EMO-RL: Emotion-Rule-Based Reinforcement Learning Enhanced Audio-Language Model for Generalized Speech Emotion Recognition

Pengcheng Li[1,2†], Botao Zhao[1†], Zuheng Kang[1], Junqing Peng[1], Xiaoyang Qu[1], Yayun He[1], Jianzong Wang[1*]
[1]Ping An Technology (Shenzhen) Co., Ltd., Shenzhen, China,
[2]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China,

Abstract

Although Large Audio-Language Models (LALMs) have exhibited outstanding performance in auditory understanding, their performance in affective computing scenarios, particularly in emotion recognition, reasoning, and subtle sentiment differentiation, remains suboptimal. Recent advances in Reinforcement Learning (RL) have shown promise in improving LALMs' reasoning abilities. However, two critical challenges hinder the direct application of RL techniques to Speech Emotion Recognition (SER) tasks: (1) convergence instability caused by ambiguous emotional boundaries and (2) limited reasoning ability when using relatively small models (e.g., 7B-parameter architectures). To overcome these limitations, we introduce EMO-RL, a novel framework incorporating reinforcement learning with two key innovations: Emotion Similarity-Weighted Reward (ESWR) and Explicit Structured Reasoning (ESR). Built upon pretrained LALMs, our method employs group-relative policy optimization with emotion constraints. Comprehensive experiments demonstrate that our EMO-RL training strategies can significantly enhance the emotional reasoning capabilities of LALMs, attaining state-of-the-art results on both the MELD and IEMOCAP datasets, and cross-dataset experiments prove the strong superiority of generalization.

1 Introduction

Speech Emotion Recognition (SER) is a significant research direction in the field of affective computing, aiming to map speech signals to corresponding emotional labels through computational analysis. It plays an important role in various applications, including intelligent customer service (Li and Lin, 2021), mental health assessment (Madanian et al., 2022), and human-computer interaction (Alsabhan, 2023).

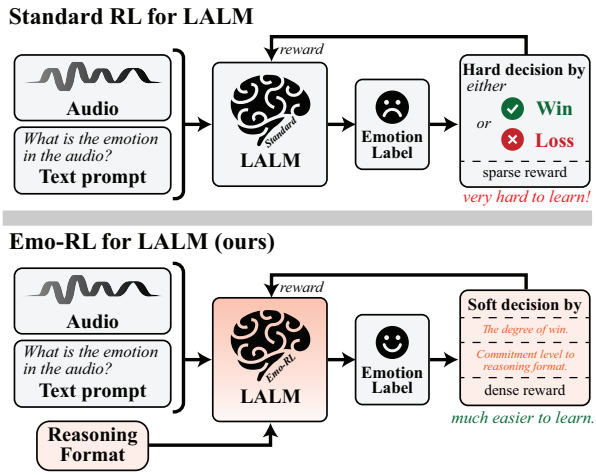

Figure 1: The key ideas of our proposed Emo-RL. Compared with the standard RL, Emo-RL exploited emotion similarity-weighted reward and the explicit structured reasoning to improve the emotion recognition performance of LALM.

In SER task, prior studies predominantly rely on pre-trained speech models or perform fine-tuning on affective corpora to derive emotional representations, then train a classification head to implement emotion classification (Li et al., 2023b; Chen et al., 2023). However, the emotional representations extracted by these models can only capture the acoustic expressions of emotions, but lack a collaborative analysis of text semantics. These models have very limited generalization capability and lack explainability.

With the development of multi-modal large models, many powerful Large Audio-Language Models (LALMs) have emerged (Kong et al., 2024), among which Qwen2-Audio (Chu et al., 2024) is a representative example. It can follow user instructions to perform many downstream tasks, such as speech recognition, transcription, sound classification, and more. Although Qwen2-Audio demonstrates strong speech understanding capabilities, its performance on SER tasks remains limited. This

† These authors contributed equally to this work.
* Corresponding author: jzwang@188.com.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18744–18754
November 4-9, 2025

limitation stems from its tendency to rely on shallow associations rather than multi-step reasoning that integrates textual semantics and auditory features across modalities, resulting in its limited accuracy and generalization on SER. Speech emotion recognition inherently constitutes a cognitive reasoning process that necessitates comprehensive analysis from multiple perspectives through step-by-step reasoning. When humans recognize emotions in speech, they often understand the specific content and keywords of speech and integrate this understanding based on acoustic features (such as pitch, voice quality, speech rate). For example, when someone says "fed up" with rapid speech, high volume, and sharp intonation, anger can be inferred.

These reasoning steps are beyond the capability of traditional audio feature extraction and classification head frameworks. Extensive research has shown that reinforcement learning can enhance the reasoning capabilities of LLM (Guo et al., 2025; Team et al., 2025). The effective deployment of reinforcement learning (RL) in SER encounters two fundamental limitations: convergence reliability issues primarily arising from ill-defined inter-class affective boundaries that induce gradient conflict during policy updates, compounded by insufficient affective reasoning capacity in under-parameterized architectures (e.g., 7B-parameter configurations).

