Towards Comprehensive Emotional Characterization: Ambiguity-Aware Emotion Reasoning in Speech

Anonymous ACL submission

Abstract

Recent progress in *speech large language models* (SLLMs) has advanced *speech emotion recognition* (SER) from single-label classification toward fine-grained emotion reasoning. Emotions are complex and often co-occur. Therefore, modeling perceptual variability across annotators is essential. This observation suggests that we should push SER toward comprehensive emotional characterization, where reasoning is grounded in interpretable acoustic and lexical cues rather than abstract label prediction. We propose a hierarchical supervision framework that models Primary, Minor, and All Perceived Emotions, combined with rationale distillation to enable SLLMs to reason about perceptual ambiguity. Experiments on the MSP-Podcast corpus show a 3.6% improvement in Primary Macro-F1 and stronger performance on low-frequency emotion categories. Human evaluation further indicates that the generated explanations are more coherent and better aligned with the underlying audio evidence.

1 Introduction

Speech emotion recognition (SER) plays a critical role in enabling empathetic human–computer interaction systems (Jordan et al., 2025; Wani et al., 2021). Recent advances in *speech large language models* (SLLM) have shifted SER toward explainable and reasoning-based emotion modeling (Lian et al., 2025a; Lai et al., 2025; Li et al., 2025). However, the supervision paradigm in most existing SER systems remains consensus-driven (Li et al., 2025; Su et al., 2026; Wang et al., 2026), where a single ground-truth label is determined via plurality/majority voting and annotator disagreement is treated as annotation noise (Chou et al., 2024b). However, human emotional expression is inherently diverse and perceptually ambiguous (Larsen et al., 2001). A single utterance often contains multi-layered affective cues, leading to diverse yet valid emotional interpretations across listeners (Devillers et al., 2005; Lotfian and Busso, 2017, 2019). Collapsing multi-annotator perception into a single label not only oversimplifies emotional perception but also prevents SLLMs from reasoning about co-occurring emotional states (Devillers et al., 2005; Sethu et al., 2019; Chou and Lee, 2025). Beyond improving prediction accuracy, emotion understanding requires explaining why a particular affective interpretation arises and how multiple emotions may co-exist within the same utterance. Such grounded emotion reasoning is essential for trustworthy and interpretable human–computer interaction.

In this work, we further investigate whether SER can effectively leverage hierarchical multi-perception supervision within a reasoning-based modeling framework. Unlike approaches that focus solely on the dominant emotion prediction or flat multi-label modeling (Kang and Cho, 2025; Chou et al., 2022b), we explicitly model annotator disagreement as a hierarchical perceptual structure and leverage SLLM to jointly model perceptual ambiguity and subjective interpretation. We hypothesize that annotator disagreement encodes structured information about the distribution of human emotional perception and can help align model predictions with human perceptual variability.

To systematically model the multi-annotator perceptual structure, we propose a hierarchical emotion supervision framework that decomposes emotional perception into dominant and secondary emotional categories. This strategy is consistent with the concept of mixed emotions proposed in Plutchik's Psychoevolutionary Theory and its wheel of emotions representation (Plutchik, 2001). We adopt a particular emotion hierarchy dictated by the annotation process used in the MSP-Podcast corpus (Busso et al., 2026). For a given sentence, we consider the (1) Primary Emotion, defined as the plurality class, (2) Minor Emotion, defined as

the set of all other emotions selected as the dominant emotion that are different from the plurality class, and (3) *All Perceived Emotions* (APEs), defined as all the set of secondary emotions perceived in the recording. This formulation enables simultaneous modeling of dominant perception and structured perceptual variability. Building on this representation, we use SLLM to generate hierarchical rationales conditioned on different perceptual aspects, modeling the perceptual mechanisms underlying emotional ambiguity. We further transfer this hierarchical reasoning capability into an efficient student model via parameter-efficient distillation (Hsieh et al., 2023), enabling joint emotion prediction and perception-grounded explanation generation.

Experimental results on the MSP-Podcast dataset (Busso et al., 2026) demonstrate that incorporating minority emotion supervision improves SER performance and mitigates majority-class bias, yielding a 3.6% improvement in Primary Macro-F1 over the consensus-only baseline. In particular, minority supervision improves primary emotion recognition and enhances the model's ability to recover the full range of emotions perceived by annotators. Moreover, hierarchical supervision further improves distribution-level perception modeling. Human evaluation further shows that hierarchical minority-aware supervision produces more coherent and evidence-grounded rationales. To summarize, the main contributions of this paper are as follows:

- **Validation of Minority Emotions:** We demonstrate that minority annotations encode structured perceptual signals rather than noise. Incorporating minority-annotated emotion labels as additional training supervision not only boosts primary emotion recognition but also significantly enhances broad perceptual coverage.

- **Recovering Rare Emotion Characterization:** We establish that minority supervision is particularly effective for rare and low-consensus emotions (e.g., Fear, Contempt). This result suggests that minority labels capture the essential uncertainty structure within the human emotional perception space.

