AffectGPT-RL: Revealing Roles of Reinforcement Learning in Open-Vocabulary Emotion Recognition

Zheng Lian[1*], **Fan Zhang**[2], **Lan Chen**[1], **Yazhou Zhang**[3], **Rui Liu**[4], **Jinyang Wu**[5], **Haoyu Chen**[6], **Xiaobai Li**[7], **Xiaojiang Peng**[8], **Bin He**[1], **Jianhua Tao**[5]

[1]State Key Laboratory of Autonomous Intelligent Unmanned Systems, Tongji University
[2]The Chinese University of Hong Kong, [3]Tianjin University, [4]Inner Mongolia University
[5]Tsinghua University, [6]CMVS, University of Oulu, [7]Zhejiang University
[8]Shenzhen Technology University

Abstract

Open-Vocabulary Multimodal Emotion Recognition (OV-MER) aims to predict emotions without being constrained by predefined label spaces, thereby enabling fine-grained emotion understanding. Unlike traditional discriminative methods, OV-MER leverages generative models to capture the full spectrum of emotions and employs emotion wheels (EWs) for metric calculation. Previous approaches primarily rely on token-level loss during training. However, this objective is misaligned with the metrics used in OV-MER, and these metrics cannot be directly optimized via gradient backpropagation. To address this limitation, we turn our attention to reinforcement learning, as this strategy can optimize non-differentiable objectives. We term this framework **AffectGPT-RL**. Furthermore, we conduct extensive experiments to elucidate the role of reinforcement learning in this task, revealing the necessity of the reasoning process, the impact of different rewards, and the generalizability to other emotion tasks such as sentiment analysis and basic emotion recognition. Experimental results demonstrate that AffectGPT-RL yields significant performance improvements on OV-MER. Beyond this task, we also achieve remarkable performance gains on basic emotion recognition, attaining state-of-the-art results on MER-UniBench. To the best of our knowledge, this is the pioneering work exploring the role of reinforcement learning in OV-MER, providing valuable guidance for subsequent researchers. Our code is provided at: https://github.com/zeroQiaoba/AffectGPT/tree/master/AffectGPT-RL.

1 Introduction

Emotion is highly related to human cognition, decision-making, and behavior, playing a vital role in our daily life [18, 47, 65]. Naturally, humans express emotions through multiple modalities [28], including (micro)-gestures, (micro)-expressions, audio tones [40, 17], linguistic content [29], and other cues [2, 4, 10, 36]. To better understand human emotions, it is essential to integrate these clues, enabling models to capture emotional nuances [42]. This has led to the development of Multimodal Emotion Recognition (MER) [33, 74, 29]. Traditional MER approaches primarily rely on Ekman's theory [9, 27, 41], which categorizes human emotions into six basic labels: *anger*, *disgust*, *happiness*, *sadness*, *fear*, and *surprise*. However, human emotions extend far beyond these basic categories [8, 48], and such a restricted label space inevitably leads to imprecise emotion representations.

*Corresponding Author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Recently, open-vocabulary MER (OV-MER) has emerged as a promising research direction, shifting emotion recognition from constrained categories to the full spectrum, thereby enabling more nuanced emotion representations [30, 32]. To support this paradigm shift, OV-MER introduces new solutions and metrics. Regarding solutions, it transitions from discriminative to generative approaches, leveraging the extensive vocabulary of large language models (LLMs) to expand the recognition scope. Regarding metrics, it employs emotion wheels (EWs) to capture semantic relationships between distinct emotions. Appendix F.2 provides a detailed illustration of these metrics. Current works primarily address this task by employing token-level loss to align predicted and ground-truth labels [6, 30]. However, these methods suffer from a critical misalignment: token-level loss exhibits limited correlation with EW-based metrics. For instance, *joyful* and *awful* share high token-level similarity yet possess low semantic similarity in terms of their underlying emotions. Furthermore, since EW-based metrics are non-differentiable (see Appendix F.3), they cannot be directly optimized via gradient backpropagation, posing significant challenges for existing solutions. Figure 1 illustrates the motivation behind our work.

