical formulation without harming recognition performance.

REPRODUCIBILITY STATEMENT

To promote transparency and reproducibility, we plan to release our training and dataset-creation code, datasets (if copyright allows), model checkpoints, and Weights & Biases (W&B) artifacts containing full training details and evaluation results throughout the training process. By making these resources publicly available, we hope to encourage other researchers to do the same, helping make research more transparent and reproducible and allowing the field to move forward.

AI USE STATEMENT

AI tools, specifically ChatGPT, were used solely to assist with coding and manuscript refinement. For coding, ChatGPT was used to help implement functions according to clearly specified requirements and intended functionality. Generative AI was also used during data annotation, as described in Section 3 and Appendix A.3. FlamingoNext generated the initial speech and tone descriptions, which were then extracted and structured with Qwen3. ChatGPT was used to refine the tone descriptions using the initial descriptions and acoustic features. Qwen3 was also used to convert baseline model outputs into a structured format for evaluation. For manuscript preparation, ChatGPT was used to correct grammatical errors, improve wording, assist with table editing, and create early figure drafts. AI was not used to identify the research gap, formulate the proposed methodology, develop the underlying research ideas, interpret scientific contributions, or generate any novel aspects of the work. All research concepts, methodological decisions, and scientific contributions were developed by the authors.

REFERENCES

Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*, 2018. URL https://arxiv.org/abs/1806.09514.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023. URL https://arxiv.org/abs/2303.00747.

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.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. URL https://arxiv.org/abs/2407.10759.

Florian Eyben, Martin Wöllmer, and Björn Schuller. openSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462, 2010.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio Flamingo Next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026. URL https://arxiv.org/abs/2604.10905.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Sonal Kumar, Prem Seetharaman, Ke Chen, Oriol Nieto, Jiaqi Su, Zhepei Wang, Rithesh Kumar, Dinesh Manocha, Nicholas J. Bryan, Zeyu Jin, and Justin Salamon. TAC: Timestamped audio captioning. *arXiv preprint arXiv:2602.15766*, 2026. URL https://arxiv.org/abs/2602.15766.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36993–37014. PMLR, 2025a. URL https://proceedings.mlr.press/v267/lian25a.html.

Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, Rui Liu, Haoyu Chen, Xiaobai Li, and Bin He. AffectGPT-R1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2508.01318*, 2025b. URL https://arxiv.org/abs/2508.01318.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. URL https://arxiv.org/abs/2503.20783.

Abdelrahman Mohamed and Yova Kementchedjhieva. The devil is in the EOS: Sequence training for detailed image captioning. *arXiv preprint arXiv:2507.20077*, 2025. URL https://arxiv.org/abs/2507.20077.

Robert Plutchik. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman (eds.), *Emotion: Theory, Research, and Experience, Volume 1: Theories of Emotion*, pp. 3–33. Academic Press, 1980. doi: 10.1016/B978-0-12-558701-3.50007-7.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Hyeongseop Rha, Jeong Hun Yeo, Yeonju Kim, and Yong Man Ro. Emotion-coherent reasoning for multimodal LLMs via emotional rationale verifier. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2029–2037, 2026. URL https://arxiv.org/abs/2510.23506.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Zheng-Hua Tan, Achintya kr. Sarkar, and Najim Dehak. rVAD: An unsupervised segment-based robust voice activity detection method. *Computer Speech & Language*, 59:1–21, 2020. doi: 10.1016/j.csl.2019.06.005. URL https://github.com/zhenghuatan/rVAD.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.13289.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yancheng Wang, Osama Hanna, Ruiming Xie, Xianfeng Rui, Maohao Shen, Xuedong Zhang, Christian Fuegen, Jilong Wu, Debjyoti Paul, Arthur Guo, Zhihong Lei, Ozlem Kalinli, Qing He, and Yingzhen Yang. VowelPrompt: Hearing speech emotions from text via vowel-level prosodic augmentation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.06270.

Yingzhi Wang, Mirco Ravanelli, and Alya Yacoubi. Speech emotion diarization: Which emotion appears when? In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2023. URL https://arxiv.org/abs/2306.12991.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989. doi: 10.1162/neco.1989.1.2.270.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. Beyond silent letters: Amplifying LLMs in emotion recognition with vocal nuances. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2202–2218, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.117. URL https://aclanthology.org/2025.findings-naacl.117/.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-Reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23829–23851, 2025. URL https://arxiv.org/abs/2503.02318.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. SECap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19323–19331, 2024. doi: 10.1609/aaai.v38i17.29902. URL https://arxiv.org/abs/2312.10381.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

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, Jieping Ye, and Pheng-Ann Heng. MME-Emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.09210.

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Weixuan Chen, Xihan Wei, and Liefeng Bo. HumanOmni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025. URL https://arxiv.org/abs/2501.15111.

APPENDICES

Section	Contents	Page
A	**Implementation Details**	14
A.1	Model	14
A.2	Audio Preprocessing	14
A.3	Target Annotations	14
A.4	Multi-Span Design Choices	15
B	**Training Ablation**	16
B.1	Context Information and Masking	16
B.2	Masking Probability	16
B.3	Trainable Components	16
C	**Additional Experiments**	18
C.1	Emotion-Only Data	18
C.2	GRPO Fine-tuning	18
D	**Datasets**	20
D.1	Data Statistics	20
D.2	Tone Descriptions	20
E	**Qualitative Results**	23
E.1	Example 1: Emotion recognition and temporal localization	23
E.2	Example 2: Emotion recognition with transcription differences	24
E.3	Example 3: Recovering three emotion spans	25
E.4	Example 4: Correct boundaries but incorrect emotions	26
E.5	Example 5: Merging utterances with different emotions	27
E.6	Example 6: Incomplete temporal coverage	29
E.7	Example 7: Emotion and temporal accuracy	31
E.8	Example 8: Short emotional utterances	32
E.9	Example 9: Three-emotion sequence	33
E.10	Example 10: Emotion confusion	34
E.11	Example 11: Merged utterances	35
E.12	Example 12: Premature onset	37
F	**Limitations**	39
G	**Prompts**	40
G.1	Model Training and Context Ablations	40
G.2	Model Evaluation	42
G.3	Tone Extraction	43
G.4	Tone Refinement	44

A Implementation Details

A.1 Model

We use Qwen2-Audio-7B-Instruct[1] as our backbone. We apply LoRA (Hu et al., 2022) to both the language-model and audio-encoder components, while also fine-tuning the multimodal projection layer. We use LoRA with rank $r = 16$, scaling parameter $\alpha = 32$, and dropout 0.05. The newly added timestamp-token rows in the input embedding matrix and output head are also trainable. We use an effective batch size of 48 and optimize with AdamW with a learning rate of 10^{-5}, using a cosine scheduler over 50 epochs.

The masking is introduced after two epochs, while the multi-span training starts at epoch 13. The loss weights are set to $\lambda_{\text{emo}} = 1.5$, $\lambda_{\text{start}} = 0.25$, and $\lambda_{\text{end}} = 0.25$. The auxiliary loss weights are gradually increased to their respective maximum values over the first three training epochs. We repeat this warm-up over the first two epochs after introducing multi-span samples to reduce training instability. For the auxiliary losses, we apply context masking with probability $p_{\text{mask}} = 1.0$, using tone-only masking in 60% of cases and tone-and-speech masking in the remaining 40%. For timestamp loss, we use δ=1, $c_{\min}$=0.25 and $c_{\max}$=3.0 We do not use early stopping and only evaluate the final checkpoint. Sections B.2 and B.3 discuss the masking probability and LoRA configuration, respectively.

Training used eight AMD MI250X GPUs and took approximately 46 hours per run. The total compute used for training, ablations, and unsuccessful experiments was approximately $30{,}000$ GPU-hours.

A.2 Audio Preprocessing

We process audio as mono waveforms at 16 kHz. For transcripts, we lowercase the text, remove punctuation except apostrophes, and normalize whitespace. Each candidate speech region is transcribed using WhisperX with Whisper large-v3[2] and compared with the reference using token overlap. For segments between selected speech regions, we compute RMS level, peak amplitude, frame-level energy variation, and the proportion of voiced frames. We retain intervals as natural pauses or sighs when their RMS and peak levels are at most -35 and -25 dBFS, respectively. If the selected regions and intervening intervals can be retained continuously, we crop a single interval. Otherwise, we concatenate the retained regions and update their positions in the resulting audio clip.

To remove residual noise or extra speech at the boundaries, we use WhisperX again for forced alignment. We crop from the start of the first aligned word to the end of the last aligned word, with small margins on both sides. We compare transcriptions before and after trimming and accept the refined crop only if it preserves the same transcription.

For MAFW and DEFW, in addition to removing empty samples, we remove non-English and unclear samples. Since there are no ground-truth annotations for these datasets, we compare transcriptions from FlamingoNext and Whisper, and if there is less than 60% alignment, we discard the samples.

A.3 Target Annotations

Emotion labels: We standardize emotion labels across MELD, IEMOCAP, EMOVDB, MAFW, and DEFW using eight emotion categories: neutral, happiness, sadness, anger, surprise, fear, disgust, and frustration. Labels outside these categories are mapped to the closest emotion where appropriate; otherwise, the corresponding samples are removed. Specifically, we map both "joy" and "excitement" to "happiness" and remove samples labeled "other," "sleepy," "worried," "contempt," "anxiety," "helplessness," or "disappointment."

Tone Extraction and Refinement The full FlamingoNext[3] response is passed to Qwen3-14B[4], which returns a JSON object with `speech` and `tone` fields. The extraction prompt in Section G.3

[1] https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct
[2] https://huggingface.co/openai/whisper-large-v3
[3] https://huggingface.co/nvidia/audio-flamingo-next-think-hf
[4] https://huggingface.co/Qwen/Qwen3-14B

asks Qwen3 to combine speech fragments and summarize the main speaker's delivery. The reference emotion is included in the extraction input for consistency checking.