To address these challenges, we adopt a psychological perspective by transforming the original right/wrong classification problem into a regression problem that accommodates varying degrees of correctness and error, through the introduction of an emotion-state-transition matrix (As illustrated in Figure 1). We implement an Emotion Similarity-Weighted Reward (ESWR) mechanism that progressively guides the policy model from simpler to more complex tasks. This approach initially teaches the model to distinguish between basic positive and negative emotions before advancing to finer-grained emotional distinctions. To further enhance the model's emotional reasoning capabilities, we incorporate Explicit Structured Reasoning (ESR) strategies during RL training. These strategies provide the model with guiding clues to help it more effectively differentiate between emotions, thereby improving its overall reasoning ability in SER tasks.

Based on the ESWR and ESR, we exploited our emotion-rule-based RL method to fine-tune the LALM, and the contributions of this paper are summarized as follows:

- We propose a SER pipeline via RL fine-tuning of a large audio and language model.
- We introduce Emotion-rule based RL to improve the emotion recognition ability of LALM, leveraging emotion similarity-weighted rewards and explicit structured reasoning strategies.
- Extensive experiments demonstrate that the proposed approach exhibits strong generalizability and achieves state-of-the-art performance.

2 Related Works

2.1 Generalized Speech Emotion Recognition

For SER, traditional approaches have focused on designing novel network architectures (Zou et al., 2022; Li et al., 2023b) based on classical neural networks. With the advancement of self-supervised learning, researchers have increasingly utilized pre-trained audio models like WavLM (Chen et al., 2022), Emotion2vec (Ma et al., 2024b), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2023) to extract speech features or fine-tune these models on speech emotion datasets to obtain emotion-specific features (Morais et al., 2022; Chen and Rudnicky, 2023). Subsequently, a linear classification head is trained to perform emotion classification. These models have significantly enhanced speech emotion perception capabilities (Li et al., 2023a). For instance, the Vesper model (Chen et al., 2024), obtained by distilling the WavLM-large model with emotion data, has achieved promising results in SER tasks. However, the generalization capabilities of these models remain limited, and they lack collaborative analysis of text semantics and explainability.

2.2 Large Audio-Language Models

Recent progress in multimodal large-scale language modeling has led to the emergence of numerous LALMs, such as Audio Flamingo (Kong et al., 2024), Qwen2-Audio (Chu et al., 2024),and SALMONN (Tang et al., 2023), which have demonstrated strong audio understanding capabilities, with Qwen2-Audio even outperforming previous methods across the vast majority of audio-focused evaluation benchmark. These models typically comprise three main components: an Audio Encoder, a Large Language Model, and a modality connector that bridges them. These models are

capable of directly processing cross-modal inputs, including audio (such as speech, environmental sounds, and music) and text prompts, and can generate the corresponding textual output. They are able to follow user instructions to perform a variety of downstream tasks, such as transcription, SER, and sound classification (Wang et al., 2025; Waheed et al., 2024). However, current LALM training mainly focuses on perception and basic QA tasks, lacking explicit multi-step reasoning. Thus, the potential of LALMs like Qwen2-Audio in complex audio reasoning tasks such as SER remains untapped. Enhancing their reasoning abilities in these advanced tasks is crucial.

2.3 Reinforcement Learning and Reasoning

Reinforcement learning (RL) plays a crucial role in advancing the reasoning abilities of LLMs and MLLMs. RLHF employs proximal policy optimization (PPO)(Schulman et al., 2017) alongside a trained reward mechanism to align LLMs with human preferences. Direct Preference Optimization (DPO)(Rafailov et al., 2023) bypasses reward modeling by learning from preference data directly, whereas Rejection Sampling Fine-tuning (RFT)(Yuan et al., 2023) strengthens reasoning through curated self-produced reasoning chains. Group Relative Policy Optimization (GRPO)(Shao, 2024) refines PPO by eliminating the critic component and utilizing group-level baseline averaging for advantage computation, achieving enhanced LLM reasoning with reduced computational overhead. The Hybrid GRPO variant (Sane, 2025) integrates GRPO's sampling mechanism with a trained value estimator, improving training stability and data utilization efficiency. Contemporary research demonstrates that Chain-of-Thought (COT) combined with RL substantially elevates LALM reasoning capabilities. Audio-CoT (Ma et al., 2025) pioneered COT integration in LALMs, though gains were modest without model parameter optimization. Audio-Reasoner (Xie et al., 2025) developed CoTA, an extensive synthetic corpus containing millions of question-answer instances with detailed reasoning trajectories, markedly advancing extended-context reasoning abilities. Xiaomi's implementation utilized GRPO optimization on the Qwen2-Audio-7B architecture for audio question-answering applications (Li et al., 2025), achieving notable improvements in reasoning precision. SARI (Wen et al., 2025) additionally combines systematic reasoning frameworks with progressive reinforcement training curricula, establishing new benchmarks on MMAU and MMSU evaluations. Reward-based optimization frameworks have proven effective in boosting reasoning precision, demonstrating that reinforcement-driven strategies can maximize learning efficiency from constrained training datasets. Nevertheless, existing RL methodologies remain overly broad for specialized speech emotion applications. Consequently, developing emotion-rule-guided reinforcement strategies specifically tailored for SER tasks becomes essential.