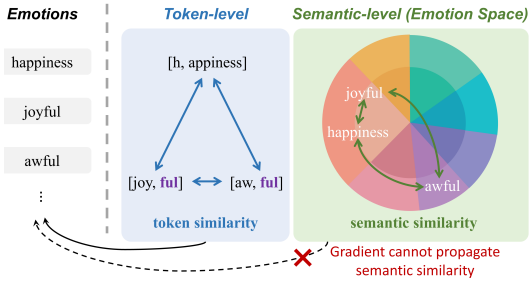


Figure 1: **Motivation.** Token similarity differs from semantic similarity. Existing work relies on token similarity, whereas the evaluation metrics for OV-MER depend on semantic similarity. Moreover, semantic similarity is inherently non-differentiable. This paper aims to address these issues through reinforcement learning.

To address these limitations, we shift our focus to reinforcement learning, as this strategy enables the optimization of non-differentiable objectives. We term this framework **AffectGPT-RL**. Specifically, we utilize EW-based metrics as reward functions and apply reinforcement learning to maximize these rewards, facilitating the direct optimization of these metrics. Meanwhile, we conduct extensive experiments to elucidate the role of reinforcement learning in OV-MER, including the necessity of the reasoning process, the influence of different reward functions, and the impact of reinforcement learning on other emotion tasks, such as sentiment analysis and basic emotion recognition. During training, we observe the issue of reward hacking [11], where the model tends to generate lengthy predictions containing redundant emotional expressions. To mitigate this, we introduce length penalties into the reward function, encouraging more concise outputs while maintaining good performance. Experimental results demonstrate that AffectGPT-RL achieves significant improvements on OV-MER. Beyond this task, our framework also yields notable performance gains in basic emotion recognition, attaining state-of-the-art performance on MER-UniBench [30]. *This paper should not be viewed merely as an application of reinforcement learning, but rather as a significant milestone in advancing open-vocabulary emotion research. This work will inspire more researchers to consider reinforcement learning as a foundational tool for addressing this task.* The main contributions of this paper are summarized below:

- **(Framework)** This is the first work to reveal the role of reinforcement learning in OV-MER.
- **(Experiment)** We conduct systematic experiments, revealing the role of the reasoning process, the impact of different rewards, and the generalization of reinforcement learning to other tasks.
- **(Performance)** Our method achieves state-of-the-art performance on fine-grained and basic emotion recognition, highlighting the potential of reinforcement learning in affective computing.

2 AffectGPT-RL

This paper aims to elucidate the role of reinforcement learning in OV-MER. Following previous works [57], our training pipeline comprises two phases: cold-start training and reinforcement learning. In the first phase, we use a *large-scale, coarse-grained* dataset to establish initial capabilities in emotion understanding and format alignment. In the second phase, we use a *high-quality, fine-grained* dataset for reward calculation and policy updates. Figure 2 illustrates the overall pipeline of AffectGPT-RL.

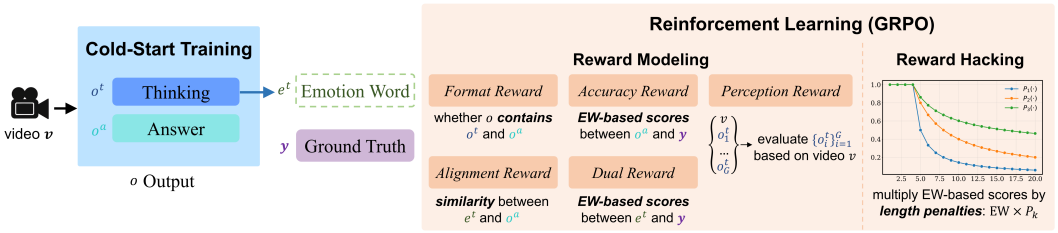


Figure 2: **AffectGPT-RL.** Our training process consists of two phases: cold-start training and reinforcement learning. In the first phase, we output both thinking and answers. In the second phase, we propose five rewards and introduce length penalties to mitigate reward hacking.