- **Explainable Ambiguity Alignment:** We propose hierarchical rationale distillation to model emotional ambiguity. Human evaluation proves this strategy produces grounded, consistent explanations, enabling an ambiguity-aware SER framework.

2 Related Work

Ambiguity-Aware Emotion Supervision. The evolution of SER is increasingly defined by the shift from consensus-driven classification to ambiguity-aware generative reasoning (Sun et al., 2026; Lian et al., 2025b). Traditional studies often rely on plurality-voted and majority-voted labels, effectively discarding the rich emotional signals embedded in annotator disagreement (Chou and Lee, 2025; Tavernor and Mower Provost, 2025; Sethu et al., 2019). Early efforts to recover this discarded signal include the *primary-and-secondary* multitask formulation of Lotfian and Busso (2018b), which jointly predicts the most-voted and second-most-voted emotion categories, and probabilistic perception models that treat annotator ratings as samples from a latent emotional distribution (Lotfian and Busso, 2017; Fayek et al., 2016). More recent work further treats disagreement as structured perceptual information rather than annotation noise (Lian et al., 2025a; Kang and Cho, 2025; Sun et al., 2026), culminating in all-inclusive aggregation strategies that preserve the full distribution of annotator ratings through multi-label supervision (Chou et al., 2024b). Together, these studies have established that soft supervision and emotion co-occurrence statistics are essential for characterizing expressive behaviors. However, these approaches remain predominantly discriminative and provide limited insight into *why* a particular mixture of emotions is perceived.

Explainable Emotion Reasoning with SLLMs. The rise of SLLMs has catalyzed a parallel transition toward explainable emotion reasoning. AER-LLM (Hong et al., 2025) was among the first to show that LLMs can perform zero- and few-shot ambiguous emotion recognition via in-context learning over dialogue history. AE-CoTD (Mai et al., 2025) later demonstrated that chain-of-thought rationales distilled from a reasoning teacher (DeepSeek-R1) into Qwen2.5-7B with *low-rank adaptation* (LoRA) substantially improve single-label SER performance. Subsequent frameworks further extended this direction using prosody-aware reinforcement learning (Wang et al., 2026; Lai et al., 2025) and agentic evidence-grounded probing (Li et al., 2025; Sun et al., 2026).

These trends have also been amplified by the *multimodal emotion recognition* (MER) challenge (Lian et al., 2025b) in 2025, which explicitly emphasizes fine-grained, multimodal, open-vocabulary emotion understanding.

From Majority Explanation to Hierarchical Perceptual Reasoning. Recent rationale-based SpeechLM frameworks have improved interpretability by generating natural-language explanations alongside emotion predictions. Among them, the closest predecessor to our proposed solution is the SpeechLM rationale framework of Su et al. (2026), which jointly predicts transcripts, emotion labels, and teacher-distilled rationales, and evaluates with annotator-aware metrics. Their work demonstrated that rationale supervision can improve explainability and prediction quality in SER. However, the framework remains fundamentally centered on majority-vote labels and does not explicitly model the hierarchical structure embedded in annotator disagreement.

Our work extends this line of research from majority-based rationale generation toward hierarchical perceptual reasoning. Rather than conditioning rationales solely on a single consensus label, we incorporate hierarchical emotion supervision consisting of Primary, Minor, and *All Perceived Emotions* (APE), allowing generated rationales to reflect multiple levels of human emotional perception. This formulation enables us to systematically investigate how incorporating different perceptual layers into rationale supervision influences emotion prediction, rationale quality, and ambiguity modeling. By integrating structured annotator perception into rationale learning, our framework shifts the objective from explaining a dominant emotional interpretation toward modeling richer perceptual variability.

3 Methodology

Our strategy aims to obtain a comprehensive emotional characterization by augmenting emotional labels beyond consensus (Sec. 3.1), improving the SLLM training process with a rationale-augmented generative prediction task (Sec. 3.2).

3.1 Label Construction

To systematically represent emotional content while preserving annotator subjectivity, we derive a hierarchical emotional characterization. In Plutchik's Psychoevolutionary Theory (Plutchik, 2001), mixed emotions are explained by the idea that emotions can blend similarly to colors on a color wheel. The theory organizes emotions into a wheel composed of eight primary emotions. Then, mixed emotions are obtained by combining primary emotions. We aim to capture this concept by considering dominant and secondary emotional perceptions provided by multiple evaluators. Our strategy is to leverage the perceptual evaluation used to emotionally annotate the sentences in the MSP-Podcast Corpus (Busso et al., 2026). As illustrated in Fig. 1, each speaking turn is assessed by multiple evaluators answering two distinct questions: (Q1) what is the most dominant emotion perceived in the audio (8-class taxonomy)? (Q2) Can you identify all the emotional classes perceived in this audio (16-class taxonomy)? With these two questions, we define the **Primary Emotion** (L_p), **Minor Emotion** (L_m), and **All Perceived Emotions** (APE) from raw multi-annotator data. The **Primary Emotion** (L_p) is determined by the plu-