3 Methodology

3.1 Problem Definition

We use emotional audio question answering in Qwen2-Audio-7B-Instruct to implement SER. SER process through LALMs constitutes a parametric mapping process where: given a speech signal x and structured textual query Q containing multiple-choice options, their temporal-contextual concatenation forms the input prompt $p = [x; Q]$. The LALM, $\pi_{\boldsymbol{\theta}}$, then generates emotion prediction $\hat{y}$ through cross-modal understanding, formally expressed as:

$$\pi_{\boldsymbol{\theta}}(x; Q) \rightarrow a \rightarrow \hat{y}, \tag{1}$$

where S is a speech audio with a sampling rate of 16kHz, Q is the textual question prompt, and a is the generated response of LALM, including thinking and reasoning contents and the final selected answer, and $\hat{y}$ denotes the predicted emotion label.

This study aim to address two core challenges in SER through LALMs: (1) Enhancing the predictive accuracy of f_θ via reinforcement learning using the training dataset $\mathcal{D} = \{(x_i, y_i)\}_i^N$, where the N means the sample number. (2) Discovering optimal prompt formulations Q to improve the inference performance. Considering the parameter space of is infinite, we defined three experimentally validated designs as the Q space, $\mathbf{Q}$, including implicit reasoning Q_{IR}, explicit unstructured reasoning Q_{EUR}, explicit structured reasoning Q_{ESR}. Therefore, the target of this study could be defined as:

$$\theta, Q = \arg\max_{\theta, Q \in \mathbf{Q}} (\mathcal{R}(\pi_{\boldsymbol{\theta}}(X; Q), Y)), \tag{2}$$

where the R denotes the reward function.

In detail, we explore three reasoning strategies in EMO-RL training to evaluate the impact of reasoning patterns. We detail three patterns below:

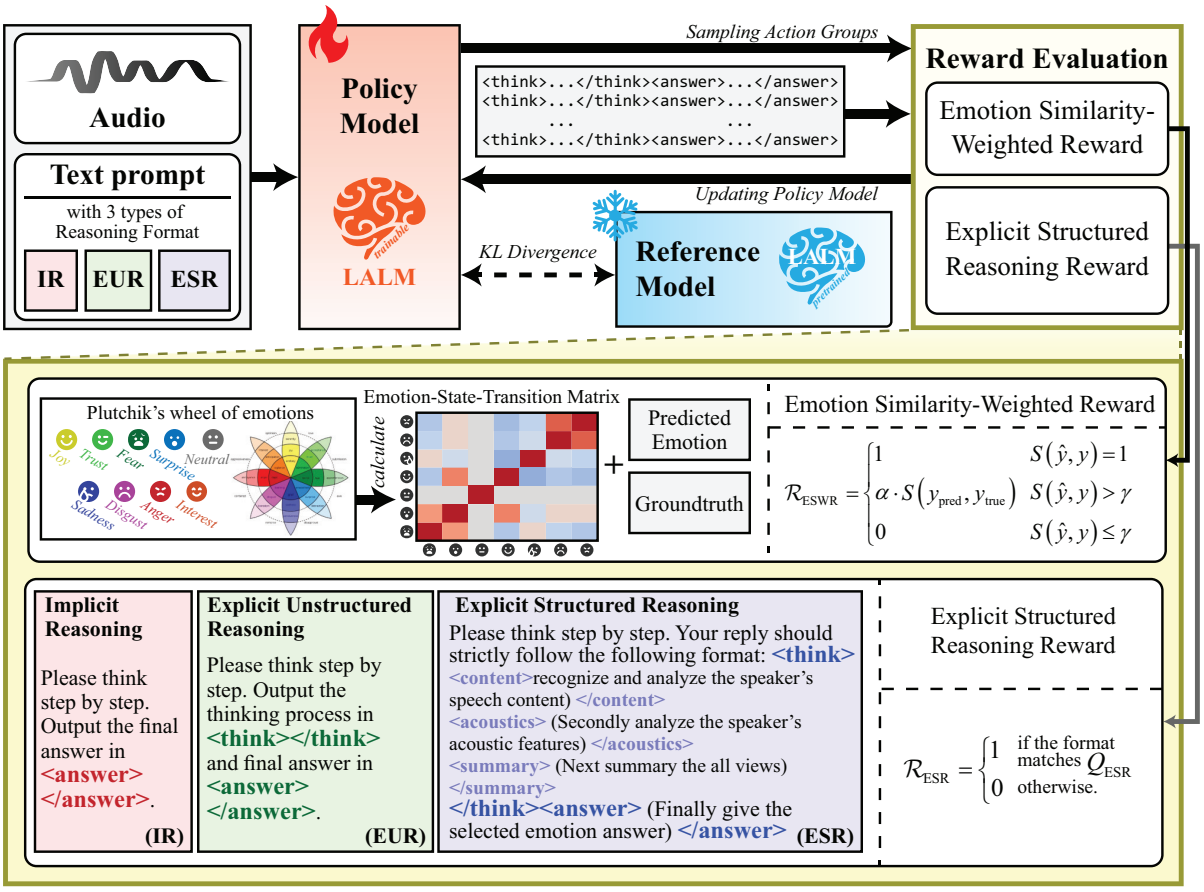

Figure 2: The overview of the Emo-RL for LALM to improve the generalized speech emotion recognition. Building on GRPO, we enhance emotion recognition via two improvements. First, we create an emotion-state-transition matrix from Plutchik's wheel of emotions (Plutchik, 1982), allowing the policy model to receive rewards for predicting similar emotions. Second, we introduce explicit structured reasoning to directly input human emotion recognition priors into the model.

