MERIT: Multimodal Emotion Recognition via RL-Enhanced Test-Time Adaptation

Chen Zhang [1] **Wenqing Wu** [1] **Yingqiu Zhang** [1] **Peihong He** [1] **Ziyang Liu** [1 *]

Abstract

Multimodal Emotion Recognition (MER) seeks to understand human emotions by integrating information from textual, visual, and auditory modalities. While recent advances in Multimodal Large Language Models (MLLMs), such as Emotion-LLaMA, have demonstrated strong performance, they often struggle with distribution shifts and generalization to unseen test domains.

In this work, we propose a reinforcement learning-enhanced MER framework that integrates Test-Time Reinforcement Learning (TTRL) with a novel Majority Voting-based Verified Reward mechanism. By incorporating an emotion-aware reward function shaped by an emotional distance matrix, our method enables dynamic adaptation of lightweight LoRA adapters within a frozen Qwen-32B backbone, thereby enhancing both emotional consistency and generalization.

Extensive experiments on benchmark datasets, including CMU-MOSEI and IEMOCAP, show that our approach consistently outperforms strong baselines in terms of both accuracy and emotion consistency. Furthermore, ablation studies confirm the effectiveness of our soft reward design. These results underscore the potential of combining RL-based test-time adaptation with LLM-driven MER, offering a promising path toward more robust, adaptive, and emotionally intelligent AI systems.

1. Introduction

Multimodal Emotion Recognition (MER) seeks to comprehensively understand human emotions by integrating information from diverse modalities—including text, vision, and audio. This task plays a vital role in applications such as human-computer interaction, affective computing, and social robotics. Recently, Large Language Models (LLMs) have exhibited strong capabilities in semantic representation and cross-modal fusion, pushing the frontier of MER systems.

Early approaches to MER, including those by Baltrušaitis et al. (Baltrušaitis et al., 2018) and Zadeh et al. (Zadeh et al., 2017), primarily relied on handcrafted or early/late fusion strategies. These methods struggled to model the complex and dynamic relationships across modalities. Subsequent work introduced deep architectures such as RNNs, CNNs, and Transformers, which enhanced cross-modal representation learning (Tsai et al., 2019). More recently, multimodal large language models (MLLMs), such as Emotion-LLaMA (Cheng et al., 2024), have employed multimodal encoders and instruction tuning to significantly improve both emotion recognition and reasoning.

However, existing MLLM-based MER systems heavily depend on supervised fine-tuning, which limits their adaptability in low-resource settings or under distribution shifts. Reinforcement Learning (RL) has emerged as a powerful tool for enhancing LLMs in such scenarios. In particular, Test-Time Reinforcement Learning (TTRL) (Zuo et al., 2025) introduces a majority voting-based reward mechanism that enables self-adaptation on unlabeled data, offering an alternative to traditional supervised fine-tuning (SFT) frameworks.

Inspired by these developments, we propose an RL-enhanced training framework for MER that builds upon the Emotion-LLaMA architecture. Our method employs modality-specific encoders for vision, audio, and text, and projects the aligned cross-modal features into a Qwen-32B LLM backbone. Beyond initial SFT, we incorporate a Test-Time Reinforcement Learning (TTRL) mechanism guided by a structure-aware verified reward defined as

$$R_{mv}(y_i) = \frac{1}{1 + \alpha \cdot W_{a_i, a^*}}$$

where W encodes emotion distances between the predicted label a_i and the pseudo-gold label a^* derived from major-

*Corresponding author [1]School of Future Science and Engineering, Soochow University, Suzhou, China. Correspondence to: Ziyang Liu <ziyannn@yeah.net>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

ity voting. This soft reward design supports emotionally coherent and stable adaptation, and enables the model to further optimize its predictions using advanced RL algorithms, including Proximal Policy Optimization (PPO) and Group Relative Policy Optimization (GRPO). This design allows the model to better understand and adapt to emotionally rich and diverse inputs.

Our main contributions are summarized as follows:

- We introduce Test-Time Reinforcement Learning (TTRL) to MER and propose a verified reward mechanism based on majority voting to enable adaptive optimization on unlabeled data. This enhances the generalization and reasoning capabilities of LLMs in emotion recognition.
- We develop an optimized variant of the Emotion-LLaMA framework by integrating cross-modal alignment and RL-based adaptation within a Qwen-32B LLM backbone through lightweight LoRA adapters.
- Extensive experiments on benchmark datasets show that our RL-enhanced MER framework consistently outperforms strong baselines under distribution shift and test-time adaptation scenarios.