For tone refinement, each clip is divided into five segments and processed with openSMILE. The segment-level features, speaker-specific reference statistics, and initial tone description are supplied to ChatGPT to produce the refined description. The refinement input includes speaker references for characteristics such as pitch, intensity, and speaking pace. Speaker-specific acoustic references help interpret features relative to the speaker's usual delivery. For example, a high absolute pitch does not necessarily indicate that the speaker has raised their pitch. We provide these references during refinement so that descriptions can account for individual vocal characteristics. Segment-level measurements additionally provide evidence about changes within an utterance that may be shadowed by a single clip average. These inputs support more consistent descriptions of pitch, intensity, pace, and voice quality.

A.4 Multi-Span Design Choices

Sampling and silence intervals. Source utterances are sampled without replacement from our refined sample pool. Their order is randomized without requiring an emotion change at every boundary. Consecutive utterances with the same emotion remain separate annotated spans. Compositions that exceed the supported recording duration are resampled. An interval classified as silence in our audio preprocessing is checked again after sampling the actual waveform crop. We inspect its overall energy, peaks, and the occurrence of active frames to reduce the chance of introducing unannotated speech or background activity. Silence is selected to match the sampled gap duration. When a suitable single interval is unavailable, two intervals with similar energy levels are joined to reach that duration.

Speaker numbering and timestamp updates. Speaker tags are assigned within each constructed recording in order of first appearance. The first distinct speaker identifier receives `<speaker1>`, the next receives `<speaker2>`, and so on. When a speaker identifier occurs again, its original number is reused. For example, a sequence of speakers A, B, A is annotated as `<speaker1>`, `<speaker2>`, and `<speaker1>`. Numbering restarts for each recording, so the same speaker may receive different numbers in different examples. We track the starting sample index of each utterance and calculate the temporal offset δ_i. A source span $[s_i, e_i]$ is then recorded as $[s_i + \delta_i, e_i + \delta_i]$. Offsets also account for the inserted silence and overlap.

B TRAINING ABLATION

We examine how training context, loss components, masking probability, and trainable parameters affect performance.

B.1 CONTEXT INFORMATION AND MASKING

Table 3 shows a per-dataset breakdown of the context ablation and our objective.

Table 3: Effect of training context and loss components. Single-span results are Emo-Hit Accuracy on the raw test sets. Temporal metrics are Span F1 at a tolerance of 0.25 s and mean timestamp IoU (mIoU).

	MELD	IEMOCAP	MELD Multi-Span				IEMOCAP Multi-Span			
Setting	**Emo-Hit Acc.**	**Emo-Hit Acc.**	**Emo-Duration F1**	**Emotion F1**	**Span F1 @ 0.25 s**	**mIoU**	**Emo-Duration F1**	**Emotion F1**	**Span F1 @ 0.25 s**	**mIoU**
Partial Context SFT										
Emotion only	54.44	66.36	52.23	54.40	–	–	62.85	60.11	–	–
Speech + Emotion	52.98	66.00	49.09	49.50	–	–	63.25	60.20	–	–
Tone + Emotion	51.87	56.30	49.91	52.15	–	–	52.75	51.23	–	–
Refined Tone + Emotion	53.60	60.12	49.98	53.44	–	–	56.21	53.86	–	–
Full context										
SFT	53.91	61.09	50.78	53.58	94.49	95.11	57.76	55.39	74.65	86.71
SFT+TAG	53.72	61.09	52.01	54.35	96.08	96.41	59.83	55.33	83.00	89.54
SFT+M-TAG (EMO-TAG)	**56.13**	**66.97**	**56.60**	**58.79**	**98.45**	**98.32**	**65.65**	**62.26**	**85.23**	**91.96**
w/o emotion loss	54.02	61.39	51.47	54.90	98.01	97.68	56.01	53.34	83.69	90.11
w/o temporal loss	55.21	65.88	52.83	56.57	95.38	95.89	61.64	59.83	73.08	87.22

As was shown in the main paper, emotion-only training achieves surprisingly strong performance, outperforming all other settings except our approach and speech-plus-emotion training on IEMOCAP. The bottom part of the table presents an ablation of our objective. Without context masking, adding the auxiliary losses yields modest and inconsistent improvements over SFT, suggesting that simply increasing the weight of emotion and timestamp tokens is insufficient. Introducing context masking considerably improves this performance.

As expected, removing the emotion loss mainly hurts emotion recognition, while removing the timestamp loss reduces temporal accuracy. Since Emotion-Duration F1 also depends on emotion accuracy, we additionally report Span F1 @0.25 and mean intersection over union (mIoU) to assess temporal performance. Span F1 @0.25 is the harmonic mean of span precision and recall, with a prediction considered correct when both its start and end times are within 0.25 seconds of the reference boundaries. mIoU measures the average intersection over union between predicted and reference time spans.

B.2 MASKING PROBABILITY

Figure 3 compares five context masking probabilities, from $p = 0$ (no masking) to $p = 1$ (masking in every auxiliary-loss pass).

The effect of masking remains consistent across the evaluated probabilities. We use $p = 1$ for Emo-TAG in the main experiments.

B.3 TRAINABLE COMPONENTS

Table 4 compares LoRA on the language model alone with additional adaptation of the audio encoder and projection layer.

Adding audio-encoder LoRA improves multi-span and MME performance, with little change on single-span IEMOCAP. Training the projection layer as well gives the best results across all four settings.

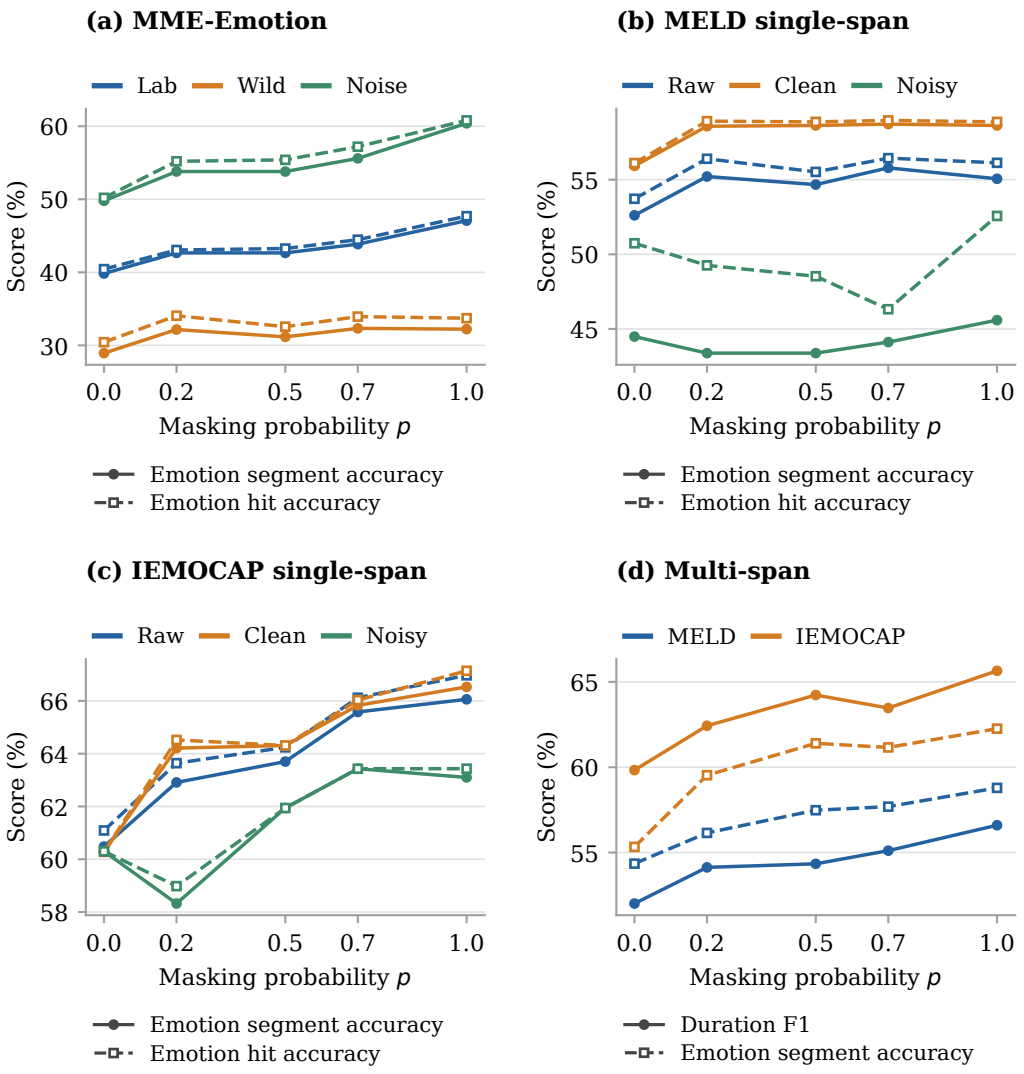


Figure 3: Effect of context masking probability across emotion recognition benchmarks. All runs use our full objective with encoder–decoder LoRA and a trainable projection layer; $p = 0$ disables context masking. Markers show the evaluated probabilities $\{0, 0.2, 0.5, 0.7, 1\}$. All scores are percentages.

Table 4: Ablation of trainable components. Components are added cumulatively. All configurations use the full objective with probability $p = 1$. MME, MELD, and IEMOCAP report average emotion hit accuracy; Multi-span reports average emotion-duration F1.

Configuration	**MELD** Hit	**IEMOCAP** Hit	**MME** Hit	**Multi-span** Dur. F1
LLM only LoRA	54.03	63.40	42.35	54.11
+ Audio-encoder LoRA	54.64	63.24	44.75	58.26
+ Projection layer (Emo-TAG)	**55.86**	**65.84**	**47.40**	**61.13**

C Additional Experiments

C.1 Emotion-Only Data

Table 5 compares emotion-only and full supervision when adding DEFW and MAFW training samples. For emotion-only supervision, we use teacher forcing and apply the masked affective loss only to the first token of each emotion label. All other tokens in these samples contribute zero loss.