- Implicit Reasoning, Q_{IR}: The foundational configuration involves training the system to directly produce conclusive responses, bypassing any obligation to articulate underlying cognitive mechanisms or intermediate analytical steps.

- Explicit Unstructured Reasoning, Q_{EUR}: This approach facilitates organic and unconstrained thought expression by employing prompting techniques that eschew rigid organizational templates or prescribed divisions. While permitting flexible formulation, the framework necessitates generation of logically consistent interpretations culminating in unambiguous determinations.

- Explicit Structured Reasoning, Q_{ESR}: The methodology enforces systematic generation of transparently organized cognitive pathways. Implementation requires adherence to a bifurcated analytical framework encompassing textual dimensions (verbatim transcriptions, pivotal terminology) and prosodic characteristics (intonation patterns, temporal cadence, acoustic intensity, vocal texture). Through synthesis of these dual information streams, the system derives its conclusive assessment.

3.2 Emotion-rule based RL framework

We built our Emo-RL based on the GRPO framework for its efficiency and scalability. Unlike proximal policy optimization, which requires a computationally expensive value network, GRPO calculates relative advantages by comparing rewards within a group of sampled actions, reducing computational overhead and simplifying optimization. This makes GRPO particularly suitable for speech reasoning tasks. Similar to GRPO, the Emo-RL also has three main steps, sampling action groups, reward evaluation, and updating policy network with relative advantage and KL divergence (As shown in Figure 2).

Sampling Action Groups For each input state $s = (x, Q)$, where x is the speech encoding of the input audio and Q the textual encoding of the

question, GRPO samples a group of actions (the generated response of LALM), $\{a_1, a_2, \ldots, a_G\}$, from the current policy $\pi_{\boldsymbol{\theta}}$. The sampling process is:

$$a_i \sim \pi_\theta(a \mid x, Q), \quad \text{for} \quad i = 1, 2, \ldots, G. \quad (3)$$

This strategy ensures diverse responses, promoting exploration and preventing premature convergence.

Reward Evaluation. In our reinforcement learning framework, each sampled action a_i is assigned a reward $\mathcal{R}(a_i)$ based on verifiable criteria, resulting in a reward set $\{r_1, r_2, \ldots, r_G\}$. For emotional speech reasoning tasks, the reward function $\mathcal{R}(a_i)$ combines two components: the reasoning format reward $\mathcal{R}_{\text{format}}(a_i)$ and emotion accuracy reward $\mathcal{R}_{\text{acc}}(a_i)$. The format reward ensures that the responses adhere to a structured format, thereby guiding the reasoning strategy of the policy network, π_θ. The accuracy reward evaluates the correctness of the action a_i, providing feedback to π_θ on the extent to which the answer aligns with the correct response. The overall reward function is:

$$\mathcal{R}(a_i) = \mathcal{R}_{\text{format}}(a_i) + \mathcal{R}_{\text{acc}}(a_i). \quad (4)$$

Updating Policy Network with Relative Advantage and KL divergence. The π_θ is optimized by the Relative Advantage of rewards and KL divergence between π_θ and reference model π_{ref}. Firstly, policy Rewards are normalized within the sampled group to compute relative advantages $\{A_1, A_2, \ldots, A_G\}$, defined as:

$$A_i = \frac{r_i - \text{mean}\{r_1, r_2, \ldots, r_G\}}{\text{std}\{r_1, r_2, \ldots, r_G\}}. \quad (5)$$

Based on these advantages, the policy is updated to reinforce actions with positive advantages and reduce the probability of less effective ones. To ensure stable RL learning, π_θ updates are further constrained by minimizing the KL divergence between the updated and reference models.

3.3 Rewards Mechanism Design

The EMO-RL framework implements dual reward mechanisms synergistically combining structural compliance enforcement and affective alignment optimization. Specifically, domain-specific response schemata are enforced through regular-expression pattern matching that validates three distinct reasoning pattern compliance rates ($Q_{\text{IR}}, Q_{\text{EUR}}, Q_{\text{ESR}}$), systematically enhancing explainability via cognitive transparency in decision pathways. Complementarily, the emotion similarity-weighted reward employs an emotion-state-transition matrix constructed through Plutchik's wheel of emotions (Plutchik, 1982), generating dense reward signals that precisely guide policy gradients through convex optimization landscapes.