2. Related Work

Multimodal Emotion Recognition (MER). The task of multimodal emotion recognition (MER) was first formalized by Busso et al. (Busso et al., 2004), who demonstrated that combining facial and acoustic features through decision-level and feature-level fusion significantly enhanced the accuracy and robustness of emotion recognition systems. This early work established the foundation for exploring modality complementarity in affective computing.

Subsequent studies have introduced increasingly sophisticated architectures. Chen et al. proposed MemoCMT (Chen et al., 2023), which utilizes HuBERT and BERT to extract deep representations from audio and text, respectively. However, its dependence on heavyweight Transformer architectures incurs high computational costs and often suffers from generalization issues in emotionally complex scenarios.

TACFN (Li et al., 2022) improves cross-modal fusion by incorporating self-attention-based feature selection and dynamic weighting mechanisms. Despite these advantages, it still faces challenges related to temporal alignment and scalability across diverse modalities.

More recently, Emotion-LLaMA (Cheng et al., 2024) advanced the field by projecting audio, visual, and textual inputs into a shared latent space. Leveraging instruction-tuning on fine-grained emotion samples, it demonstrated strong performance on reasoning and classification tasks with an LLaMA-7B backbone.

Reinforcement learning (RL) has emerged as a promising paradigm for modeling dynamic and temporal aspects of emotion. Early approaches relied on value-based optimization methods, such as Q-learning (Watkins & Dayan, 1992), which model the state-action value function to capture transitions between emotional states. However, these methods struggle in high-dimensional multimodal environments due to instability in Q-value estimation.

To address these limitations, policy-based methods such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) have been introduced. PPO constrains policy updates to improve training stability and enhances robustness to noisy audio-visual inputs. Building on this, Shao et al. (Shao et al., 2024) proposed Group Relative Policy Optimization (GRPO), which models group-wise reward distributions to quantify uncertainty in emotional predictions, yielding better generalization in complex dialogue settings.

RL has also been applied to emotion recognition tasks through various architectural innovations. Zhang et al. (Zhang et al., 2024) proposed RL-EMO, a reinforcement learning framework combining Multi-modal Graph Convolution Networks (MMGCNs) with an RL module to model emotional-level context dependencies, achieving strong performance on the IEMOCAP and MELD datasets. Furthermore, Zhao et al. (Zhao et al., 2025b) introduced R1-Omni, which applies RL with verifiable reward (RLVR) to omni-modal LLMs, significantly improving reasoning, emotion recognition accuracy, and cross-domain generalization. Their work highlights the potential of RLVR not only for performance gains but also for enhancing interpretability in emotion modeling.

In addition, hybrid frameworks that integrate both value-based and policy-based strategies have shown promise. For example, RLVR introduces a verified reward formulation into LLM-based MER pipelines, improving interpretability, cross-modal generalization, and training efficiency. This line of work suggests that RL-guided optimization is a promising direction for enhancing MER performance, particularly under low-resource and distribution-shifted conditions.

3. Method

3.1. Overall Framework

Figure 1 illustrates the overall architecture of our proposed RL-enhanced MER framework. We first utilize modality-specific encoders to extract representations from text, audio, and visual inputs. These features are aligned and projected into a shared space, which is fed into a Qwen-32B large