Table 5: Effect of training with additional data. Single-span scores are mean Emo-Hit Accuracy over raw, clean, and noisy test sets; multi-span scores are Emotion-Duration F1. Multi-Span means using multi-span data in addition to single-span data. Full means using the full training objective rather than only the emotion loss for the additional data.

	SFT	**Emo-TAG**	**Emo-TAG + DEFW + MAFW**			
Supervision Training spans			Emotion only (Multi-Span)	Full (Multi-Span)	Emotion only Single	Full Single
MME-Emotion						
Lab	42.05	**47.69**	44.06	44.47	<u>45.27</u>	42.05
Wild	31.83	33.72	35.33	**37.00**	35.44	<u>36.44</u>
Noise	53.20	60.80	<u>61.20</u>	59.40	**61.60**	56.60
Single-span						
MELD	51.91	**55.86**	54.56	53.61	<u>55.23</u>	52.96
IEMOCAP	60.18	**65.84**	64.53	<u>64.97</u>	64.65	64.40
DEFW	41.76	46.37	52.70	**53.62**	51.78	<u>53.36</u>
MAFW	33.56	34.79	**38.17**	36.56	35.48	<u>38.10</u>
Multi-span						
MELD	50.78	**56.60**	53.82	52.66	<u>54.17</u>	53.90
IEMOCAP	57.76	**65.65**	<u>63.86</u>	63.58	63.57	62.68

As shown in the table, the additional data from DEFW and MAFW improve performance on their respective test sets, as well as on the Wild subset of MME-Emotion and, with emotion-only supervision, its Noise subset. However, adding these data slightly harms performance on IEMOCAP and MELD, which may reflect differences between the nature of these datasets.

The important takeaway from this experiment is that using only emotion supervision for these additional datasets maintains competitive performance and sometimes even surpasses full supervision. This shows the flexibility of our proposed objective.

C.2 GRPO Fine-tuning

We also ran a preliminary GRPO experiment, initializing the policy from our SFT model and using a frozen copy as the reference policy. We continued training the SFT LoRA adapter, keeping the same output format: speech, timestamps, tone, and emotion.

For each generated response, we computed a weighted combination of format, timestamp, emotion, speech, and tone rewards:

$$R = 0.10R_{\text{fmt}} + 0.30R_{\text{ts}} + 0.30R_{\text{emo}} + 0.15R_{\text{sp}} + 0.15R_{\text{tone}}. \tag{7}$$

The format reward measures compliance with the required output structure, including the expected number of spans and the absence of additional text. The timestamp reward decreases linearly with the sum of absolute start- and end-time errors, reaching zero at a specified error tolerance. Span scores are summed and divided by the larger of the predicted and reference span counts, penalizing missing or additional spans. The emotion reward measures the proportion of reference spans assigned the correct emotion, comparing predictions and references in sequence order. The speech reward gives full credit below a specified word error rate threshold and decreases linearly to zero at

Table 6: Comparison of SFT, GRPO, and EMO-TAG. Single-span scores are mean Emo-Hit Accuracy over the Original, Refined, and Noisy sets; multi-span scores are Emotion-Duration F1 on the Multiple sets. MME scores are mean Emo-Segment Accuracy over Noise, Wild, and Lab.

	IEMOCAP		**MELD**		**MME**
Method	Single-span	Multi-span	Single-span	Multi-span	Average
SFT	60.18	57.76	51.91	50.78	42.36
SFT + GRPO	61.15	58.21	52.33	51.01	43.16
EMO-TAG	**65.84**	**65.65**	**55.86**	**56.60**	**47.40**

a word error rate of one. The tone reward uses the cosine similarity between the tone embeddings extracted with FlagEmbedding [5]

For each audio input, we sample three responses across eight distributed processes and computed advantages using the pooled rewards with gradient accumulation of four. We centered rewards by their group mean and normalized them by the group standard deviation, with a lower bound on the denominator to avoid amplifying small reward differences. The optimization objective also included a KL penalty relative to the frozen SFT reference policy.

As shown in table 6, GRPO yielded a modest improvement over the SFT model. In early experiments, it also occasionally switched to Chinese during generation.

[5] https://huggingface.co/BAAI/bge-large-en-v1.5

D DATASETS

Our filtering pipeline is aggressive in order to remove as many low-quality samples as possible. This comes with the cost of losing many good samples, especially very short ones.

D.1 DATA STATISTICS

Table 7 reports sample counts by span setting and the number of samples removed from each dataset split.

Table 7: Dataset composition and filtering statistics. Top: composition of our training and test datasets, where MELD represents a more natural setting and EmoV-DB and IEMOCAP represent controlled and acted settings. * Additional samples with overlapping speech spans. Bottom: filtering statistics for the original datasets.

(a) Temporal dataset composition

Spans	**MELD**	**EmoV-DB**	**IEMOCAP**	**DFEW**	**MAFW**
1	6,818	2,565	3,484	1,911	1,311
2	2,233 + 703*	1,066 + 502*	1,331 + 555*	241	163
3	1,606 + 520*	822 + 411*	967 + 411*	188	132
4	186 + 61*	105 + 54*	108 + 42*	20	10
Total	**10,843 + 1,284***	**4,558 + 967***	**5,890 + 1,008***	**2,360**	**1,616**

(b) Filtering statistics

Dataset	**Split**	**Original**	**Removed**	**Retained (%)**
EmoV-DB	Train	5,256	2,691	48.80%
MELD	Train	9,989	3,171	68.26%
	Test	2,610	541	79.27%
IEMOCAP	Train	5,810	2,326	59.97%
	Test	1,623	631	61.12%
DFEW	Train	9,356	7,445	20.43%
	Test	2,341	1,582	32.42%
MAFW	Train	7,333	6,022	17.88%
	Test	1,839	797	56.66%

Figure 4 summarizes the dataset cleaning and tone generation pipeline.

D.2 TONE DESCRIPTIONS

Table 8 compares five original tone descriptions with their refined versions. The originals often describe emphasis on particular words and changes in delivery. The refined versions use a more consistent acoustic structure but sometimes lose these details.

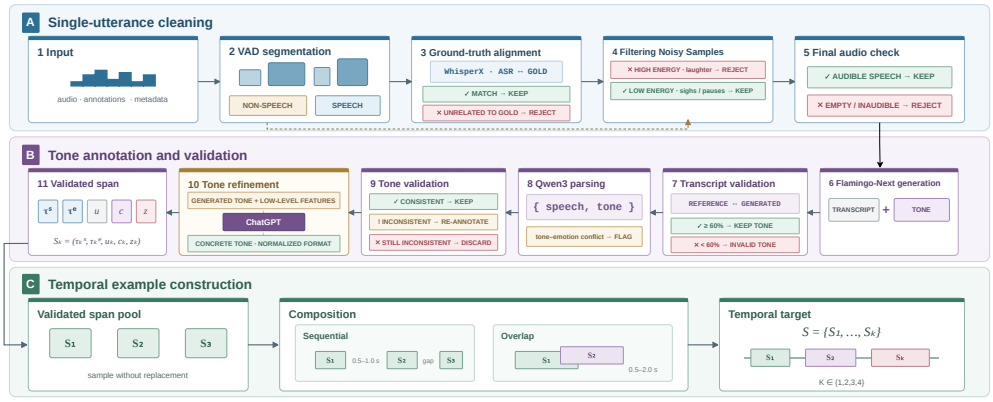


Figure 4: Overview of the data-processing pipeline, from speech and tone validation to the construction of cleaned single- and multi-span datasets.

Table 8: Selected comparisons showing the complete `target_text` from the original and refined files.

Version	Complete target text
Sample	`meld_train_dia613_utt8`
Original	`[speech]` `<speaker1>`: ”Yes!! Oh.” from `<|0.00|>s` to `<|1.23|>s`, `[tone]` Starts with a high pitch, fast, loud exclamation of ’Yes!’, then transitions to a soft, slow, low-pitched ’Oh.’ with subdued energy., `[emotion]` surprise.
Refined	`[speech]` `<speaker1>`: ”Yes!! Oh.” from `<|0.00|>s` to `<|1.23|>s`, `[tone]` Deliberate, restrained delivery with high, broadly varying pitch, strong, sustained loudness, slow, continuous pacing, and unstable pitch control; pitch gradually falls while loudness remains largely steady, `[emotion]` surprise.
Sample	`meld_train_dia1010_utt9`
Original	`[speech]` `<speaker1>`: ”No-o-o! No way!” from `<|0.00|>s` to `<|3.39|>s`, `[tone]` Deep, drawn-out ’No’ with strain and a sigh, then rapid, loud ’No way’ without pitch change, showing escalating frustration and disbelief., `[emotion]` disgust.
Refined	`[speech]` `<speaker1>`: ”No-o-o! No way!” from `<|0.00|>s` to `<|3.39|>s`, `[tone]` Alert, startled delivery with high, narrowly varying pitch, soft, even loudness, slow pacing with frequent pauses, and breathy-leaning voicing; pitch gradually rises while loudness remains largely steady, `[emotion]` disgust.
Sample	`meld_train_dia791_utt11`
Original	`[speech]` `<speaker1>`: ”I’m not gonna be able to do that for so long, and it’s so much fun!” from `<|0.00|>s` to `<|4.25|>s`, `[tone]` Voice remains bright and forward, pitch rises emphatically when discussing duration, peaks louder and faster during ’so much fun!’ with energetic intensity and no signs of hesitation or distress., `[emotion]` joy.
Refined	`[speech]` `<speaker1>`: ”I’m not gonna be able to do that for so long, and it’s so much fun!” from `<|0.00|>s` to `<|4.25|>s`, `[tone]` Animated, energetic delivery with mid-level, broadly varying pitch, strong, strongly fluctuating loudness, measured pacing with frequent pauses, and clear, stable voicing; pitch gradually rises while loudness remains largely steady, `[emotion]` joy.