3.3.1 Reasoning Format Reward

This component ensures adherence to specific response formats across different reasoning strategies by implementing tailored format reward. We define three distinct reasoning format functions, including Implicit Reasoning (IR), Explicit Unstructured Reasoning (EUR), and Explicit Structured Reasoning (ESR), each requiring different format constraints.

For IR, which targets only answer generation, the reward is granted only when the final answer is both correct and correctly delimited by `<answer>` tags, as shown in Figure 2. For EUR, which require explicit reasoning display, the format reward is granted when the response contains reasoning within `<think>` tags and the final answer within `<answer>` tags. ESR is similar to EUR, but with four additional format constraint tags.

The format reward function of ESR is as follows, and all format reward follow this rule. Each format reward function employs binary scoring based on regex pattern matching, where strict adherence to the specified format yields a reward score of 1, while any deviation results in a score of 0. This ensures consistent formatting across different reasoning strategies while maintaining the flexibility to accommodate strategy-specific requirements.

$$\mathcal{R}_{\text{ESR}} = \begin{cases} 1, & \text{if the format matches } Q_{\text{ESR}} \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

3.3.2 Emotion Accuracy Reward

The conventional approach uses binary classification rewards (BCR), allocating a score of 1 to fully accurate responses and 0 to all others. However, this has limitations when applied to emotions, as it ignores the relationships between different emotion types. Emotions are inherently continuous and complex. Drawing from psychological emotion dimension theories (Plutchik, 1982) and other psychological knowledge, we comprehensively consider emotion valence (the positive or negative of emotions) and arousal (the intensity or activation level of emotions). We have meticu-

lously designed an emotion-state-transition matrix $S \in \mathbb{R}^{C \times C}$ (C denote emotion categories) as:

$$S_{i,j} = \begin{cases} \frac{1}{2}, & \text{if } y_i \text{ or } y_i\text{='neutral'} \\ \frac{1}{2}\left(\cos\left(\mathrm{Pl}\left(y_i, y_j\right)\right) + 1\right), & \text{otherwise} \end{cases} \tag{7}$$

Here, the y_i and y_j denotes two emotion types, and $\mathrm{Pl}(\cdot,\cdot)$ means the angles from each other on the Plutchik's wheel of emotions. Based on the S, we have the Emotion Similarity-Weighted Reward function, formulated as:

$$\mathcal{R}_{\mathrm{ESWR}} = \begin{cases} 1 & S(\hat{y}, y) = 1 \\ \alpha \cdot S(\hat{y}, y) & S(\hat{y}, y) > \gamma \\ 0 & S(\hat{y}, y) <= \gamma \end{cases} \tag{8}$$

where α is the partial matching coefficient, dynamically adjusting from 1 to 0 during training, and γ is the threshold of the contradictory emotion, which was set as 0.7 in this paper.

4 Experiment

4.1 Dataset

We evaluated model capabilities and generalization in speech emotion recognition (SER) using four datasets: **MELD** (Poria et al., 2019) (13 708 utterances from *Friends*, 7 emotions), **IEMOCAP** (Busso et al., 2008) (5,531 utterances from conversations, 4 emotions), **RAVDESS** (Livingstone and Russo, 2018) (4 800 audio-video recordings of speech and song, 8 emotions), and **SAVEE** (Jackson and Haq, 2014) (480 samples, 7 emotions).

4.2 Implementation Details

We use Qwen2-Audio-7B-instruct as the foundational backbone model for our experiments. The RL models are trained using eight NVIDIA RTX A6000 GPUs, each processing a per-device batch-size of 1 with gradient accumulation over 2 steps. Training proceeds for 300 optimisation steps under a learning rate of 1×10^{-6} and a softmax temperature of 1.0. Each reinforcement learning optimization step generates 6 responses per sample. SFT models are optimized with AdamW at a learning rate of 1×10^{-5} for five complete epochs. The optimal iteration results are selected for final analysis.

4.3 Baselines and Metrics

We benchmark the proposed approach against state-of-the-art methods, which we group into three distinct categories: **W/o-LALM**, **LALM**, and **LALM-FT**. W/o-LALM and LALM-FT refer to models post-trained on the MELD training set, while LALM involves zero-shot inference using prompt strategies without task-specific fine-tuning.

- **W/o-LALM**: We selected four advanced self-supervised pre-trained audio models: HuBERT large (Hsu et al., 2021), data2vec 2.0 large (Baevski et al., 2023), WavLM large (Chen et al., 2022), Whisper large v3 (Radford et al., 2023) and Emotion2vec (Ma et al., 2024b). Features from the last Transformer layer of these frozen pre-trained models were extracted to train the downstream linear layers with a hidden dimension of 256.

- **LALM**: We directly use Qwen2-Audio (Chu et al., 2024) for SER tasks without additional training or fine-tuning, employing two prompt patterns: direct inference and chain-of-thought inference.

- **LALM-FT**: We further trained Qwen2-Audio. To evaluate different training methods, we compare models trained with supervised fine-tuning (SFT), GRPO (Shao, 2024), and EMO-RL. Additionally, we assess the impact of different reasoning strategies in EMO-RL: implicit reasoning, unstructured explicit reasoning, and structured explicit reasoning.