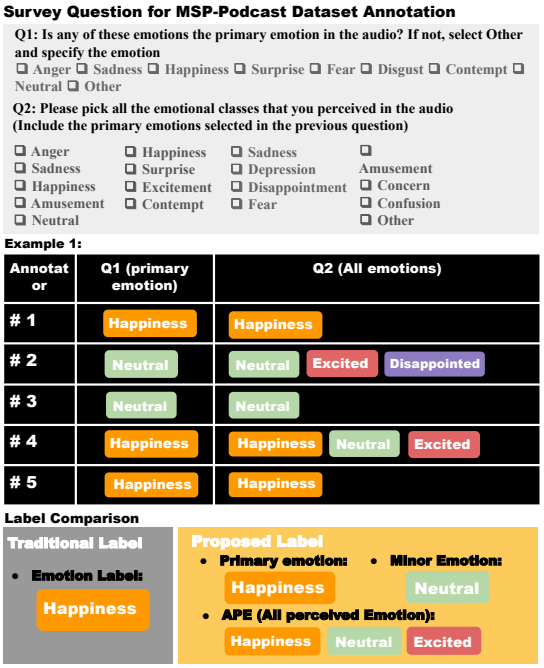


Figure 1: Illustration of hierarchical label construction from multi-annotator perception. Annotators first select a Primary emotion (Q1) and then select all perceived emotions (Q2). From these annotations, we derive three representations: Primary emotion (plurality Q1 vote), Minor emotions (non-primary Q1 selections), and APE (union of all Q2 selections). To improve annotation reliability, we use a filtered APE variant in experiments, where emotions appearing in Primary or Minor are always retained, and remaining APE-only emotions are kept only if selected by more than two annotators.

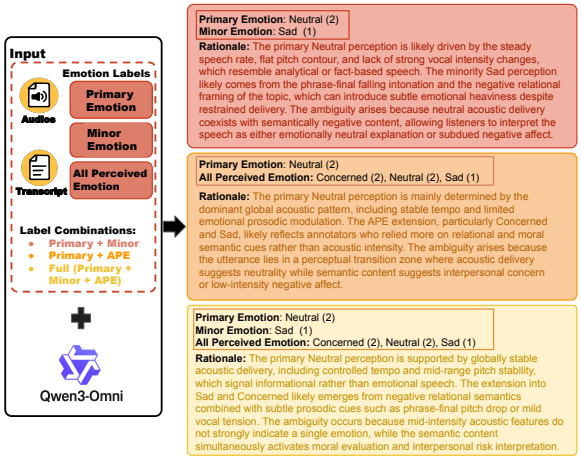


Figure 2: Example of label-conditioned hierarchical rationale generation. Given the same input audio and transcript, different supervision settings (Primary+Minor, Primary+APE, and Primary+Minor+APE) guide the model to generate rationales explaining why the Primary emotion is perceived, why Minor emotions may arise, and why perceptual ambiguity exists.

rality of Q1 votes, with *Other* responses discarded to ensure categorical focus:

$$L_p = \arg\max_{e \in E}(\text{votes}_e) \tag{1}$$

where E denotes the set of primary emotion categories and votes_e denotes the number of annotators who selected emotion e in Q1. To model perceptual ambiguity, we define **Minor Emotion** (L_m) as the set of non-plurality categories selected in Q1, thereby preserving subtle emotional signals perceived by subsets of annotators rather than treating them as noise. Finally, **APE** is defined as the union of all Q2 selections, refined via frequency-based filtering to improve reliability. Specifically, we always retain emotions appearing in L_p or L_m. Then, we retain other emotional classes mentioned in Q2 if at least two evaluators mark those emotions (e.g., the example in Fig. 1 retains Excited but not Disappointed in APE).

3.2 Rationale-Augmented SLLM Training Pipeline

We formulate SER as a rationale-augmented generative prediction task for hierarchical, multi-class emotional predictions. The model is trained to jointly predict emotional labels and generate natural-language rationales explaining both the dominant emotion perception and perceptual ambiguity. Given input audio, transcript, and supervision labels (L_p, L_m, L_{APE}), we use a teacher SLLM to generate structured rationales.

As illustrated in Fig. 2, different supervision configurations guide the model to generate rationales conditioned on different perceptual layers. Specifically, we design *chain-of-thought* (CoT) prompting templates that explicitly decompose reasoning into hierarchical perceptual explanations. The model is first prompted to explain why the Primary emotion is perceived based on acoustic and lexical evidences. It is then prompted to explain why Minor emotions or additional APE labels may arise from alternative perceptual interpretations. Finally, the model is prompted to explain why perceptual ambiguity exists, focusing on overlapping acoustic cues, mixed emotional signals, or contextual uncertainty. We evaluate different emotion hierarchies in our implementation. First, we use the *Primary + Minor* supervision, which exposes the model to disagreement patterns in Q1, encouraging it to consider both dominant- and minority-emotion interpretations. Second, we consider *Primary + APE* and *Primary + Minor + APE*. These implementations introduce multi-label emotion sets from Q2 that capture co-occurrence patterns and perceptual structure. This additional signal helps the model treat ambiguity as a structured phenomenon rather than a simple label variation. To prevent unimodal reasoning (e.g., relying solely on transcripts), we explicitly require that rationales be grounded in both acoustic and lexical evidence during prompt design, thereby improving evidence consistency and perceptual reliability.