Continued on next page

Table 8 (continued)

Version	Complete target text
Sample	meld_train_dia116_utt3
Original	[speech] <speaker1>: "Wow, you ah, you really like all this space stuff, huh?" from <|0.00|>s to <|4.41|>s, [tone] Calm, conversational delivery with relaxed, mildly curious mood. Starts with hesitation, steady mid-range pitch, moderate pace, slight emphasis on 'really,' and rising intonation on 'space stuff,' conveying casual interest and gentle surprise., [emotion] surprise.
Refined	[speech] <speaker1>: "Wow, you ah, you really like all this space stuff, huh?" from <|0.00|>s to <|4.41|>s, [tone] Restrained, startled delivery with low, broadly varying pitch, soft, even loudness, slow pacing with frequent pauses, and firm, slightly pressed voicing; pitch and loudness fluctuate unevenly across the utterance, [emotion] surprise.
Sample	meld_train_dia139_utt0
Original	[speech] <speaker1>: "I really thought you making a good point. I mean y'know, until you got cut off." from <|0.00|>s to <|3.79|>s, [tone] Calm, neutral, and reflective with slight emphasis on 'good point' and a marginal increase in volume and speed when mentioning being cut off., [emotion] neutral.
Refined	[speech] <speaker1>: "I really thought you making a good point. I mean y'know, until you got cut off." from <|0.00|>s to <|3.79|>s, [tone] Calm, controlled delivery with low, moderately varying pitch, soft, strongly fluctuating loudness, fast pacing with frequent pauses, and clear, stable voicing; pitch gradually falls while loudness remains largely steady, [emotion] neutral.

E QUALITATIVE RESULTS

Green marks exact reference matches; red marks errors or extra predictions. Baseline models do not predict tone descriptions. One important observation from the results is how multiple samples in the MELD benchmark have incorrect annotations. One such instance is E.4. All models agree that the delivery is natural, which can also be verified by listening to the audio, but the annotation is surprise. Similarly, all models agree on the part missing from the ground-truth transcription in E.3.

E.1 EXAMPLE 1: EMOTION RECOGNITION AND TEMPORAL LOCALIZATION

Success `MELD_test_concat_000240_k2_1ea2e4937f`

Emo-TAG identifies both emotions and their boundaries correctly. Audio-Reasoner also recognizes the emotions but misplaces the boundaries.

Source	Emotion–span sequence
Reference	anger [0.00, 2.00]; sadness [2.75, 4.25]
Emo-TAG	anger [0.00, 2.00]; sadness [2.75, 4.25]
FlamingoNext	frustration [0.38, 2.79]; sadness [5.13, 6.00]
Audio-Reasoner	anger [0.00, 3.47]; sadness [3.47, 5.08]

Span / emotion / time	Speech and tone
Reference annotation	
Reference / 1 <speaker1> **anger** <|0.00|>s–<|2.00|>s	**Speech:** "People have got to finish their stories!" **Tone:**
Reference / 2 <speaker2> **sadness** <|2.75|>s–<|4.25|>s	**Speech:** "Okay, I'm sorry." **Tone:**
Emo-TAG prediction	
Emo-TAG / 1 <speaker1> **anger** <|0.00|>s–<|2.00|>s	**Speech:** "People have got to finish their stories!" **Tone:** High pitch, fast pace, and loud delivery with clear articulation, conveying urgency and frustration through heightened intensity
Emo-TAG / 2 <speaker2> **sadness** <|2.75|>s–<|4.25|>s	**Speech:** "Okay, I'm sorry." **Tone:** Very quickly, with a noticeably higher-than-normal pitch and a sense of mild strain or urgency rather than relaxed normal talking. She doesn't hesitate long between phrases, so the entire exchange feels rushed and somewhat pressured
FlamingoNext prediction	
FlamingoNext / 1 <speaker1> **frustration** <|0.50|>s–<|2.75|>s	**Speech:** "People have got to finish their stories." **Tone:**
FlamingoNext / 2 <speaker2> **sadness** <|5.25|>s–<|6.00|>s	**Speech:** "Okay, I'm sorry" **Tone:**
Audio-Reasoner prediction	
Audio-Reasoner / 1 <speaker1> **anger** <|0.00|>s–<|3.50|>s	**Speech:** "People have got to finish their stories!" **Tone:**

Continued on next page

Example 1: annotations (continued)

Span / emotion / time	Speech and tone
Audio-Reasoner / 2 <speaker2> **sadness** <|3.50|>s–<|5.00|>s	**Speech:** "Okay, I'm sorry." **Tone:**

E.2 Example 2: Emotion recognition with transcription differences

Success `MELD_test_concat_000477_k2_948feb96e7`

FlamingoNext transcribes "How neat is that?" and labels it happiness; the reference says "how needy is that?" and is labeled disgust. Emo-TAG gets both emotions and their boundaries right despite errors in its transcript.

Source	Emotion–span sequence
Reference	disgust [0.00, 1.25]; neutral [2.00, 3.50]
Emo-TAG	disgust [0.00, 1.25]; neutral [2.00, 3.50]
FlamingoNext	happiness [0.00, 1.00]; neutral [3.00, 5.00]
Audio-Reasoner	neutral [0.00, 2.08]; neutral [2.08, 4.08]

Span / emotion / time	Speech and tone
Reference annotation	
Reference / 1 <speaker1> **disgust** <|0.00|>s–<|1.25|>s	**Speech:** "I mean,... how needy is that?" **Tone:**
Reference / 2 <speaker2> **neutral** <|2.00|>s–<|3.50|>s	**Speech:** "Okay, well then bring her in." **Tone:**
Emo-TAG prediction	
Emo-TAG / 1 <speaker1> **disgust** <|0.00|>s–<|1.25|>s	**Speech:** "The boy? How needy is that?" **Tone:** Noticeably higher-than-normal pitch, sounding startled and slightly incredulous instead of neutral or casual. Her speaking speed stays steady around mid-range, neither rushed nor drawn out, conveying mild surprise through inflection rather than dramatic signs
Emo-TAG / 2 <speaker2> **neutral** <|2.00|>s–<|3.50|>s	**Speech:** "Okay, well then bring her with you." **Tone:** Neutral, composed delivery with steady pace, even stress distribution, and no emotional inflection or vocal variation
FlamingoNext prediction	
FlamingoNext / 1 <Speaker 1> **happiness** <|0.00|>s–<|1.00|>s	**Speech:** "How neat is that?" **Tone:**
FlamingoNext / 2 <Speaker 2> **neutral** <|3.00|>s–<|5.00|>s	**Speech:** "...will then bring her..." **Tone:**
Audio-Reasoner prediction	
Audio-Reasoner / 1 <speaker1> **neutral** <|0.00|>s–<|2.00|>s	**Speech:** "How needy is that?" **Tone:**

Continued on next page

Example 2: annotations (continued)

Span / emotion / time	Speech and tone
Audio-Reasoner / 2 <speaker2> **neutral** <|2.00|>s–<|4.00|>s	**Speech:** "Okay, well then bring her." **Tone:**

E.3 Example 3: Recovering three emotion spans

Success `MELD_test_concat_000285_k3_6033afb40c`

Emo-TAG recovers all three emotion spans but adds words to the final transcript. FlamingoNext splits the first utterance, while Audio-Reasoner misclassifies the first and last emotions.

Source	Emotion–span sequence
Reference	disgust [0.00, 3.25]; neutral [3.75, 7.75]; neutral [8.25, 10.50]
Emo-TAG	disgust [0.00, 3.25]; neutral [3.75, 7.75]; neutral [8.25, 10.50]
FlamingoNext	disgust [0.50, 2.40]; disgust [3.60, 7.80]; neutral [9.00, 12.00]; fear [12.50, 15.00]
Audio-Reasoner	frustration [0.00, 2.98]; neutral [2.98, 8.08]; fear [8.08, 10.56]

Span / emotion / time	Speech and tone
Reference annotation	
Reference / 1 <speaker1> **disgust** <|0.00|>s–<|3.25|>s	**Speech:** "No ... the leather sticks to my ass." **Tone:**
Reference / 2 <speaker2> **neutral** <|3.75|>s–<|7.75|>s	**Speech:** "All-all right, well make sure you tell him that Joey Tribbiani stopped by to drop off all of these clothes." **Tone:**
Reference / 3 <speaker3> **neutral** <|8.25|>s–<|10.50|>s	**Speech:** "I just have to clean that up." **Tone:**
Emo-TAG prediction	
Emo-TAG / 1 <speaker1> **disgust** <|0.00|>s–<|3.25|>s	**Speech:** "No! The leather sticks to my ass!" **Tone:** Flat, resigned delivery with low pitch, moderate volume, slow pace, heavy drawl, and noticeable breathing, conveying irritation and weariness
Emo-TAG / 2 <speaker2> **neutral** <|3.75|>s–<|7.75|>s	**Speech:** "All right well, make sure you tell him that Joey Tribbiani stopped by and dropped off all these clothes." **Tone:** Steady, even manner with normal pitch and moderate volume, showing no emotional variation
Emo-TAG / 3 <speaker3> **neutral** <|8.25|>s–<|10.50|>s	**Speech:** "I don't mean to scare you, I just have to clean that up." **Tone:** Calm, steady, and controlled with moderate pace and normal volume, showing mild concern without panic or distress
FlamingoNext prediction	
FlamingoNext / 1 <Speaker_1> **disgust** <|0.50|>s–<|2.50|>s	**Speech:** "No." **Tone:**

Continued on next page

Example 3: annotations (continued)

Span / emotion / time	Speech and tone
FlamingoNext / 2 <Speaker_1> **disgust** <|3.50|>s–<|7.75|>s	**Speech:** "The leather sticks to my ass." **Tone:**
FlamingoNext / 3 <Speaker_1> **neutral** <|9.00|>s–<|12.00|>s	**Speech:** "All right, well, make sure you tell them Joey Tribbiani stopped by and dropped off all these clothes" **Tone:**
FlamingoNext / 4 <Speaker_2> **fear** <|12.50|>s–<|15.00|>s	**Speech:** "I didn't mean to scare ya, I just have to clean that up." **Tone:**
Audio-Reasoner prediction	
Audio-Reasoner / 1 <speaker1> **frustration** <|0.00|>s–<|3.00|>s	**Speech:** "No. The leather sticks to my ass." **Tone:**
Audio-Reasoner / 2 <speaker2> **neutral** <|3.00|>s–<|8.00|>s	**Speech:** "Alright, well, make sure you tell him that Joey Tribbiani stopped by and dropped off all these clothes." **Tone:**
Audio-Reasoner / 3 <speaker3> **fear** <|8.00|>s–<|10.50|>s	**Speech:** "I didn't mean to scare you. I just have to clean that up." **Tone:**

E.4 Example 4: Correct boundaries but incorrect emotions

Failure `MELD_test_concat_000026_k2_f016dc1781`

All three models label both utterances as neutral, missing surprise and sadness. Emo-TAG still places both boundaries correctly.