In our evaluation, we utilize three key metrics: Unweighted Accuracy (UA), Weighted Accuracy (WA), and Macro F1 Score (F1), to assess the performance of the SER task. WA reflects the overall accuracy of the model across all emotion classes. UA measures the average accuracy by considering each emotion class equally, regardless of its frequency in the dataset. The macro-F1 score, harmonic mean of precision and recall, furnishes a balanced and class-agnostic gauge of model efficacy, particularly in scenarios where there is an imbalance in the distribution of emotion classes.

5 Results

5.1 Main Performance

As shown in Table 1, the results demonstrate the effectiveness of COT in LALM, our proposed ESWR, and ESR. **Effectiveness of COT in LALM**: Using CoT prompts significantly enhances the zero-shot SER performance of LALMs. In fact, CoT enables Qwen2-Audio to approach the performance

Table 1: The comparison of the main performance metrics for various methods on the MELD and IEMOCAP datasets. Results for W/o-LALM methods are cited from the Emobox benchmark (Ma et al., 2024a) and (Ma et al., 2024b). The **bold** font indicates the best results among all models. The baseline here denotes the GRPO+IR, and the SOTA means the best results among W/o-LALM methods.

Model Type	Model	Method	MELD			IEMOCAP		
			UA(%)	WA(%)	F1(%)	UA(%)	WA(%)	F1(%)
W/o-LALM	HuBERT large (Hsu et al., 2021)	Classification Head	24.13	46.37	24.99	67.42	66.69	67.24
	WavLM large (Chen et al., 2022)	Classification Head	28.18	49.31	29.11	69.47	69.07	69.29
	data2vec 2.0 large (Baevski et al., 2023)	Classification Head	26.33	47.72	27.35	57.30	56.23	56.70
	Whisper large V3 (Radford et al., 2023)	Classification Head	31.54	51.89	32.95	73.54	72.86	73.11
	Emotion2vec+ large (Ma et al., 2024b)	Classification Head	28.03	51.88	/	70.70	67.30	/
LALM	Qwen2-Audio (Chu et al., 2024)	Direct Inference	18.96	39.83	19.84	53.76	51.52	47.68
		CoT Inference	26.89	50.57	28.05	64.33	60.37	61.61
LALM-FT	Qwen2-Audio (Chu et al., 2024))	SFT + IR	33.26	57.39	35.77	85.70	83.87	84.53
		BCR + IR	31.60	55.41	33.22	81.74	80.00	80.71
		BCR + EUR	34.87	59.63	36.46	83.34	82.56	82.88
		BCR + ESR	34.43	60.78	36.28	84.74	83.96	84.26
		ESWR + IR	36.23	63.85	38.57	84.12	83.90	83.11
		ESWR + EUR	37.81	66.17	39.19	85.97	84.85	85.50
		ESWR + ESR	**39.46**	**69.56**	**41.87**	**87.42**	**87.28**	**87.40**
/	Comparasion	Ours VS SOTA	↑25.1	↑34.0	↑27.1	↑18.9	↑19.8	↑19.6
		Ours VS Baseline	↑24.9	↑25.5	↑26.0	↑6.95	↑10.9	↑10.8

of the best W/o-LALM pre-trained audio models without any task-specific post-training. **Effectiveness of ESWR**: When training and testing on the same dataset, the direct use of GRPO achieves similar accuracy to SFT. This may be due to the MELD dataset containing considerable noise, resulting in the model's lower ability to recognize correct emotions. This leads to GRPO's binary rewards being too sparse, with 60% of accuracy rewards being 0, making it difficult for the model's update policy to stabilize. However, ESWR provides more dense and psychologically grounded reward signals, improving the model's emotional reasoning capability by consistently guiding it toward the correct emotional direction.

The effectiveness of ESR training strategies. Besides the ESWR method, compared to models trained with IR, models trained with EUR and ESR can both enhance emotional reasoning capabilities, improving accuracy on the MELD and IEMOCAP test sets. Moreover, models with structured thinking capabilities achieve superior accuracy relative to models lacking structured reasoning mechanisms, indicating that structured reasoning helps models avoid errors. Through the above experiments, we have demonstrated that using the EMO-RL algorithm can significantly enhance the emotional reasoning capabilities of LALMs, achieving SOTA performance in SER tasks. Additionally, we found that our method yields greater improvements on datasets with more complex emotion labels, for example, the improvement of MELD, compared to IEMOCAP.

Table 2: Weighted Accuracy (WA, %) across RAVDESS, SAVEE, and IEMOCAP Datasets. The model was trained on the MELD training dataset. The baseline here denotes the SFT+IR.