We use Qwen-3-Omni (Xu et al., 2025) as the teacher model to generate CoT rationales grounded in acoustic and lexical evidence. The generated rationales serve as intermediate supervision signals for training a student SLLM. We construct training samples consisting of audio, rationale, and emotion labels, and perform *supervised fine-tuning* (SFT) on Qwen-2-Audio (Chu et al., 2024) using parameter-efficient *low-rank adaptation* (LoRA) (Hu et al., 2021). The student model is trained to jointly predict the transcript, emotion labels, and evidence-grounded rationales, encouraging alignment between speech content understanding, prediction decisions, and interpretable reasoning behavior. During inference, only raw audio is provided as input. The trained model outputs emotion label predictions across perceptual levels (e.g., primary and minor emotions), together with natural-language rationales explaining both label evidence and perceptual ambiguity. To further evaluate the quality and perceptual consistency of the generated

Table 1: Performance comparison of different supervision modes on the MSP-Podcast test set across primary fidelity, perceptual coverage, and distribution-level recall metrics. Minority-aware and hierarchical supervision improve primary emotion prediction, perceptual set recovery, and ambiguity-aware recall compared to consensus-based training.

Model Mode	Set	GT Size	Pred Size	P-Macro-F1	Soft Rec.	Set Rec.	Jaccard
Consensus	8-class	1.0	1.0	30.87%	27.71%	–	–
Primary + Minor	8-class	2.55	2.48	**34.53%**	54.32%	**61.50%**	**47.21%**
Primary + APE	16-class	3.89	3.21	31.98%	48.75%	59.27%	35.11%
Primary + Minor + APE	16-class	3.89	**3.57**	33.72%	**59.16%**	60.34%	34.58%

rationales, we design a human evaluation protocol to compare rationale generation across different supervision configurations.

4 Experiments

4.1 Experimental Setup

We evaluate our framework on the MSP-Podcast dataset (Busso et al., 2026), which is one of the largest spontaneous speech databases developed for the study of emotions. The recordings are sourced from podcasts featuring conversational speech across a broad range of topics. Relevant to this study are the annotations of emotional categories, collected using the questionnaire presented in Fig. 1. Each speaking turn was annotated by five or more annotators. The corpus contains recordings from more than 3,641 speakers. We use version 2.0 of the corpus, which includes 267,905 speaking turns (409 hours). The training set contains 169,190 speaking turns, while the test set contains 46,294 speaking turns (set Test1 in the release). Busso et al. (2026) provides a detailed description of the corpus.

We use **Qwen2-Audio-7B-Instruct** as the backbone model. The model is fine-tuned using SFT with parameter-efficient adaptation implemented with LoRA (rank = 8, α = 16). The maximum sequence length is set to 2,048 tokens.

We compare the conventional *Consensus (Plurality-label)* baseline against three hierarchical supervision variants: (1) $L_p + L_m$, (2) $L_p + L_{APE}$, and (3) $L_p + L_m + L_{APE}$.

4.2 Evaluation Metrics

Following the evaluation framework introduced in Sun et al. (2026), we utilize a multi-level evaluation suite to assess model performance under ambiguity-aware emotion modeling. Specifically, we evaluate performance at three complementary levels: (i) primary emotion fidelity, (ii) perceptual set coverage, and (iii) perceptual density calibration.

Primary Emotion Fidelity. We report *Primary Macro-F1* to evaluate strict primary emotion prediction accuracy. In addition, we report *Soft Recall*, defined as the proportion of samples for which the ground-truth primary emotion is included in the model's predicted emotion set under each supervision configuration.

Perceptual Set Coverage. Let Pred denote the set of emotion labels predicted by the model for a given sample, and let GT denote the ground-truth multi-annotator emotion set (e.g., APE). We report *Set-based Recall* defined as $\frac{|\text{Pred} \cap \text{GT}|}{|\text{GT}|}$ and the *Jaccard Index (IoU)* defined as $\frac{|\text{Pred} \cap \text{GT}|}{|\text{Pred} \cup \text{GT}|}$.

Perceptual Density Calibration. We define *human perceptual density* as the average number of distinct emotion labels perceived per sample. The goal is that our model predicts similar density as the one observed in the ground-truth multi-annotator emotion set (i.e., the average cardinality of the GT across the dataset). To assess whether the model captures perceptual ambiguity at a similar granularity, we report the *Average Prediction Size*, defined as the average cardinality of the Pred set across samples.

4.3 Effectiveness in Primary Emotion Prediction

As shown in Table 1, incorporating Minor Emotions (L_m) as an auxiliary supervision signal significantly improves primary emotion recognition. In particular, Primary Macro-F1 increases from 30.87% under the consensus baseline to 34.53% under the Primary + Minor configuration, representing an absolute improvement of 3.66%.