2.1 Notation Definition

This section first clarifies the necessary notations. In Figure 2, we use v to represent the video and q to denote the user message. The output o consists of two components: the thinking o^t and the answer o^a. Specifically, o^t and o^a correspond to the content enclosed within the <think> and </think> tags and the <answer> and </answer> tags, respectively. We denote the ground-truth open-vocabulary labels as y. Our main goal is to predict the emotion y based on the video content v. This paper focuses on utterance-level emotion recognition, ensuring fair comparison with prior work [30, 32]. Appendix J.4 provides additional clarification regarding our task formulation.

2.2 Cold-Start Training

Human emotions are highly dependent on subtle, multimodal cues [22, 25], posing significant challenges for current models to capture all these signals and accurately interpret emotional states [37, 45]. To enhance emotion understanding, we employ large-scale descriptive emotion datasets for cold-start training. Specifically, we output both reasoning (enclosed within <think> and </think> tags) and answers (enclosed within <answer> and </answer> tags). Since *descriptive* emotion datasets already follow a reasoning-like format [38], we treat these descriptions as o^t. Given our primary focus on OV-MER, we use *open-vocabulary* emotions as o^a. This output format unifies *descriptive* and *open-vocabulary* emotions into a single task, enabling us to investigate whether these two tasks can mutually benefit each other. As demonstrated in Section 4.5, the quantity of cold-start data significantly affects both cold-start training performance and subsequent reinforcement learning, which further underscores our rationale for utilizing large-scale datasets in this phase.

2.3 Reinforcement Learning

Reinforcement learning relies on policy optimization and reward design. Specifically, we employ GRPO [57] for policy updates and propose five reward functions, all designed to guide the reasoning and answers. In this paper, we focus on designing reward models that better address OV-MER and leave the exploration of optimization methods to future work.

2.3.1 Reward Modeling

Format Reward. To ensure structured outputs, we define a binary reward function that evaluates whether the model's response contains both thinking and answer components.

$$\mathcal{R}_{\text{format}}(o|v,q,y) = \begin{cases} 1, \text{is_required_format}(o) \\ 0, \text{otherwise} \end{cases} \tag{1}$$

Accuracy Reward. We employ the official metrics of OV-MER to compute *accuracy rewards* [32]. Unlike discriminative MER, which centers on a fixed label space, OV-MER imposes no constraints on the label space. Consequently, models may generate synonyms, i.e., emotions with identical meanings but different expressions, which poses challenges for evaluation. To address this, OV-MER introduces EW-based metrics, which leverage the structural information of emotion wheels for clustering. Appendix F provides a detailed calculation procedure for these metrics.

$$\mathcal{R}_{\text{accuracy}}(o|v,q,y) = \text{EW}\left(o^a, y\right). \tag{2}$$

Alignment Reward. *Accuracy reward* ensures the correctness of o^a. However, without constraints on the reasoning process, the generated output may include irrelevant o^a and o^t. To address this, we introduce *alignment reward*, ensuring that the reasoning process is consistent with the final answer. As illustrated in Figure 2, the model output consists of both the thinking o^t and the answer o^a. Specifically, we first extract emotion words e^t from o^t. We then calculate the similarity between e^t and o^a, serving as the alignment reward. By optimizing this reward, we encourage correlation between the reasoning process o^t and the predicted answer o^a. Appendix G provides the detailed methods for emotion word extraction and similarity calculation.

$$\mathcal{R}_{\text{alignment}}(o|v,q,y) = \text{is_similar}\left(o^t \rightarrow e^t, o^a\right). \tag{3}$$