Source	Emotion–span sequence
Reference	surprise [0.00, 1.75]; sadness [2.75, 4.00]
Emo-TAG	neutral [0.00, 1.75]; neutral [2.75, 4.00]
FlamingoNext	neutral [0.24, 3.56]; neutral [4.78, 5.94]; neutral [6.54, 7.34]
Audio-Reasoner	neutral [0.00, 2.98]; neutral [3.06, 4.08]

Span / emotion / time	Speech and tone
Reference annotation	
Reference / 1 <speaker1> **surprise** <|0.00|>s–<|1.75|>s	**Speech:** "Ooh, do I sense a little bit of resentment?" **Tone:**
Reference / 2 <speaker2> **sadness** <|2.75|>s–<|4.00|>s	**Speech:** "No. Sorry." **Tone:**
Emo-TAG prediction	
Emo-TAG / 1 <speaker1> **neutral** <|0.00|>s–<|1.75|>s	**Speech:** "Do I sense a little bit of resentment?" **Tone:** Calm, conversational delivery with subtle rising intonation and deliberate pace, suggesting mild curiosity and\u5C0F\u5FC3\u601D

Continued on next page

Example 4: annotations (continued)

Span / emotion / time	Speech and tone
Emo-TAG / 2 <speaker2> **neutral** <|2.75|>s–<|4.00|>s	**Speech:** "No. Sorry." **Tone:** calm, controlled, soft-to-moderate volume, steady cadence, no emotional inflection
FlamingoNext prediction	
FlamingoNext / 1 <Speaker_0> **neutral** <|0.25|>s–<|3.50|>s	**Speech:** "Do I sense a little bit of resentment?" **Tone:**
FlamingoNext / 2 <Speaker_1> **neutral** <|4.75|>s–<|6.00|>s	**Speech:** "No." **Tone:**
FlamingoNext / 3 <Speaker_1> **neutral** <|6.50|>s–<|7.25|>s	**Speech:** "Sorry" **Tone:**
Audio-Reasoner prediction	
Audio-Reasoner / 1 <speaker1> **neutral** <|0.00|>s–<|3.00|>s	**Speech:** "Do I sense a little bit of resentment?" **Tone:**
Audio-Reasoner / 2 <speaker2> **neutral** <|3.00|>s–<|4.00|>s	**Speech:** "No, sorry." **Tone:**

E.5 Example 5: Merging utterances with different emotions

Failure `MELD_test_concat_000658_k3_65d90b2c2e`

Emo-TAG merges the last two utterances into one surprise span, missing the change to neutral and including the gap between them. The baselines produce duplicate or extra spans, and all three models misclassify the first emotion.

Source	Emotion–span sequence
Reference	happiness $[0.00, 4.25]$; surprise $[5.25, 6.00]$; neutral $[6.75, 7.00]$
Emo-TAG	neutral $[0.00, 4.25]$; surprise $[5.00, 7.00]$
FlamingoNext	neutral $[0.00, 5.00]$; surprise $[5.00, 5.00]$; neutral $[6.00, 7.00]$; neutral $[6.00, 7.00]$
Audio-Reasoner	neutral $[0.00, 5.47]$; surprise $[5.47, 6.08]$; neutral $[6.08, 6.78]$; surprise $[6.80, 7.08]$; neutral $[7.08, 7.78]$; surprise $[7.80, 8.08]$; neutral $[8.08, 8.78]$; surprise $[8.80, 9.08]$

Span / emotion / time	Speech and tone
Reference annotation	
Reference / 1 <speaker1> **happiness** <|0.00|>s–<|4.25|>s	**Speech:** "Unless, we give her all gifts she can use after she's done being pregnant." **Tone:**
Reference / 2 <speaker2> **surprise** <|5.25|>s–<|6.00|>s	**Speech:** "Oh my god." **Tone:**

Continued on next page

Example 5: annotations (continued)

Span / emotion / time	Speech and tone
Reference / 3 <speaker3> **neutral** <|6.75|>s–<|7.00|>s	**Speech:** "You think?" **Tone:**
Emo-TAG prediction	
Emo-TAG / 1 <speaker1> **neutral** <|0.00|>s–<|4.25|>s	**Speech:** "Unless we give her all gifts she can use after she's done being pregnant." **Tone:** Steady, even delivery with consistent pitch and moderate volume, showing no emotional inflection or variation in pace
Emo-TAG / 2 <speaker2> **surprise** <|5.00|>s–<|7.00|>s	**Speech:** "Oh my God! What'd you think?" **Tone:** Startled exclamation followed by urgent, inquiry-filled questioning with rising intonation, conveying shock and urgency
FlamingoNext prediction	
FlamingoNext / 1 <Speaker_1> **neutral** <|0.00|>s–<|5.00|>s	**Speech:** "Unless we give her all gifts that she can use after she's done being pregnant." **Tone:**
FlamingoNext / 2 <Speaker_2> **surprise** <|5.00|>s–<|5.00|>s	**Speech:** "Oh my god." **Tone:**
FlamingoNext / 3 <Speaker_1> **neutral** <|6.00|>s–<|7.00|>s	**Speech:** "Do you think?" **Tone:**
FlamingoNext / 4 <Speaker_2> **neutral** <|6.00|>s–<|7.00|>s	**Speech:** "Do you think?" **Tone:**
Audio-Reasoner prediction	
Audio-Reasoner / 1 <speaker1> **neutral** <|0.00|>s–<|5.50|>s	**Speech:** "Unless we give her all gifts that she can use after she's done being pregnant." **Tone:**
Audio-Reasoner / 2 <speaker2> **surprise** <|5.50|>s–<|6.00|>s	**Speech:** "Oh my God!" **Tone:**
Audio-Reasoner / 3 <speaker3> **neutral** <|6.00|>s–<|6.75|>s	**Speech:** "Do you think?" **Tone:**
Audio-Reasoner / 4 <speaker4> **surprise** <|6.75|>s–<|7.00|>s	**Speech:** "Oh my God!" **Tone:**
Audio-Reasoner / 5 <speaker5> **neutral** <|7.00|>s–<|7.75|>s	**Speech:** "Do you think?" **Tone:**
Audio-Reasoner / 6 <speaker6> **surprise** <|7.75|>s–<|8.00|>s	**Speech:** "Oh my God!" **Tone:**
Audio-Reasoner / 7 <speaker7> **neutral** <|8.00|>s–<|8.75|>s	**Speech:** "Do you think?" **Tone:**

Continued on next page

Example 5: annotations (continued)

Span / emotion / time	Speech and tone
Audio-Reasoner / 8 <speaker8> **surprise** <|8.75|>s–<|9.00|>s	**Speech:** "Oh my God!" **Tone:**
Audio-Reasoner / 9 <speaker9> **unknown** *Time not supplied*	**Speech:** "Do you think?" **Tone:**

E.6 Example 6: Incomplete temporal coverage

Failure `MELD_test_concat_000089_k3_045055ccdc`

Emo-TAG gets the emotion sequence right but ends the second span 2.25 seconds early. Both baselines miss the initial anger and add spans beyond the final reference utterance.

Source	Emotion–span sequence
Reference	anger $[0.00, 4.75]$; neutral $[5.50, 10.25]$; neutral $[10.75, 11.25]$
Emo-TAG	anger $[0.00, 4.75]$; neutral $[5.50, 8.00]$; neutral $[10.75, 11.25]$
FlamingoNext	neutral $[0.00, 2.00]$; neutral $[3.00, 4.00]$; neutral $[6.00, 7.00]$; surprise $[8.00, 19.00]$
Audio-Reasoner	neutral $[0.00, 2.00]$; neutral $[2.00, 8.00]$; neutral $[8.00, 10.00]$; neutral $[10.00, 14.00]$; neutral $[14.00, 18.00]$; neutral $[18.00, 20.00]$; neutral $[20.00, 21.00]$; neutral $[21.00, 22.00]$; neutral $[22.00, 23.00]$