We also observe substantial gains in Soft Recall, which measures whether the ground-truth primary emotion is included in the predicted emotion set. Soft Recall increases from 27.71% to 54.32%, demonstrating that minority supervision helps preserve dominant emotional perception even under multi-label prediction settings. While the Primary+Minor+APE configuration further improves distribution-level recall (59.16% Soft Recall), the

Contrast	Δ(P-Macro-F1)	Δ(Soft Rec.)	Δ(Set Rec.)	Δ(Jaccard)
P+M vs. Consensus	$3.87 \pm 0.84^{***}$	$26.27 \pm 3.62^{***}$	–	–
P+APE vs. Consensus	$1.15 \pm 1.27^{*}$	$20.85 \pm 4.01^{***}$	–	–
P+M+APE vs. Consensus	$2.64 \pm 1.09^{**}$	$31.09 \pm 4.36^{***}$	–	–
P+M vs. P+APE	$2.71 \pm 1.39^{**}$	$5.42 \pm 2.14^{**}$	$2.04 \pm 2.37^{*}$	$12.13 \pm 3.26^{***}$
P+M vs. P+M+APE	$1.22 \pm 1.41^{*}$	$-4.82 \pm 2.52^{**}$	1.23 ± 2.58	$12.80 \pm 3.54^{***}$

Table 2: Paired significance tests across 10 sub-sample retraining runs. Each cell reports the mean paired difference $\Delta\pm$ standard deviation across the 10 matched runs, where $\Delta_i = M_i^A - M_i^B$ is computed between two configurations trained on the same subset i. Statistical significance is determined using paired t-tests over the per-subset differences: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Set Recall and Jaccard are undefined for the single-label Consensus baseline.

Primary + Minor setting achieves the strongest primary prediction accuracy and superior perceptual set coverage, as reflected by 61.50% Set-based Recall and 47.21% Jaccard. These results support our hypothesis that minority labels encode structured perceptual information rather than stochastic annotation noise.

Statistical Significance. To verify that the observed improvements reflect genuine, statistically significant effects rather than sampling variance, we construct 10 random 70% subsets of the training set using fixed seeds. We train all four supervision configurations (Consensus, P+M, P+APE, P+M+APE) on each subset using identical hyperparameters. Each model is evaluated on the full MSP-Podcast test set, yielding ten matched observations per configuration. We then perform paired t-tests across all comparisons of interest.

Table 2 reports the resulting pairwise contrasts. The entries in the table correspond to the mean difference between two configurations across the 10 matched runs. All three hierarchical configurations significantly outperform the Consensus baseline on Primary Macro-F1, confirming that incorporating richer perceptual supervision consistently improves primary emotion recognition. Among them, Primary + Minor achieves the largest gain over Consensus ($\Delta = +3.87$, $p < 0.001$), while Primary + APE yields a more modest improvement ($\Delta = +1.15$, $p < 0.05$). These results suggest that minority emotion labels provide more informative and targeted supervision than broader perceptual aggregation alone.

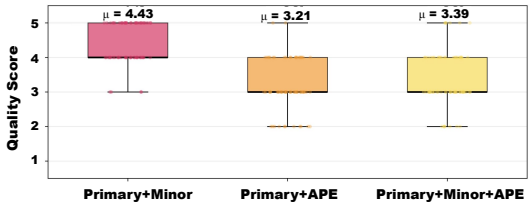
More importantly, comparisons among hierarchical supervision settings reveal a clear trade-off between primary fidelity and perceptual coverage. Compared with Primary + Minor + APE, Primary + Minor achieves significantly stronger primary prediction accuracy (Macro-F1 $+1.22$, $p < 0.05$) and substantially better set-level alignment (Jaccard $+12.80$, $p < 0.001$), whereas Primary + Minor + APE achieves stronger distribution-level recall (Soft Recall -4.82, $p < 0.01$). This finding suggests that introducing broader perceptual supervision improves emotion coverage but may dilute dominant emotional signals. Therefore, the choice of hierarchical supervision should depend on whether the task prioritizes primary emotion fidelity or comprehensive perceptual density modeling.

Table 3: Per-emotion Primary F1 (%, 8-class) across supervision modes, showing improvements on rare and ambiguity-sensitive emotions under minority-aware supervision.