Dual Reward. Similar to *alignment reward*, this reward also imposes constraints on o^t. Specifically, we require the emotion words extracted from o^t to match the true label y:

$$\mathcal{R}_{\text{dual}}(o|v,q,y) = \text{EW}\left(o^t \rightarrow e^t, y\right). \tag{4}$$

Perception Reward. This reward constrains the thinking output o^t to ensure its alignment with the video content. To achieve this, we draw inspiration from EmoPrefer [35], which proves that the MLLM can serve as a judge to compare two descriptions (d_1 and d_2) and determine which better reflects the emotional state of characters in a given video. To extend this to multiple descriptions, we propose to use the Bradley-Terry algorithm. Specifically, suppose we have N descriptions $\{d_i\}_{i=1}^N$. We first compute all pairwise preferences using the MLLM, and then apply the Bradley-Terry algorithm to estimate the relative strength of each description. Finally, outputs ranked in the top 50% receive a reward of 1, while the rest receive 0, thereby encouraging the generated reasoning to remain faithful to the video. Appendix H provides further details regarding the *perception reward*.

$$\mathcal{R}_{\text{perception}}(\{o_i\}_{i=1}^G|v,q,y) = \begin{cases} 1, \text{is_top50}(o_i^t|v) \\ 0, \text{otherwise} \end{cases} \tag{5}$$

2.3.2 Reward Hacking

During inference, we observe that the model tends to generate lengthy o^a containing numerous emotional words. This is primarily because EW-based metrics automatically disregard synonyms during calculation (see Appendix F). Consequently, synonyms do not influence the metric scores, whereas predicting more words increases the likelihood of matching the ground truth. This results in verbose outputs, which in turn increases inference time and reduces the readability of o^a. This phenomenon, also known as *reward hacking* [11], occurs when a model exploits loopholes or flaws in the reward function design to achieve high rewards while exhibiting undesired behavior. To address this, we introduce length penalties, encouraging the model to maximize rewards with fewer prediction words. Specifically, we propose three types of penalties as follows:

$$P_1 = \begin{cases} 1, & |o^a| \le |y| \\ \frac{1}{|o^a|-|y|+1}, & |o^a| > |y| \end{cases}, \; P_2 = \begin{cases} 1, & |o^a| \le |y| \\ \frac{|y|}{|o^a|}, & |o^a| > |y| \end{cases}, \; P_3 = \begin{cases} 1, & |o^a| \le |y| \\ \frac{\log|y|}{\log|o^a|}, & |o^a| > |y| \end{cases} \tag{6}$$

where $|o^a|$ and $|y|$ denote the number of emotion words in the predictions and ground truth, respectively. Figure 2 visualizes the values of these penalties (see the *reward hacking* part). From $P_1(\cdot)$ to $P_3(\cdot)$, the penalty on prediction length progressively loosens. We then apply these length penalties to the reward functions that include EW-based metrics: *accuracy reward* and *dual reward*.

$$\mathcal{R}_{\text{accuracy}}(o|v,q,y) = P_k(o^a, y) \times \text{EW}(o^a, y), \tag{7}$$

$$\mathcal{R}_{\text{dual}}(o|v,q,y) = P_k(o^t \rightarrow e^t, y) \times \text{EW}(o^t \rightarrow e^t, y). \tag{8}$$

3 Experimental Setup

Corpus Description. AffectGPT-RL consists of two training phases. In the cold-start phase, we utilize a *large-scale* dataset to endow the model with emotion understanding and format alignment capabilities. For this purpose, we select MER-Caption+ [30], which contains 31K samples annotated with rich emotion descriptions and open-vocabulary labels. In the reinforcement learning phase, we require a *high-quality* dataset specifically designed for OV-MER. Accordingly, we choose MER2025-OV [31], which features human-annotated, high-quality open-vocabulary labels. For evaluation,

Table 1: **Dataset statistics.** This table provides statistics for the training and testing sets. We rigorously verify that no data leakage exists between these two sets (see Appendix J.3).