Span / emotion / time	Speech and tone
Reference annotation	
Reference / 1 <speaker1> **anger** <|0.00|>s–<|4.75|>s	**Speech:** "Oh yeah? If I can't, I'll knock five bucks off the price off the unit." **Tone:**
Reference / 2 <speaker2> **neutral** <|5.50|>s–<|10.25|>s	**Speech:** "Ohh, well. Y'know we were always together, in fact the had a nickname for the three of us." **Tone:**
Reference / 3 <speaker3> **neutral** <|10.75|>s–<|11.25|>s	**Speech:** "Diverse." **Tone:**
Emo-TAG prediction	
Emo-TAG / 1 <speaker1> **anger** <|0.00|>s–<|4.75|>s	**Speech:** "Oh yeah? If I can't, I'll knock five bucks off the price of the unit." **Tone:** Starts with high pitch, fast-paced excitement, shifts to calm explanation, then ends with low-pitched, slow, grand statement
Emo-TAG / 2 <speaker2> **neutral** <|5.50|>s–<|8.00|>s	**Speech:** "Oh well, we were always together. You know, the three of us." **Tone:** Calm, steady delivery with steady pitch and volume, slight pauses between phrases, reflective and sincere affect
Emo-TAG / 3 <speaker3> **neutral** <|10.75|>s–<|11.25|>s	**Speech:** "Diversity." **Tone:** Neutral, steady delivery with moderate pace, normal volume, and clear articulation; no emotional inflection or vocal strain detected

Continued on next page

Example 6: annotations (continued)

Span / emotion / time	Speech and tone
FlamingoNext prediction	
FlamingoNext / 1 <speaker1> **neutral** <|0.00|>s–<|2.00|>s	**Speech:** "Oh yeah?" **Tone:**
FlamingoNext / 2 <speaker1> **neutral** <|3.00|>s–<|4.00|>s	**Speech:** "If I can't." **Tone:**
FlamingoNext / 3 <speaker1> **neutral** <|6.00|>s–<|7.00|>s	**Speech:** "I'll knock five bucks off the price of the unit." **Tone:**
FlamingoNext / 4 <speaker1> **surprise** <|8.00|>s–<|19.00|>s	**Speech:** "We were always together... In fact, the other kids had a nickname for the three of us... Divorce." **Tone:**
Audio-Reasoner prediction	
Audio-Reasoner / 1 <speaker1> **neutral** <|0.00|>s–<|2.00|>s	**Speech:** "Oh yeah." **Tone:**
Audio-Reasoner / 2 <speaker2> **neutral** <|2.00|>s–<|8.00|>s	**Speech:** "If I can't, I'll knock five bucks off the price of the unit." **Tone:**
Audio-Reasoner / 3 <speaker3> **neutral** <|8.00|>s–<|10.00|>s	**Speech:** "Oh well." **Tone:**
Audio-Reasoner / 4 <speaker4> **neutral** <|10.00|>s–<|14.00|>s	**Speech:** "No, we were always together." **Tone:**
Audio-Reasoner / 5 <speaker5> **neutral** <|14.00|>s–<|18.00|>s	**Speech:** "In fact, the other kids had a nickname for the three of us." **Tone:**
Audio-Reasoner / 6 <speaker6> **neutral** <|18.00|>s–<|20.00|>s	**Speech:** "Divided." **Tone:**
Audio-Reasoner / 7 <speaker7> **neutral** <|20.00|>s–<|21.00|>s	**Speech:** "Oh." **Tone:**
Audio-Reasoner / 8 <speaker8> **neutral** <|21.00|>s–<|22.00|>s	**Speech:** "Yeah." **Tone:**
Audio-Reasoner / 9 <speaker9> **neutral** <|22.00|>s–<|23.00|>s	**Speech:** "Oh." **Tone:**

E.7 Example 7: Emotion and temporal accuracy

Success `IEMOCAP_train_concat_000198_k2_674c73dbf8`

Emo-TAG matches both emotions and their boundaries. Audio-Reasoner shifts the boundary between them, while FlamingoNext splits the first utterance.

Source	Emotion–span sequence
Reference	anger $[0.00, 4.25]$; neutral $[5.00, 6.75]$
Emo-TAG	anger $[0.00, 4.25]$; neutral $[5.00, 6.75]$
FlamingoNext	anger $[1.00, 5.00]$; anger $[6.00, 9.00]$; neutral $[9.00, 11.00]$
Audio-Reasoner	anger $[0.00, 4.75]$; neutral $[4.75, 6.75]$

Span / emotion / time	Speech and tone
Reference	
Reference / 1 <speaker1> **anger** <|0.00|>s–<|4.25|>s	**Speech:** "I have been a good son for too long. A good sucker. I'm through with it." **Tone:**
Reference / 2 <speaker2> **neutral** <|5.00|>s–<|6.75|>s	**Speech:** "Um- How many people do you want to live with?" **Tone:**
Emo-TAG	
Emo-TAG / 1 <speaker1> **anger** <|0.00|>s–<|4.25|>s	**Speech:** "I have been a good son for too long. A good sucker. I'm through with it." **Tone:** Low pitch, fast pace, and high-volume delivery with clear signs of anger, conveying defiant frustration and decisiveness
Emo-TAG / 2 <speaker2> **neutral** <|5.00|>s–<|6.75|>s	**Speech:** "How many people do you want to live with?" **Tone:** Calm but tensely, with slight hesitation on 'with,' rising intonation, and restrained urgency suggesting mild impatience beneath neutral phrasing
FlamingoNext	
FlamingoNext / 1 <Speaker_0> **anger** <|1.00|>s–<|5.00|>s	**Speech:** "I have been a good son for too long" **Tone:**
FlamingoNext / 2 <Speaker_0> **anger** <|6.00|>s–<|9.00|>s	**Speech:** "a good sucker!" **Tone:**
FlamingoNext / 3 <Speaker_1> **neutral** <|9.00|>s–<|11.00|>s	**Speech:** "How many people do you want to live with?" **Tone:**
Audio-Reasoner	
Audio-Reasoner / 1 <speaker1> **anger** <|0.00|>s–<|4.75|>s	**Speech:** "I've been a good son for too long. A good sucker. I'm through with it." **Tone:**
Audio-Reasoner / 2 <speaker2> **neutral** <|4.75|>s–<|6.75|>s	**Speech:** "How many people do you want to live with?" **Tone:**

E.8 Example 8: Short emotional utterances

Success `IEMOCAP_train_concat_000230_k2_9a9af06851`

Emo-TAG correctly identifies frustration followed by sadness and matches both spans. The baselines predict sadness followed by neutral.

Source	Emotion–span sequence
Reference	frustration $[0.00, 1.75]$; sadness $[2.00, 2.75]$
Emo-TAG	frustration $[0.00, 1.75]$; sadness $[2.00, 2.75]$
FlamingoNext	sadness $[0.00, 1.00]$; neutral $[2.00, 3.00]$
Audio-Reasoner	sadness $[0.00, 2.00]$; neutral $[2.00, 3.00]$

Span / emotion / time	Speech and tone
Reference	
Reference / 1 <speaker1> **frustration** <|0.00|>s–<|1.75|>s	**Speech:** "This is humiliating." **Tone:**
Reference / 2 <speaker1> **sadness** <|2.00|>s–<|2.75|>s	**Speech:** "I didn't say that." **Tone:**
Emo-TAG	
Emo-TAG / 1 <speaker1> **frustration** <|0.00|>s–<|1.75|>s	**Speech:** "This is humiliating." **Tone:** Voice remains high pitch, loud, and fast pace with tense urgency, reflecting clear frustration and irritation
Emo-TAG / 2 <speaker2> **sadness** <|2.00|>s–<|2.75|>s	**Speech:** "I didn't say that." **Tone:** Low-pitched voice, speaks softly, moves slowly with hesitant rhythm, conveying subdued introspection and quiet sorrow through measured pace and restrained delivery
FlamingoNext	
FlamingoNext / 1 <speaker1> **sadness** <|0.00|>s–<|1.00|>s	**Speech:** "This is humiliating." **Tone:**
FlamingoNext / 2 <speaker2> **neutral** <|2.00|>s–<|3.00|>s	**Speech:** "I didn't say that." **Tone:**
Audio-Reasoner	
Audio-Reasoner / 1 <speaker1> **sadness** <|0.00|>s–<|2.00|>s	**Speech:** "This is humiliating." **Tone:**
Audio-Reasoner / 2 <speaker2> **neutral** <|2.00|>s–<|3.00|>s	**Speech:** "I didn't say that." **Tone:**

E.9 Example 9: Three-emotion sequence

Success IEMOCAP_train_concat_000067_k3_93e3a64564

Emo-TAG follows the sequence from frustration to happiness to sadness and matches all three spans, despite transcription errors.

Source	Emotion–span sequence
Reference	frustration $[0.00, 3.50]$; happiness $[4.25, 7.50]$; sadness $[8.50, 10.75]$
Emo-TAG	frustration $[0.00, 3.50]$; happiness $[4.25, 7.50]$; sadness $[8.50, 10.75]$
FlamingoNext	neutral $[0.00, 3.00]$; neutral $[1.00, 2.00]$; surprise $[4.00, 7.00]$; disgust $[8.00, 9.00]$; sadness $[9.00, 10.00]$
Audio-Reasoner	frustration $[0.00, 4.00]$; surprise $[4.00, 7.00]$; neutral $[7.25, 9.00]$

Span / emotion / time	Speech and tone
Reference	
Reference / 1 <speaker1> **frustration** <|0.00|>s–<|3.50|>s	**Speech:** "I'm growing quite tired and bored of this conversation." **Tone:**
Reference / 2 <speaker2> **happiness** <|4.25|>s–<|7.50|>s	**Speech:** "oh, God I kissed Annie. I kissed Annie. I can't How long have I waited to do that?" **Tone:**
Reference / 3 <speaker2> **sadness** <|8.50|>s–<|10.75|>s	**Speech:** "And I guess that included you." **Tone:**
Emo-TAG	
Emo-TAG / 1 <speaker1> **frustration** <|0.00|>s–<|3.50|>s	**Speech:** "I am growing quite tired and bored of this conversation." **Tone:** Mildly frustrated and weary, with a steady, conversational pace and slight strain, indicating irritation rather than intense anger or sadness
Emo-TAG / 2 <speaker2> **happiness** <|4.25|>s–<|7.50|>s	**Speech:** "God you're standing. How long have I waited to do that?" **Tone:** High pitch, fast pace, energetic, with quick articulation and rising intonation, showing excitement and urgency
Emo-TAG / 3 <speaker2> **sadness** <|8.50|>s–<|10.75|>s	**Speech:** "Can I guess that included you?" **Tone:** Soft, hesitant start with high pitch, escalating to faster pace and louder volume, showing nervous tension and anxiety
FlamingoNext	
FlamingoNext / 1 <speaker1> **neutral** <|0.00|>s–<|3.00|>s	**Speech:** "she is growing quite tired and bored of this conversation" **Tone:**
FlamingoNext / 2 <speaker2> **neutral** <|1.00|>s–<|2.00|>s	**Speech:** "" **Tone:**
FlamingoNext / 3 <speaker3> **surprise** <|4.00|>s–<|7.00|>s	**Speech:** "" **Tone:**
FlamingoNext / 4 <speaker4> **disgust** <|8.00|>s–<|9.00|>s	**Speech:** "" **Tone:**