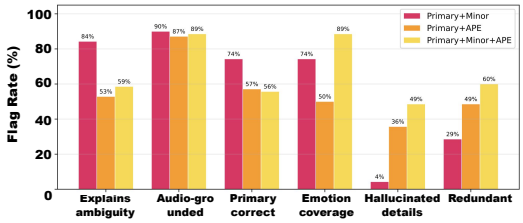
Emotion	Consensus	P+M	P+APE	P+M+APE
Happy	67.42	68.03	67.59	**68.23**
Sad	36.03	39.56	43.04	**44.58**
Angry	65.55	65.22	65.84	**66.41**
Neutral	65.11	**67.81**	61.80	63.87
Surprise	9.22	**14.07**	10.69	12.14
Fear	1.18	**7.49**	2.54	5.03
Disgust	2.03	**8.09**	1.45	4.47
Contempt	0.40	**5.99**	3.08	5.03

4.4 Emotion Coverage and Rare Emotion Recognition

Another key objective of our framework is to align model predictions with the multi-annotator perceptual density of human emotions, defined as the average cardinality of the ground-truth multi-annotator emotion set. As shown in Table 1, hierarchical supervision produces prediction sets closer to the ground-truth perceptual density than consensus-based single-label predictions. The *Primary + Minor + APE* configuration increases prediction density to 3.57 and achieves the highest Soft Recall (59.16%), indicating improved recovery of the ground-truth multi-annotator emotion set. Compared to the *Primary + APE* supervision, incorporating minority perception provides more balanced performance across primary fidelity and perceptual coverage.



(a) Rationale Quality Score Distribution



(b) Attribute Flag Rate by Rationale Type

Figure 3: Human evaluation results on rationale quality and attributes. (a) Q1: Distribution of overall rationale quality scores (0–5) across three training settings; mean scores are μ=4.43 (Primary+Minor), μ=3.21 (Primary+APE), and μ=3.39 (Primary+Minor+APE). (b) Q2: Attribute flag rates (%) across rationale properties related to perceptual reasoning.

Table 3 shows that, incorporating Minor Emotions (L_m) substantially improves recognition for rare categories such as Fear, Disgust, and Contempt, suggesting that minority annotations encode essential perceptual evidence for subtle affective states. By modeling these signals through hierarchical supervision and rationale distillation, our framework mitigates majority-class bias and enables SLLM to better capture the full spectrum of human emotional perception. Interestingly, the *Primary + Minor + APE* configuration does not consistently achieve the best performance across all metrics. While hierarchical supervision improves distribution-level recall, the *Primary + Minor* configuration achieves stronger primary fidelity and set-level alignment. This result suggests that our strategy to incorporate APE supervision via text rationale SFT may not fully exploit structured perceptual information and may introduce low-confidence perceptual signals when treated uniformly. Other studies have shown that secondary emotions are useful to characterize emotional behaviors (Harvill et al., 2023, 2019; Lotfian and Busso, 2018a; Chou et al., 2024a, 2025, 2022a). Therefore, we left a future work to identify more effective strategies to use secondary emotions to improve emotional reasoning in SLLM.

4.5 Human Evaluation on Rationales

We conduct a controlled human evaluation to assess the perceptual quality and faithfulness of generated rationales under three supervision settings: Primary + Minor, Primary + APE, and Primary + Minor + APE. Ten annotators evaluate 30 samples (10 per setting), yielding 300 total judgments. To examine whether rationale quality depends on prediction correctness, we deliberately balance *ground-truth* (GT) aligned and misaligned cases. Samples are grouped based on whether the predicted Primary emotion matches the GT label (16 aligned, 14 misaligned), allowing us to compare rationale quality under correct and incorrect prediction conditions.

For each sample, annotators provide (i) an overall rationale quality score on a 1–5 ordinal scale, and (ii) binary attribute annotations indicating whether the rationale satisfies six perceptual criteria: *ambiguity explanation* (whether the rationale convincingly explains why multiple emotional interpretations may arise), *audio grounding* (whether the explanation cites acoustic evidence consistent with the perceived speech signal), *primary emotion correctness* (whether the predicted primary emotion aligns with the annotator's own perception), *emotion coverage* (whether the predicted emotion set captures the annotator's perceived emotions), *hallucinated details* (whether the rationale introduces unsupported or unverifiable evidence), and *redundancy* (whether the explanation contains repetitive or unnecessary content). Figure 3 reports the quality score distribution and the attribute selection rate (the proportion of annotators endorsing a given attribute) under each supervision setting.

Inter-Evaluator Agreement. To ensure reliability of subjective judgments, we compute inter-evaluator agreement on the full set of 300 judgments. We use Krippendorff's α (ordinal) for the 1–5 quality score and Fleiss' κ for each binary attribute flag. As shown in Table 4, the overall quality score achieves $\alpha = 0.66$, indicating substantial agreement for a subjective generative-text evaluation. Objective attributes such as *primary correct* ($\kappa = 0.90$) and *audio-grounded* ($\kappa = 0.72$) yield high agreement, while more subjective attributes such as *redundant* ($\kappa = 0.46$) and *hallucinated details* ($\kappa = 0.50$) show moderate agreement, consistent with the inherent subjectivity of rationale-style judgments.

Table 4: Inter-evaluator agreement on 300 human judgments (30 samples $\times$ 10 annotators). Krippendorff's α (ordinal) is reported for the 1–5 quality score; Fleiss' κ is reported for each binary attribute flag.