Dataset	#Samples	Emotion Recognition Task			
		Sentiment Analysis	Basic Emotion	Open-Vocabulary Emotion	Descriptive Emotion
Training Phase 1: Cold-start training					
MER-Caption+ [30]	31,327	×	×	√	√
Training Phase 2: Reinforcement learning					
MER2025-OV [31]	1,000	×	×	√	×
Testing					
OV-MERD+ [30]	532	×	×	√	×
MER-UniBench [30]	12,799	√	√	√	×

we assess the model's performance on OV-MERD+ [30] and MER-UniBench [30] to measure its capabilities in open-vocabulary and other emotion tasks. We conduct a rigorous overlap check to ensure no data leakage exists between the training and testing sets (see Appendix J.3). Table 1 provides dataset statistics. During cold-start training, we leverage both descriptive and open-vocabulary emotions. However, in reinforcement learning [64], we exclusively use open-vocabulary emotions. This decision stems from the fact that the reward calculation does not rely on human-annotated thinking results. Therefore, our dataset selection is well-suited to the distinct requirements of each training stage. For testing, besides open-vocabulary emotions, we also evaluate performance on other emotion tasks to reveal the generalization of reinforcement learning to these tasks.

Implementation Details. During training, we set the learning rate to 1e-5 and the maximum number of epochs to 60. The entire implementation is conducted using PyTorch, and the code is executed on an NVIDIA A100 GPU. Due to memory constraints, we use a batch size of 3 for cold-start training and a batch size of 1 for reinforcement learning. For the model architecture, we adopt the default hyperparameters of AffectGPT [30]. This setup ensures a fair comparison with the baselines, thereby isolating the performance gains attributable to our proposed framework.

4 Results and Discussion

In this section, we conduct a series of experiments examining the impact of different rewards (Section 4.1), the necessity of the thinking process (Section 4.2), the role of different length penalties (Section 4.3), generalization to other emotion tasks (Section 4.4), and various ablation studies regarding hyper-parameter sensitivity and data scaling in cold-start training and reinforcement learning (Section 4.5). Ultimately, we aim to answer the key question: *Is reinforcement learning a promising approach for addressing open-vocabulary emotion recognition?*

4.1 Impact of Different Rewards

This section evaluates the effectiveness of different reward functions. In Figure 3, we report performance under three scenarios: (1) *without reinforcement learning*, (2) *with reinforcement learning using a single reward*, and (3) *with reinforcement learning using multiple rewards*.

Single-reward RL. In Figure 3 (a), the *accuracy reward* and *dual reward* are the most effective, significantly improving performance compared to the model without reinforcement learning. These gains stem from the direct optimization of EW-based metrics via either the thinking o^t or the answer o^a, highlighting the importance of aligning reinforcement learning objectives with evaluation metrics in OV-MER. In contrast, other rewards lead to performance degradation. For instance, the *format reward* and *alignment reward* show limited impact, as these aspects are already well-learned during the cold-start phase. The suboptimal performance of the *perception reward* may be attributed to current MLLMs' inability to decode human preferences accurately, which can introduce reward noise [35]. Interestingly, the *accuracy reward* and *dual reward* yield similar performance, suggesting that while EW-based metrics are crucial, the choice of optimizing o^t or o^a has minimal impact. Therefore, in subsequent experiments, we use the *accuracy reward* by default.