Continued on next page

Example 9: annotations (continued)

Span / emotion / time	Speech and tone
FlamingoNext / 5 <speaker5> **sadness** <|9.00|>s–<|10.00|>s	**Speech:** “” **Tone:**
Audio-Reasoner	
Audio-Reasoner / 1 <speaker1> **frustration** <|0.00|>s–<|4.00|>s	**Speech:** “I am growing quite tired and bored of this conversation.” **Tone:**
Audio-Reasoner / 2 <speaker2> **surprise** <|4.00|>s–<|7.00|>s	**Speech:** “God, you’re standing. How long have I waited to do that?” **Tone:**
Audio-Reasoner / 3 <speaker3> **neutral** <|7.25|>s–<|9.00|>s	**Speech:** “And I guess it included you.” **Tone:**

E.10 Example 10: Emotion confusion

Failure `IEMOCAP_train_concat_000380_k2_9373a1a283`

Emo-TAG mistakes the neutral utterance for frustration despite matching both spans. FlamingoNext identifies both emotions but misplaces their boundaries.

Source	Emotion–span sequence
Reference	neutral [0.00, 2.25]; sadness [3.25, 4.25]
Emo-TAG	frustration [0.00, 2.25]; sadness [3.25, 4.25]
FlamingoNext	neutral [0.00, 3.00]; sadness [4.00, 5.00]
Audio-Reasoner	neutral [0.00, 4.00]; neutral [4.00, 6.00]; neutral [6.00, 7.00]

Span / emotion / time	Speech and tone
Reference	
Reference / 1 <speaker1> **neutral** <|0.00|>s–<|2.25|>s	**Speech:** “Are you taking the late fee off of my account then?” **Tone:**
Reference / 2 <speaker2> **sadness** <|3.25|>s–<|4.25|>s	**Speech:** “He cried hard?” **Tone:**
Emo-TAG	
Emo-TAG / 1 <speaker1> **frustration** <|0.00|>s–<|2.25|>s	**Speech:** “Are you taking the late fee off of my account then?” **Tone:** Slightly raised pitch, quickened tempo, and subtle tension suggest mild frustration, with controlled delivery maintaining polite composure
Emo-TAG / 2 <speaker2> **sadness** <|3.25|>s–<|4.25|>s	**Speech:** “He cried hard.” **Tone:** Low pitch, soft-spoken, slow pace with steady volume, conveying quiet sorrow through subdued vocal delivery

Continued on next page

Example 10: annotations (continued)

Span / emotion / time	Speech and tone
FlamingoNext	
FlamingoNext / 1 <speaker1> **neutral** <|0.00|>s–<|3.00|>s	**Speech:** "Are you taking the late fee off my account then?" **Tone:**
FlamingoNext / 2 <speaker2> **sadness** <|4.00|>s–<|5.00|>s	**Speech:** "He cried hard." **Tone:**
Audio-Reasoner	
Audio-Reasoner / 1 <speaker1> **neutral** <|0.00|>s–<|4.00|>s	**Speech:** "Are you taking the late fee off of my account then?" **Tone:**
Audio-Reasoner / 2 <speaker2> **neutral** <|4.00|>s–<|6.00|>s	**Speech:** "He cried hard." **Tone:**
Audio-Reasoner / 3 <speaker3> **neutral** <|6.00|>s–<|7.00|>s	**Speech:** "He cried hard." **Tone:**

E.11 Example 11: Merged utterances

Failure `IEMOCAP_train_concat_000080_k3_73b9cca132`

Emo-TAG combines the first two utterances into one frustration span, missing the change from neutral to frustration.

Source	Emotion–span sequence
Reference	neutral [0.00, 1.25]; frustration [2.00, 3.25]; frustration [3.75, 5.75]
Emo-TAG	frustration [0.00, 3.25]; frustration [3.75, 5.75]
FlamingoNext	neutral [0.00, 1.00]; neutral [2.00, 3.00]; neutral [4.00, 5.00]; anger (untimed); neutral (untimed)
Audio-Reasoner	frustration [0.00, 3.00]; anger [3.00, 3.50]; neutral [3.50, 4.00]; neutral [4.00, 4.00]; neutral [4.00, 4.00]; neutral [4.00, 4.25]; neutral [4.25, 4.25]; neutral [4.25, 4.25]; unknown [4.25, 4.25]

Span / emotion / time	Speech and tone
Reference	
Reference / 1 <speaker1> **neutral** <|0.00|>s–<|1.25|>s	**Speech:** "oh, She's back at work. yeah." **Tone:**
Reference / 2 <speaker1> **frustration** <|2.00|>s–<|3.25|>s	**Speech:** "And you couldn't find it?" **Tone:**
Reference / 3 <speaker2> **frustration** <|3.75|>s–<|5.75|>s	**Speech:** "It's not fair." **Tone:**

Continued on next page

Example 11: annotations (continued)

Span / emotion / time	Speech and tone
Emo-TAG	
Emo-TAG / 1 <speaker1> **frustration** <|0.00|>s–<|3.25|>s	**Speech:** "Oh, she's back at work. And you couldn't find it." **Tone:** A subtle frustration tinged with resignation, conveyed through low pitch, slow delivery and moderate volume
Emo-TAG / 2 <speaker2> **frustration** <|3.75|>s–<|5.75|>s	**Speech:** "It's not fair." **Tone:** High-pitched voice with strong emphasis on 'not,' fast pace, slight rising intonation, and steady volume, conveying clear frustration and exasperation
FlamingoNext	
FlamingoNext / 1 <speaker1> **neutral** <|0.00|>s–<|1.00|>s	**Speech:** "Oh" **Tone:**
FlamingoNext / 2 <speaker1> **neutral** <|2.00|>s–<|3.00|>s	**Speech:** "she's back at work." **Tone:**
FlamingoNext / 3 <speaker1> **neutral** <|4.00|>s–<|5.00|>s	**Speech:** "And you couldn't find her?" **Tone:**
FlamingoNext / 4 <speaker2> **anger** *Time not supplied*	**Speech:** "It's not fair!" **Tone:**
FlamingoNext / 5 <speaker1> **neutral** *Time not supplied*	**Speech:** "Well..." **Tone:**
Audio-Reasoner	
Audio-Reasoner / 1 <speaker1> **frustration** <|0.00|>s–<|3.00|>s	**Speech:** "Oh, she's back at work. And you couldn't find it." **Tone:**
Audio-Reasoner / 2 <speaker2> **anger** <|3.00|>s–<|3.50|>s	**Speech:** "It's not fair." **Tone:**
Audio-Reasoner / 3 <speaker3> **neutral** <|3.50|>s–<|4.00|>s	**Speech:** "Well, you have to be." **Tone:**
Audio-Reasoner / 4 <speaker4> **neutral** <|4.00|>s–<|4.00|>s	**Speech:** "Clippers." **Tone:**
Audio-Reasoner / 5 <speaker5> **neutral** <|4.00|>s–<|4.00|>s	**Speech:** "Clippers." **Tone:**
Audio-Reasoner / 6 <speaker6> **neutral** <|4.00|>s–<|4.25|>s	**Speech:** "Clippers." **Tone:**
Audio-Reasoner / 7 <speaker7> **neutral** <|4.25|>s–<|4.25|>s	**Speech:** "Clippers." **Tone:**

Continued on next page

Example 11: annotations (continued)

Span / emotion / time	Speech and tone
Audio-Reasoner / 8 <speaker8> **neutral** <|4.25|>s–<|4.25|>s	**Speech:** "Clippers." **Tone:**
Audio-Reasoner / 9 <speaker9> **unknown** <|4.25|>s–<|4.25|>s	**Speech:** "Clippers." **Tone:**

E.12 Example 12: Premature onset

Failure `IEMOCAP_train_concat_000263_k3_2a1b8f8fcc`

Emo-TAG starts the happiness span 1.75 seconds early, creating a 1.25-second overlap with the preceding neutral span.