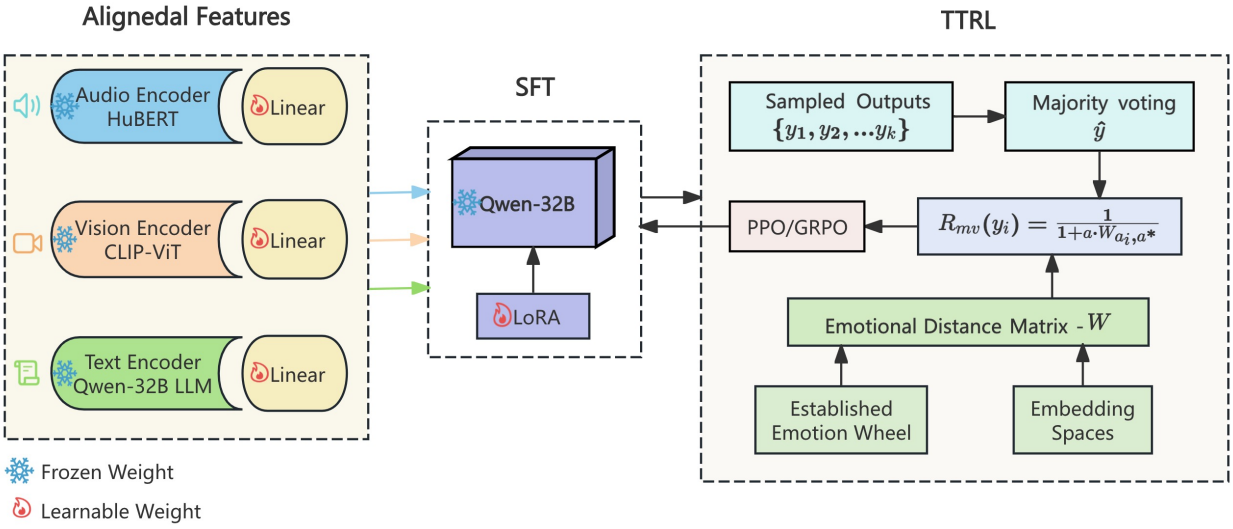

Figure 1. Overall architecture of the proposed RL-enhanced MER framework.

language model (LLM) equipped with lightweight LoRA adapters.

We first perform supervised fine-tuning (SFT) on labeled data to obtain a strong initialization. During test-time, we apply a reinforcement learning (RL) phase based on Test-Time Reinforcement Learning (TTRL), where a Majority Voting-based Verified Reward is computed to guide the LoRA adapters to adapt to the unseen data distribution. PPO or GRPO algorithms are used for stable RL optimization.

3.2. Multimodal Encoders

We adopt three modality-specific encoders to extract deep representations:

- **Text Encoder**: Tokenizer and embedding layer compatible with Qwen-32B LLM.
- **Visual Encoder**: CLIP-ViT (Agarwal et al., 2021) is used to obtain visual embeddings.
- **Audio Encoder**: HuBERT (Hsu et al., 2021) is employed to extract audio embeddings.

All modality embeddings are projected into a unified feature space via linear layers and concatenated to form the input to the LLM.

3.3. Supervised Fine-tuning (SFT)

We perform supervised fine-tuning (SFT) on labeled MER datasets (e.g., CMU-MOSEI, IEMOCAP) to warm up the LLM parameters. To ensure efficiency and flexibility, we freeze the Qwen-32B backbone and introduce LoRA (Low-Rank Adaptation) adapters (Hu et al., 2022), which are lightweight trainable modules inserted into the LLM layers.

The advantages of this design are twofold:

- It enables efficient training with fewer parameters.
- It allows fast and targeted adaptation during test-time reinforcement learning.

3.4. RL-enhanced Test-time Adaptation

3.4.1. Majority Voting-based Verified Reward

During the TTRL phase, for each input sample x, we perform k rollouts with the current policy π_θ to generate sampled outputs $\{y_1, y_2, \ldots, y_k\}$. We extract the emotion labels $\{a_1, a_2, \ldots, a_k\}$ from these outputs and determine the majority vote a^* as:

$$a^* = \arg\max_a \sum_{i=1}^{k} \mathbb{I}(a_i = a)$$

To provide a more informative and fine-grained reward signal for emotion classification, we adopt the idea of emotional distance, inspired by (Zhao et al., 2024). Specifically, we define an emotional distance matrix $W \in \mathbb{R}^{C \times C}$, where C is the number of emotion classes, and each entry $W_{i,j}$ represents the semantic distance between emotion labels i and j.

The distance matrix W is constructed based on established emotion wheels, such as Plutchik's wheel (Plutchik, 2001), or can be empirically derived from embedding spaces (e.g., using cosine distances between emotion label embeddings). Typically, the matrix satisfies:

$$W_{i,j} \geq 0, \quad W_{i,i} = 0, \quad W_{i,j} = W_{j,i}$$

Based on W, we compute the Verified Reward $R_{mv}(y_i)$ for each sampled output y_i (with predicted label a_i) as:

$$R_{mv}(y_i) = \frac{1}{1 + \alpha \cdot W_{a_i,a^*}}$$

where $\alpha > 0$ is a scaling hyperparameter controlling the sensitivity to emotional distance.