Model	RAVDESS	SAVEE	IEMOCAP
W/o-LALM Baselines			
HuBERT large	25.02	31.54	44.60
WavLM large	33.90	34.10	48.59
data2vec 2.0 large	34.21	37.79	47.43
Whisper large v3	40.68	42.18	46.14
LALM-FT (Qwen2-Audio (Chu et al., 2024))			
SFT+ IR	59.83	71.52	82.74
BCR + IR	62.07	72.38	82.66
ESWR + IR	66.21	74.69	83.05
ESWR + EUR	70.43	78.57	86.11
ESWR + ESR	**73.99**	**80.83**	**87.86**
Ours VS Baseline	↑23.67	↑13.02	↑6.19

5.2 Generalizability

In practical scenarios, a model's ability to generalize emotion recognition to unseen individuals and unknown recording conditions is of paramount importance. To evaluate this capability, cross-dataset zero-shot testing offers an effective means of assessing a model's generalization in emotion recognition. We meticulously selected three diverse datasets: IEMOCAP, RAVDESS, and SAVEE. These datasets encompass a variety of sources, accents, and recording environments, enabling a comprehensive evaluation of the model's generalization

Table 3: Performance of quantitative ablation of the reward mechanism alone on MELD dataset. The model was trained based on Qwen2-Audio

Method	UA(%)	WA(%)	F1(%)
GRPO+IR	18.22	38.32	18.96
GRPO+EUR	25.55	49.12	26.36
GRPO+ESR	29.19	53.53	30.61
GRPO+BCR	33.62	55.42	35.57
GRPO+ESWR	35.02	62.93	37.73

and robustness across real-world scenarios.

As shown in Table 2, the results of cross-datasets evaluation demonstrate that **(1)** Reinforcement learning methods, including GRPO and ESWR, demonstrate superior generalization capabilities compared to SFT methods. Notably, ESWR exhibits better generalization than GRPO. Additionally, **(2)** Explicit Reasoning strategies show enhanced generalization over Implicit Reasoning, and Structured Reasoning strategies outperform their unstructured counterparts. In conclusion, the combination of ESWR and ESR surpasses all baseline and alternative training methods, achieving the highest performance in emotional reasoning generalization.

5.3 Ablation

To explore the quantitative ablation of the effects of the reward mechanism alone. As shown in Table 3, we have supplemented the relevant quantitative ablation experiments on MELD dataset. For the methods in the first three rows, we only used the corresponding format reward without the accuracy reward. For the last two rows, we only used the corresponding accuracy reward without the format reward. These results demonstrate the efficiency of our proposed ESWR and ESR methods for speech emotion recognition.

5.4 Case Study

In Figure 3, we show a case study that demonstrates the response results when testing the same speech sample after training with different methods. Models trained with IR seem to have lost many other abilities, such as not trying to think and reasoning, even though I asked them to do so.

Models trained with EUR can generate flexible reasoning based on different speech inputs. They often analyze emotions primarily through acoustic features such as pitch, rhythm, and speed. While this approach is effective for simpler cases, it faces challenges with more complex scenarios due to the omission of critical semantic emotional details.

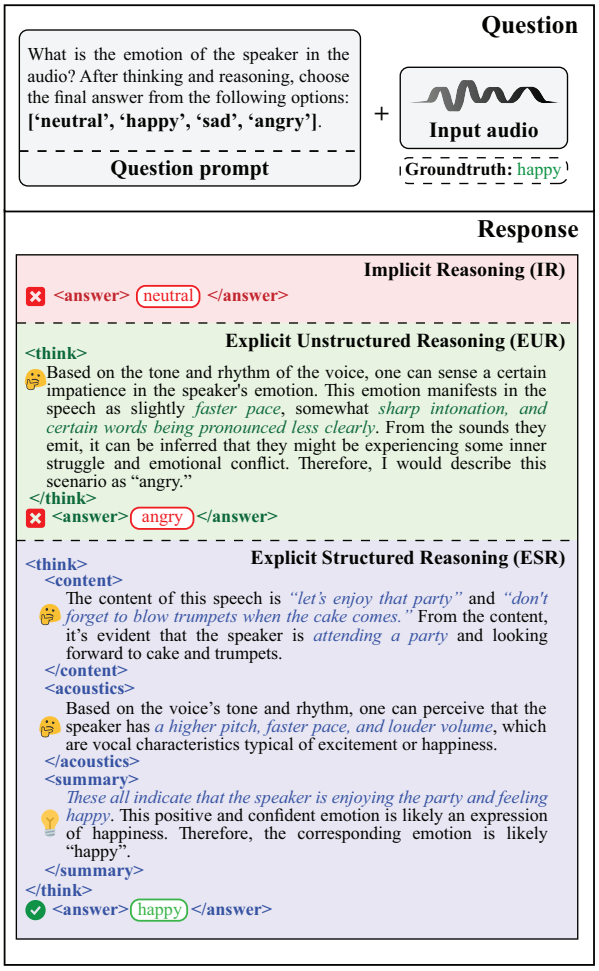

Figure 3: An example of the reasoning results of IR, EUR, and ESR

In contrast, models trained with ESR explicitly document the speaker's key content and acoustic features, followed by a comprehensive analysis of both semantic and auditory information. This structured approach reduces the likelihood of overlooking key details, thereby enhancing the model's emotional reasoning capabilities.