Source	Emotion–span sequence
Reference	neutral [0.00, 4.50]; happiness [5.00, 6.50]; neutral [7.25, 8.75]
Emo-TAG	neutral [0.00, 4.50]; happiness [3.25, 6.50]; neutral [7.25, 8.75]
FlamingoNext	surprise [0.50, 3.75]; neutral [3.00, 7.00]; neutral [7.00, 7.25]; happiness [8.00, 9.50]
Audio-Reasoner	happiness [0.00, 7.00]; neutral [7.00, 9.00]

Span / emotion / time	Speech and tone
Reference	
Reference / 1 <speaker1> **neutral** <|0.00|>s–<|4.50|>s	**Speech:** "Wow. Well, that's really important, obviously. But, um –" **Tone:**
Reference / 2 <speaker1> **happiness** <|5.00|>s–<|6.50|>s	**Speech:** "Yeah. We'll have so much fun." **Tone:**
Reference / 3 <speaker2> **neutral** <|7.25|>s–<|8.75|>s	**Speech:** "Well We will see them this year." **Tone:**
Emo-TAG	
Emo-TAG / 1 <speaker1> **neutral** <|0.00|>s–<|4.50|>s	**Speech:** "Wow. Well that's really important obviously, but um–" **Tone:** Calm and composed with slight hesitation, moderate volume, steady pitch, and steady pace
Emo-TAG / 2 <speaker1> **happiness** <|3.25|>s–<|6.50|>s	**Speech:** "Yeah, we'll have so much fun." **Tone:** Bright, enthusiastic delivery with consistently high pitch, fast pace, and strong volume, conveying eager anticipation through animated delivery
Emo-TAG / 3 <speaker2> **neutral** <|7.25|>s–<|8.75|>s	**Speech:** "Well, we'll see them this year." **Tone:** Steady mid-range pitch with slight rising intonation on 'we'll' and flat dynamics throughout, indicating mild curiosity without emotional emphasis or vocal strain

Continued on next page

Example 12: annotations (continued)

Span / emotion / time	Speech and tone
FlamingoNext	
FlamingoNext / 1 <Speaker A> **surprise** <|0.50|>s–<|3.75|>s	**Speech:** "Wow" **Tone:**
FlamingoNext / 2 <Speaker A> **neutral** <|3.00|>s–<|7.00|>s	**Speech:** "well that's really important obviously but um yeah we'll have so much fun ever" **Tone:**
FlamingoNext / 3 <Speaker A> **neutral** <|7.00|>s–<|7.25|>s	**Speech:** "ever" **Tone:**
FlamingoNext / 4 <Speaker B> **happiness** <|8.00|>s–<|9.50|>s	**Speech:** "Well we'll see them this year" **Tone:**
Audio-Reasoner	
Audio-Reasoner / 1 <speaker1> **happiness** <|0.00|>s–<|7.00|>s	**Speech:** "Wow. Well, that's, really important obviously. But um, yeah, we'll have so much fun." **Tone:**
Audio-Reasoner / 2 <speaker2> **neutral** <|7.00|>s–<|9.00|>s	**Speech:** "Well, we'll see them this year." **Tone:**

F LIMITATIONS

Our training data are imbalanced. In MELD, neutral emotion accounts for about 42% of training samples, compared with 3% for fear. The evaluation split is similarly skewed: 47% neutral and 1.5% fear. Filtering preserves much of this imbalance. We did not manually verify or evaluate tone descriptions due to limited resources and the difficulty in defining a consistent quality criterion. Finally, we focus on eight basic emotions. Extending the label set to finer distinctions, such as those in emotion wheels (Plutchik, 1980), could make the approach more useful in real-world settings. We leave these questions for future work.

G PROMPTS

The prompts below were used for training, context ablations, evaluation, and tone extraction and refinement. Full-context training includes speech, timestamps, tone, and emotion. The partial-context settings in Table 3 vary the inclusion of speech and tone; all retain timestamps.

G.1 MODEL TRAINING AND CONTEXT ABLATIONS

G.1.1 EMOTION ONLY

System prompt.

```
You are an affective audio expert in emotion recognition.

Guidelines:
- Focus only on the audio.
- Use timestamp tokens exactly as provided.
- Follow the requested output format exactly.
```

User prompt.

```
Analyze the input audio and generate one structured annotation for emotion
  recognition.

Output the final emotion label for each segement.

Output exactly in this format:
From <|0.00|>s to <|0.25|>s,
[emotion] emotion label.

Requirements:
- Use valid timestamp tokens such as <|0.00|>, <|0.25|>, <|0.50|>, and so
  on.
- The emotion label must be the final affective interpretation of the
  audio.
```

G.1.2 SPEECH AND EMOTION

System prompt.

```
You are an affective audio expert in emotion recognition.

Guidelines:
- Focus only on the audio.
- Infer speech before emotion.
- Use timestamp tokens exactly as provided.
- Follow the requested output format exactly.
```

User prompt.

```
Analyze the input audio and generate one structured annotation for emotion
  recognition.

First determine the speaker's speech, then output the final emotion label.

Output exactly in this format:
[speech] <speakerX>: transcript from <|0.00|>s to <|0.25|>s,
[emotion] emotion label.

Requirements:
- Use valid timestamp tokens such as <|0.00|>, <|0.25|>, <|0.50|>, and so
  on.
- The emotion label must be the final affective interpretation of the
  audio.
```

G.1.3 TONE AND EMOTION

System prompt.

```
You are an affective audio expert in emotion recognition.

Guidelines:
- Focus only on the audio.
- Infer tone before emotion.
- Use timestamp tokens exactly as provided.
- Follow the requested output format exactly.
```

User prompt.

```
Analyze the input audio and generate one structured annotation for emotion
  recognition.

First determine the speaker's tone, then output the final emotion label.

Output exactly in this format:
[tone] <speakerX>: tone description from <|0.00|>s to <|0.25|>s,
[emotion] emotion label.

Requirements:
- Use valid timestamp tokens such as <|0.00|>, <|0.25|>, <|0.50|>, and so
  on.
- The emotion label must be the final affective interpretation of the
  audio.
```

G.1.4 FULL CONTEXT

System prompt.

```
You are an affective audio expert in emotion recognition.

Guidelines:
- Focus only on the audio.
- Infer speech and tone before emotion.
- Use timestamp tokens exactly as provided.
- Follow the requested output format exactly.
```

User prompt.

```
Analyze the input audio and generate one structured annotation for emotion
  recognition.

First determine the speaker's speech and tone, then output the final
  emotion label.

Output exactly in this format:
[speech] <speakerX>: transcript from <|0.00|>s to <|0.25|>s, [tone] tone
  description, [emotion] emotion label.

Requirements:
- Use valid timestamp tokens such as <|0.00|>, <|0.25|>, <|0.50|>, and so
  on.
- The tone description must describe the speaker's audible vocal
  characteristics.
- The emotion label must be the final affective interpretation of the
  audio.
```

G.2 Model Evaluation

We keep each model's original system prompt and use the following user prompt for evaluation.

```
Listen to the audio and identify each spoken utterance.

For each utterance, provide:
- the spoken words
- the start and end time in seconds
- the speaker's emotion

Emotion labels: neutral, happiness, sadness, anger, surprise, fear,
  disgust, frustration.
```

G.2.1 Qwen3 Evaluation Extractor

System prompt.

```
You are an information extraction system. You convert a model's free-form
  audio-emotion answer into JSON for evaluation. Extract only information
  stated or clearly implied by the model answer. Do not use outside
  knowledge and do not guess missing fields.
```

User prompt.

```
Extract structured fields from the model answer below.

Allowed emotion labels:
neutral, frustration, anger, sadness, happiness, surprise, fear, disgust

Return ONLY valid JSON with this schema:
{
  segments: [
    {
      speech: transcribed speech or null,
      emotion: one allowed emotion label or null,
      start_sec: ,
      end_sec:
    }
  ]
}

Rules:
- If there is no explicit or clearly implied emotion, use null for emotion.
- If there is no timestamp or time range, use null for start_sec and
  end_sec.
- If there are multiple utterances/spans, return multiple segments in
  order.
- Convert timestamp tokens like <|0.50|> to seconds as numbers.
- Do not invent timestamps, emotions, or speech.
- Do not include markdown fences, comments, or explanations.

Model answer:
{model_output}
```

G.3 TONE EXTRACTION

G.3.1 FLAMINGO PROMPT

```
Reason step by step with timestamps before answering. Listen to the speech
  and extract the spoken words and vocal tone. Do not infer the emotion
  label. For the tone be descreptive, mention emotional cues such as
  pitch, loudness, speaking rate, pauses, tension, laughter, pace,
  delivery, sighing, uncertainty, crying, shouting, whispering, or
  breathiness if audible. Focus on how the words are spoken, not just what
  is spoken. Be descriptive but do not add details not supported by the
  audio.
```

G.3.2 QWEN3 TONE EXTRACTOR

```
Extract the spoken speech and vocal tone from below.

Return ONLY valid JSON with exactly these keys:
{
  speech: ...,
  tone: ...
}

Rules:
- Use the full text, including any <think> trace.
- The speech is the transcript/spoken words if present.
- The speech is usually scattered in segments between quotation marks;
  concatenate quoted speech spans without repetition.
- Focus on the main speaker only, ignore background laughter or other
  speakers if present.
- The tone should summarize the vocal/acoustic delivery, not the semantic
  content.
- Focus on the emotional and acoustic cues, including pitch, loudness,
  pace, pauses, tension, laughter, sighing, crying, shouting, whispering,
  breathiness, hesitation, trembling, or vocal strain, if present.
- Do not add unnecessary explanation or preamble.
- Do not add details that are not supported by the text.
- If the tone is clearly inconsistent with the given emotion, output wrong
  tone as the tone.

Emotion:
{emotion}

Text:
{flamingo_raw}
```

G.4 TONE REFINEMENT

ChatGPT prompt.

```
You are given a tone description, an emotion label, speaker-specific
reference ranges, and acoustic features from five temporal segments.

Rewrite the tone description as one concrete and informative sentence.

Requirements:

- Ground every stated cue in the supplied acoustic features.

- Interpret pitch, loudness, and other features relative to the speaker's
   low, mean, and high reference values.

- Focus on emotion-relevant characteristics: pitch level and variation,
   loudness and its variation, speaking pace and pauses, voice quality, and
   changes across the utterance.

- Use the existing tone description only as a semantic reference and
  correct
   it when it conflicts with the acoustic evidence.

- Do not describe every segment separately. Summarize only meaningful
  changes
   across the five segments.

- Follow this structure:

   ''[overall delivery] with [pitch], [loudness], [pacing], and [voice
  quality];
   [meaningful acoustic descriptions].''

- Use 15--30 words, do not explicitly name the emotion, and return only
  the refined tone description.

Emotion label:

{emotion}

Existing tone description:

{original_tone}

Speaker-specific reference values:

{speaker_reference_values}

Segments acoustic features:

{segment_features}
```