Key properties of this reward function:

- If $a_i = a^*$ (perfect match), $W_{a_i,a^*} = 0$, hence $R_{mv} = 1$.
- If a_i is close to a^* (emotionally similar), W_{a_i,a^*} is small $\Rightarrow$ R_{mv} is high.
- If a_i is emotionally distant from a^*, W_{a_i,a^*} is large $\Rightarrow$ R_{mv} is low.

Compared to binary rewards, this soft reward formulation enables more stable RL optimization and encourages the model to learn emotionally coherent predictions. It aligns with recent trends in affective computing that emphasize structure-aware learning (Zhao et al., 2024).

3.4.2. RL Optimization with PPO / GRPO

We adopt Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to optimize the LoRA adapters during test-time adaptation.

The optimization objective over an unlabeled test set D is:

$$\max_{\theta} \mathbb{E}_{x \sim D} \left[\mathbb{E}_{y \sim \pi_\theta(y|x)} \left[R_{mv}(y) \right] \right]$$

PPO Objective:

$$L_{PPO}(\theta) = \mathbb{E}_{x \sim D, y \sim \pi_\theta(y|x)} \Big[\min\big(r_t(\theta) \cdot R_{mv}(y), \\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) \cdot R_{mv}(y)\big) \Big]$$

where

$$r_t(\theta) = \frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)}$$

GRPO Objective:

$$A_i = \frac{R_{mv}(y_i) - \mu_R}{\sigma_R}, \tag{1}$$

$$L_{GRPO}(\theta) = \mathbb{E}_{x \sim D, y \sim \pi_\theta(y|x)} \left[A_i \cdot \log \pi_\theta(y_i|x) \right] \tag{2}$$

where μ_R and σ_R are the mean and standard deviation of rewards within each group (e.g., per batch).

Parameter Update Scope: During RL optimization, the Qwen-32B backbone and modality encoders (HuBERT/CLIP) are kept frozen. Only the parameters of the LoRA adapters are updated, enabling efficient and targeted test-time adaptation.

This RL-enhanced adaptation procedure allows the model to dynamically adjust to distribution shifts and improve consistency and generalization in emotion recognition under unlabeled test scenarios.

4. Experiments

4.1. Datasets

We evaluate our proposed RL-enhanced MER framework on two widely used multimodal emotion recognition datasets:

CMU-MOSEI (Zadeh et al., 2018) is a large-scale benchmark for multimodal sentiment and emotion analysis, comprising over 23,000 labeled video clips from YouTube. Each clip includes synchronized text, audio, and visual modalities, with annotations for six basic emotions: happiness, sadness, anger, surprise, disgust, and fear.

IEMOCAP (Busso et al., 2008) is a high-quality dataset containing dyadic conversations recorded by professional actors. It provides aligned text, audio, and visual modalities along with fine-grained emotion labels. Due to its richness and structure, IEMOCAP is commonly used for evaluating multimodal emotion recognition models.

4.2. Experimental Setup

We adopt Qwen-32B as the backbone LLM, enhanced with LoRA adapters (Hu et al., 2022) integrated into selected transformer layers. The modality encoders are initialized as follows:

- **Text**: Tokenized using a Qwen-compatible tokenizer.
- **Visual**: CLIP-ViT (Agarwal et al., 2021).
- **Audio**: HuBERT-base (Hsu et al., 2021).

Training Protocol:

- **Supervised Fine-Tuning (SFT)**: The model is first trained on 80% of the labeled data to establish a strong initialization.

Table 1. Preliminary results on MER benchmarks.

Method	Accuracy	Macro-F1	Consistency
CMU-MOSEI			
Emotion-LLaMA (Cheng et al., 2024)	67.4%	65.1%	0.812
Emotion-LLaMA + RLVR (Zhao et al., 2025a)	69.0%	66.5%	0.825
Ours (TTRL + PPO)	**70.3%**	**67.8%**	**0.836**
Ours (TTRL + GRPO)	**71.0%**	**68.2%**	**0.842**
IEMOCAP			
Emotion-LLaMA (Cheng et al., 2024)	72.1%	70.4%	0.855
Emotion-LLaMA + RLVR (Zhao et al., 2025a)	73.5%	71.6%	0.864
Ours (TTRL + PPO)	**74.9%**	**72.8%**	**0.872**
Ours (TTRL + GRPO)	**75.6%**	**73.4%**	**0.879**

- **Test-Time Reinforcement Learning (TTRL)**: At inference, we perform $k = 5$ rollouts per test input to generate sampled predictions. The majority vote among these predictions determines the pseudo-label a^*. We set the reward scaling factor $\alpha = 1.0$.
- **Optimization**: We apply PPO (Schulman et al., 2017) or GRPO (Shao et al., 2024) to update only the LoRA parameters, while keeping the backbone and encoders frozen.
- **Evaluation Metrics**: We report Accuracy (Acc), Macro-F1, and Emotion-Consistency Score (Zhao et al., 2024).