Criterion	**Agreement**
Overall quality (1–5, ordinal)	$\alpha = 0.66$
Primary correct	$\kappa = 0.90$
Audio-grounded	$\kappa = 0.72$
Emotion coverage	$\kappa = 0.64$
Explains ambiguity	$\kappa = 0.57$
Hallucinated details	$\kappa = 0.50$
Redundant	$\kappa = 0.46$

Quantitative and Attribute Analysis. Human evaluation reveals clear trade-offs across supervision strategies. The Primary + Minor setting achieves the highest overall rationale quality ($\mu = 4.41$) and the strongest performance in explaining ambiguity (82%) and audio evidence grounding (88%), while exhibiting the lowest hallucination rate (6%). In contrast, the Primary + Minor + APE setting achieves the highest emotion coverage (87%) but shows substantially higher hallucination (47%) and redundancy (58%). Hallucination typically manifests as generating transcript content not present in the input audio or providing emotional interpretations inconsistent with evaluators' perception. The Primary + APE setting leads to lower scores overall ($\mu = 3.25$), suggesting that Primary + APE supervision lacks sufficient hierarchical structure for coherent perceptual reasoning.

GT Alignment Analysis. To examine whether rationale quality depends on prediction correctness, we compare rationale quality scores between GT-aligned and GT-misaligned groups. As shown in Table 5, GT-aligned samples ($n = 16$) achieve a mean score of 3.78 (60% ≥ 4), while GT-misaligned samples ($n = 14$) achieve a comparable mean score of 3.61 (53% ≥ 4). These results indicate that evaluators prioritize internal coherence and evidence grounding over strict label correctness. Even when the predicted emotion does not match the ground truth, logically consistent and evidence-grounded rationales are still rated highly, supporting the effectiveness of our hierarchical rationale distillation framework.

Table 5: GT Alignment Analysis of rationale quality.

	GT-Aligned ($n = 16$)	GT-Misaligned ($n = 14$)
Mean Score	3.78	3.61
Score ≥ 4	96/160 (60%)	74/140 (53%)

5 Conclusions

In this work, we showed that minority emotion annotations provide structured perceptual signals that enhance both prediction and reasoning quality in SER. We propose a hierarchical supervision framework that integrates Primary, Minor, and APE labels with rationale distillation using SLLM. Experiments on the MSP-Podcast corpus demonstrate improved primary emotion recognition, stronger perceptual coverage, and more evidence-grounded rationales. Human evaluation further confirms that minority-aware supervision enables more coherent and ambiguity-aware emotional reasoning. Overall, our findings highlight the importance of modeling perceptual disagreement to advance ambiguity-aware emotion reasoning.

Limitations

Our work has several limitations. First, although hierarchical minority-aware supervision improves both emotion prediction and ambiguity modelling, the formulation still relies on predefined taxonomies and annotation protocols from MSP-Podcast. While Primary, Minor, and APE labels capture richer perceptual structures than majority-vote labels, they remain tied to a closed emotion inventory and may not fully reflect the open-ended, context-dependent nature of affective perception. Second, our rationale generation depends on teacher explanations from Qwen3-Omni. Although the teacher is conditioned on both speech and transcripts, generated rationales may still contain hallucinated or weakly grounded evidence. Human evaluation suggests improved coherence and perceptual alignment, but we do not explicitly measure rationale faithfulness through intervention-based or causal evaluation. Third, our framework models ambiguity through hierarchical label structures but does not actively perform evidence acquisition or interactive reasoning at inference; rationales are produced via supervised distillation rather than dynamic evidence-seeking. Integrating agentic reasoning and tool-based verification could better capture uncertainty and perceptual disagreement. Finally, experiments are conducted primarily on MSP-Podcast. Although it provides rich multi-annotator annotations, further validation across datasets, languages, and conversational settings is needed to assess generalizability.

References

C. Busso, R. Lotfian, K. Sridhar, A.N. Salman, W.-C. Lin, L. Goncalves, S. Parthasarathy, A. Reddy Naini, S.-G. Leem, L. Martinez-Lucas, H.-C. Chou, and P. Mote. 2026. The MSP-Podcast corpus. *IEEE Transactions on Affective Computing*, early access.

H.-C. Chou, L. Goncalves, S.-G. Leem, A.N. Salman, C.-C. Lee, and C. Busso. 2025. Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule. *IEEE Transactions on Affective Computing*, 16(1):41–55.

H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. Busso. 2022a. Exploiting annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, pages 7717–7721, Singapore.

H.-C. Chou, H. Wu, L. Goncalves, S.-G. Leem, A. N. Salman, C. Busso, H.-Y Lee, and C.-C. Lee. 2024a. Embracing ambiguity and subjectivity using the all-inclusive aggregation rule for evaluating multi-label speech emotion recognition systems. In *IEEE Spoken Language Technology Workshop (SLT 2024)*, pages 502–509, Macao, China.

Hsin-Chieh Chou, Chung-Cheng Lee, and Carlos Busso. 2022b. Exploiting Co-occurrence Frequency of Emotions in Perceptual Evaluations To Train A Speech Emotion Classifier. In *Proc. Interspeech 2022*, pages 161–165.