6 Conclusion

This paper introduces EMO-RL, a reinforcement learning framework that improves the emotional-reasoning capacity of large audio–language models for speech-emotion recognition. By incorporating emotion similarity-weighted reward, which integrates psychological prior knowledge into RL, and Explicit Structured Reasoning into our framework, EMO-RL effectively overcomes the challenges of convergence instability and limited reasoning ability in speech emotion recognition tasks. Comprehensive experiments demonstrate that EMO-RL not only improves the emotional reasoning capa-

bilities of LALMs on the MELD and IEMOCAP datasets (compared with SOTA, achieving an UA improvement of 25.1% and 18.9%, respectively), but also shows excellent generalization across different datasets. This work signifies a step forward in applying reinforcement learning and large audio-language models to speech emotion recognition, paving the way for future speech affective computing research. Moreover, EMO-RL shows potential for enhancing multi-modal LLMs' emotion perception, bringing us closer to building truly emotional LLMs.

7 Limitation

Our proposed method has certain limitations that warrant attention. Firstly, while our EMO-RL framework is designed to be versatile and applicable across a variety of multi-modal scenarios, including video, audio, and text, our current experimental scope has been limited to the speech modality alone. We have not yet incorporated visual elements such as images or videos into our experimental design. This restriction means that the full potential of our framework in multi-modal contexts remains unexplored. Secondly, although exploiting the LALMs for SER tasks has delivered promising results, it has also introduced challenges related to computational complexity and inference efficiency. The inference efficiency of our approach is comparatively lower than that of previous methods, which might affect its practicality for real-time applications. In the future, we will try to solve the above limitations.

8 Ethical Considerations

The deployment of SER systems raises significant ethical concerns that build upon established frameworks for sentiment and emotion analysis. Privacy and consent represent primary issues, as SER extracts sensitive psychological information from vocal patterns often without users' explicit awareness, unlike voluntary text-based sentiment analysis. Additionally, SER systems exhibit systematic biases across demographic groups and may misinterpret cultural differences in emotional expression, with training datasets often lacking diverse representation—problems shared with broader emotion analysis research. The "black box" nature of deep learning-based systems also raises accountability concerns when informing decisions affecting individuals' lives. These considerations highlight the need for robust consent frameworks, diverse datasets, and ethical guidelines specific to speech emotion recognition that address the unique challenges of speech emotion detection.

9 Acknowledgements

This work was supported by the Shenzhen-Hong Kong Joint Funding Project (Category A) under Grant No. SGDX20240115103359001.

References

Waleed Alsabhan. 2023. Human–computer interaction with a real-time speech emotion recognition with ensembling techniques 1d convolution neural network and attention. *Sensors*, 23(3):1386.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. 2023. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pages 1416–1429.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. I: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Li-Wei Chen and Alexander Rudnicky. 2023. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, and et. al. Xiao. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. 2024. Vesper: A compact and effective pretrained model for speech emotion recognition. *IEEE Transactions on Affective Computing*.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2023. Dst: Deformable speech transformer for emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, and et. al. Lin. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and et. al. Bi. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Philip Jackson and SJUoSG Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: a novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the International Conference on Machine Learning*, pages 25125–25148.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.

Xutong Li and Rongheng Lin. 2021. Speech emotion recognition for power customer service. In *International Conference on Computer and Communications*, pages 514–518.

Yuanchao Li, Yumnah Mohamied, Peter Bell, and Catherine Lai. 2023a. Exploration of a self-supervised speech model: A study on emotional corpora. In *IEEE Spoken Language Technology Workshop*, pages 868–875.

Zhipeng Li, Xiaofen Xing, Yuanbo Fang, Weibin Zhang, Hengsheng Fan, and Xiangmin Xu. 2023b. Multi-scale temporal transformer for speech emotion recognition. In *Interspeech*, pages 3652–3656.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Z Ma, M Chen, H Zhang, Z Zheng, W Chen, X Li, J Ye, X Chen, and T Hain. 2024a. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Interspeech*, pages 1580–1584.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024b. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics ACL*, pages 15747–15760.

Samaneh Madanian, David Parry, Olayinka Adeleye, Christian Poellabauer, Farhaan Mirza, Shilpa Mathew, and Sandy Schneider. 2022. Automatic speech emotion recognition using machine learning: digital transformation of mental health. In *Proceedings of the Annual Pacific Asia Conference on Information Systems*.

Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. 2022. Speech emotion recognition using self-supervised features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6922–6926.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.

Soham Sane. 2025. Hybrid group relative policy optimization: A multi-sample approach to enhancing policy optimization. *arXiv preprint arXiv:2502.01652*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Peiyi Zhu Qihao Xu Runxin Song Junxiao Bi et. al. Shao, Zhihong Wang. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. In *The International Conference on Learning Representations*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, and et. al. Liao. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Abdul Waheed, Hanin Atwany, Bhiksha Raj, and Rita Singh. 2024. What do speech foundation models not learn about speech? *arXiv preprint arXiv:2410.12948*.

Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, and et. al. Sun. 2025. Enabling auditory large language models for automatic speech quality evaluation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. 2025. Sari: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng. 2022. Speech emotion recognition with co-attention based multi-level acoustic information. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7367–7371.