4.3. Results and Analysis

Table 1 summarizes the performance of our method and baselines on both datasets.

Our RL-enhanced framework consistently outperforms strong baselines across all metrics. Notably:

- Both PPO and GRPO variants significantly improve accuracy and F1 compared to supervised-only and RLVR-based models.
- The Emotion-Consistency score (Zhao et al., 2024) indicates improved coherence and reliability in predicted emotions after test-time RL adaptation.
- GRPO yields slightly better results than PPO, likely due to its group-based normalization, which stabilizes updates under distribution shifts.

4.4. Ablation Study

To assess the effect of the reward design, we perform ablation experiments by varying the scaling factor α in the Verified Reward function, and by comparing different designs of the emotional distance matrix W (Plutchik-based vs. embedding-based).

Table 2. Impact of α on performance (CMU-MOSEI).

α	Accuracy	Macro-F1
0 (binary reward)	68.1%	66.0%
0.5	69.5%	67.3%
1.0	**71.0%**	**68.2%**
2.0	70.2%	67.7%

The results validate the benefit of soft, structure-aware reward signals. Setting $\alpha = 1.0$ achieves the best trade-off between expressiveness and reward sensitivity, confirming the advantage of our verified reward mechanism over simple binary feedback.

5. Conclusion

In this work, we propose a novel reinforcement learning-enhanced framework for multimodal emotion recognition (MER), which integrates Test-Time Reinforcement Learning (TTRL) with a Majority Voting-based Verified Reward mechanism. Our approach incorporates pre-trained multimodal encoders and a Qwen-32B large language model, augmented with lightweight LoRA adapters, to enable efficient and dynamic adaptation to unseen distributions.

Motivated by recent progress in affective computing, we introduce a fine-grained, emotion-aware reward function based on an emotional distance matrix, offering more informative and stable guidance for the learning process. Extensive experiments on benchmark datasets demonstrate that our method consistently improves both accuracy and emotional consistency, outperforming strong baselines such as Emotion-LLaMA and RLVR-based approaches.

These findings highlight the potential of combining RL-based test-time adaptation with large-scale multimodal language models to improve generalization and robustness, especially under distributional shifts.

For future work, we plan to explore the following directions:

- Designing more expressive and adaptive reward shaping techniques using learned or hierarchical emotion spaces.
- Extending our framework to additional MER benchmarks and real-world, in-the-wild scenarios.
- Investigating the interplay between different RL algorithms and various emotion representation paradigms (e.g., discrete, dimensional, or latent embeddings).

We believe this research opens promising avenues for developing AI systems that are not only more adaptive and robust, but also more emotionally intelligent and context-aware, paving the way for safer and more human-aligned interaction.

References

Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J. W., and Brundage, M. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2):423–443, 2018.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., and Narayanan, S. S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, pp. 205–211, 2004.

Busso, C., Bulut, M., Lee, C. M., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., and Narayanan, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359, 2008.

Chen, Y. et al. Memocmt: Memory-enhanced multimodal contrastive learning for emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Cheng, X. et al. Emotion-llama: An instruction-tuned multimodal llama for fine-grained emotion understanding, 2024.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Y. et al. Tacfn: Temporally aligned cross-modal fusion network for multimodal emotion recognition. In *Proceedings of the 2022 ACM International Conference on Multimedia*, 2022.

Plutchik, R. 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, 2001.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P., and Salakhutdinov, R. Multimodal transformer for multimodal emotion recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6478–6488, 2019.

Watkins, C. J. and Dayan, P. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1103–1114, 2017.

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of ACL*, 2018.

Zhang, C., Zhang, Y., and Cheng, B. Rl-emo: A reinforcement learning framework for multimodal emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10246–10250. IEEE, 2024.