Huang-Cheng Chou and Chi-Chun Lee. 2025. Revisiting Modeling and Evaluation Approaches in Speech Emotion Recognition: Considering Subjectivity of Annotators and Ambiguity of Emotions. *Preprint*, arXiv:2510.05934.

Huang-Cheng Chou, Haibin Wu, Lucas Goncalves, Seong-Gyun Leem, Ali Salman, Carlos Busso, Hung-Yi Lee, and Chi-Chun Lee. 2024b. Embracing ambiguity and subjectivity using the all-inclusive aggregation rule for evaluating multi-label speech emotion recognition systems. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 502–509.

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. 2024. Qwen2-Audio Technical Report. *Preprint*, arXiv:2407.10759.

Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in Real-Life Emotion Annotation and Machine Learning Based Detection. *Neural Networks*, 18(4):407–422.

Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon. 2016. Modeling Subjectiveness in Emotion Recognition with Deep Neural Networks: Ensembles vs Soft Labels. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 566–570.

J. Harvill, M. AbdelWahab, R. Lotfian, and C. Busso. 2019. Retrieving speech samples with similar emotional content using a triplet loss function. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, pages 7400–7404, Brighton, UK.

J. Harvill, S.-G. Leem, M. AbdelWahab, R. Lotfian, and C. Busso. 2023. Quantifying emotional similarity in speech. *IEEE Transactions on Affective Computing*, 14(2):1376–1390.

Xin Hong, Yuan Gong, Vidhyasaharan Sethu, and Ting Dang. 2025. Aer-llm: Ambiguity-aware emotion recognition leveraging large language models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *Preprint*, arXiv:2305.02301.

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

E. Jordan, R. Terrisse, V. Lucarini, M. Alrahabi, M. Krebs, J. Desclés, and C. Lemey. 2025. Speech emotion recognition in mental health: Systematic review of voice-based applications. *JMIR Mental Health*, 12(1):e74260.

Yujin Kang and Yong-Seok Cho. 2025. Beyond Single Emotion: Multi-label Approach to Conversational Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24321–24329.

Zhengqin Lai, Zhilin Zhu, Xiaopeng Hong, and Yaowei Wang. 2025. Agent-MER: A Cognitive Agent with Hierarchical Deliberation for Open-Vocabulary Multimodal Emotion Recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13864–13871.

Jeff T. Larsen, A. Peter McGraw, and John T. Cacioppo. 2001. Can People Feel Happy and Sad at the Same Time? *Journal of Personality and Social Psychology*, 81(4):684–696.

Pengcheng Li, Botao Zhao, Zuheng Kang, Junqing Peng, Xiaoyang Qu, Yayun He, and Jianzong Wang. 2025. Emo-rl: Emotion-rule-based reinforcement learning enhanced audio-language model for generalized speech emotion recognition. *Preprint*, arXiv:2509.15654.

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. 2025a. AffectGPT: A New Dataset, Model, and Benchmark

for Emotion Understanding with Multimodal Large Language Models. *Preprint*, arXiv:2501.16566.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. 2025b. Ovmer: Towards open-vocabulary multimodal emotion recognition. *Preprint*, arXiv:2410.01495.

R. Lotfian and C. Busso. 2017. Formulating emotion perception as a probabilistic model with application to categorical emotion classification. In *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, pages 415–420, San Antonio, TX, USA.

R. Lotfian and C. Busso. 2018a. Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning. In *Interspeech 2018*, pages 951–955, Hyderabad, India.

R. Lotfian and C. Busso. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Reza Lotfian and Carlos Busso. 2018b. Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning. In *Proc. Interspeech 2018*, pages 951–955.

Jialong Mai, Xiaofen Xing, Yangbiao Li, and Xiangmin Xu. 2025. Chain-of-thought distillation with fine-grained acoustic cues for speech emotion recognition. In *Proc. Interspeech 2025*. ISCA.

Robert Plutchik. 2001. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth Narayanan. 2019. The Ambiguous World of Emotion Representation. *Preprint*, arXiv:1909.00360.

Bo-Hao Su, Hui-Ying Shih, Jinchuan Tian, Jiatong Shi, Chi-Chun Lee, Carlos Busso, and Shinji Watanabe. 2026. Reasoning beyond majority vote: An explainable speechlm framework for speech emotion recognition. *Preprint*, arXiv:2509.24187.

Esther Sun, Bo-Hao Su, Abinay Reddy Naini, Shinji Watanabe, and Carlos Busso. 2026. Adept: Rl-aligned agentic decoding of emotion via evidence probing tools – from consensus learning to ambiguity-driven emotion reasoning. *Preprint*, arXiv:2602.12714.

James Tavernor and Emily Mower Provost. 2025. More Similar than Dissimilar: Modeling Annotators for Cross-Corpus Speech Emotion Recognition. *Preprint*, arXiv:2509.12295.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. 2026. Emotion-thinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. *Preprint*, arXiv:2601.15668.

Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. 2021. A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814.

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