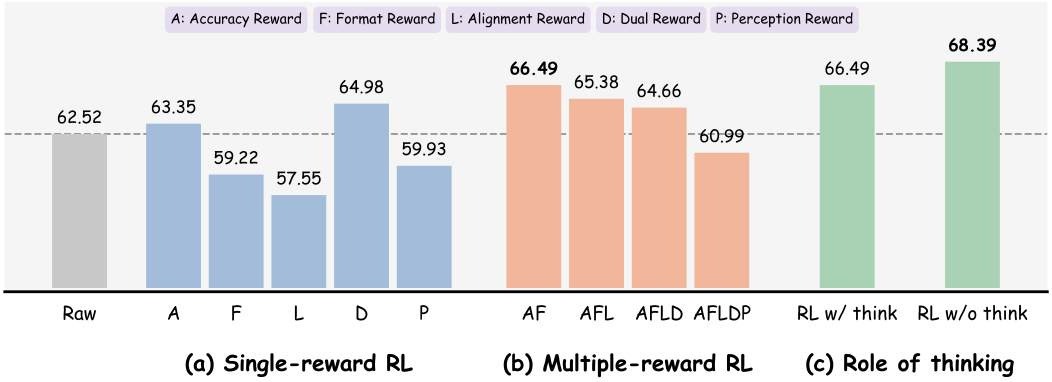


Figure 3: *(a) Single-reward RL.* This figure reports performance obtained using RL with a single reward; *(b) Multiple-reward RL.* It reports performance obtained using RL with multiple rewards; *(c) Role of thinking.* It shows that a simpler model, one that only outputs the answer o^a and calculates rewards based solely on o^a, achieves better performance.

Multiple-reward RL. In Figure 3 (b), combining the *accuracy reward* and *format reward* yields further performance improvements over single-reward setups. However, employing more rewards, such as all five simultaneously, results in performance degradation. The primary issue stems from interference between rewards: optimizing for multiple objectives can dilute their individual effectiveness, ultimately compromising the outcome. For instance, emphasizing one reward may inadvertently diminish the influence of others. These findings suggest that more rewards do not necessarily equate to better performance. Instead, selecting an appropriate combination is crucial.

4.2 Role of Thinking

In Section 4.1, the most effective rewards are the *accuracy reward* and *dual reward*, both of which are derived from evaluation metrics. This observation motivates a heuristic hypothesis: *Could we modify the output mode to generate and optimize solely on the emotion words o^a, thereby encouraging the model to focus more on EW-based metrics?* Specifically, during cold-start training, we eliminate the thinking process and require the output to contain only o^a. During reinforcement learning, we optimize the *accuracy reward* directly on o^a. We refer to the modified model as RL (w/o think) and the original model as RL (w/ think). In Figure 3 (c), removing the thinking process yields improved performance. This improvement can be attributed to two factors. First, the thinking process is derived from coarse-grained datasets, where inherent noise may lead to flawed reasoning and consequently degrade performance. Second, incorporating additional thinking content diverts the model's attention from solving OV-MER, thereby increasing training difficulty. In future work, we plan to design high-quality cold-start data and conduct an in-depth analysis to determine for which samples or emotional states the thinking process is truly necessary.

4.3 Impact of Different Penalties

Following the discussion in Section 4.2, we adopt RL (w/o think) as the default setting. During inference, we observe that the model tends to generate lengthy predictions containing redundant emotion words. This behavior stems from the suboptimal design of the EW-based metrics, where synonyms do not influence the metric scores, whereas predicting more words increases the likelihood of matching the ground truth (see Appendix F). To mitigate this issue, we introduce three types of penalties in Section 2.3.2, ordered by increasing penalty strength from $P_3(\cdot)$ to $P_1(\cdot)$. This section evaluates the effectiveness of these penalties across three di-

Table 2: **Impact of different penalties.** Our goal is to identify a penalty that maintains model performance, prevents overlong outputs, and avoids introducing reward degradation. Based on these criteria, $P_2(\cdot)$ and $P_3(\cdot)$ are more suitable choices.

Penalty	Performance	Overlong	Degradation
$P_1(\cdot)$	67.60	✗	✔
$P_2(\cdot)$	67.79	✗	✗
$P_3(\cdot)$	68.05	✗	✗
None	68.39	✔	✗

mensions: (1) *their impact on performance*; (2) *their ability to prevent overlong outputs*; and (3) *whether they introduce reward degradation*. As shown in Table 2, all three penalties successfully address overlong outputs without compromising performance. However, the strictest penalty, $P_1(\cdot)$, leads to reward degradation. Therefore, $P_2(\cdot)$ and $P_3(\cdot)$ are preferable.

4.4 Influence on Other Emotion Tasks

The preceding sections focus on open-vocabulary emotion recognition. In this section, we extend our evaluation to MER-UniBench [30], a unified benchmark comprising three key tasks: sentiment analysis, basic emotion recognition, and fine-grained emotion detection. Appendix J provides further details regarding this benchmark. This section aims to assess the generalization of reinforcement learning to other emotion tasks. To ensure a fair comparison, all methods are compared under the same input modality. Experimental results in Table 3 demonstrate that AffectGPT-RL achieves significant performance improvements in both basic and fine-grained emotion recognition, but shows a slight decline in sentiment analysis. This decline represents an interesting negative result. The underlying cause is twofold. First, the model tends to over-generate nuanced emotional terms even when a simple *positive* or *negative* label would suffice, resulting in penalties under the EW-based metrics. Second, discrepancies exist among different tasks. OV-MER focuses on recognizing discrete emotion words (e.g., *happy* and *surprise*), which aligns closely with basic and fine-grained emotion recognition but diverges from the objectives of sentiment analysis. These factors collectively limit the effectiveness of reinforcement learning in sentiment analysis. Future work will focus on balancing reinforcement learning performance across different tasks. Overall, based on the average performance across the three tasks, AffectGPT-RL still achieves state-of-the-art results on MER-UniBench, validating the efficacy of reinforcement learning in affective computing.

Table 3: **Influence on other emotion tasks.** MER-UniBench is a unified benchmark comprising three key tasks: sentiment analysis, basic emotion recognition, and fine-grained emotion detection. In this table, we aim to reveal the generalization of reinforcement learning to other emotion tasks.

	Sentiment				**Basic**				**Fine-grained**	**Mean**
	MOSI	MOSEI	SIMS	SIMS v2	MER2023	MER2024	MELD	IEMOCAP	OV-MERD+	
Input Modality: Audio, Text										
OneLLM [15]	64.01	54.09	63.39	61.98	25.52	17.21	28.32	33.44	22.25	41.14
SECap [66]	55.76	54.18	59.51	57.41	40.95	52.46	25.56	36.92	36.97	46.64
PandaGPT [59]	66.06	61.33	62.93	58.88	33.57	39.04	31.91	36.55	31.33	46.84
Qwen-Audio [7]	70.09	46.90	70.73	65.26	41.85	31.61	49.09	35.47	32.36	49.26
SALMONN [61]	81.00	67.03	68.69	65.93	55.53	45.38	45.62	46.84	45.00	57.89
AffectGPT [30]	**83.46**	**80.74**	82.99	**83.75**	72.94	73.41	**56.63**	55.68	59.98	72.18
AffectGPT-RL (Ours)	80.13	80.01	**84.49**	82.31	**81.69**	**93.49**	**63.74**	**63.85**	**65.49**	**77.24**
Input Modality: Video, Text										
Otter [21]	52.89	50.44	57.56	53.12	16.41	14.65	22.57	29.08	16.63	34.82
Video-LLaVA [39]	56.37	61.64	53.28	57.45	36.93	30.25	30.73	38.95	34.00	44.40
PandaGPT [59]	58.50	64.25	62.07	65.25	39.13	47.16	38.33	47.21	35.07	50.77
Video-ChatGPT [46]	54.42	63.12	64.82	65.80	44.86	46.80	37.33	56.83	39.80	52.64
VideoChat2 [24]	66.84	54.32	69.49	70.66	33.67	54.50	36.64	48.70	39.21	52.67
LLaMA-VID [26]	61.78	63.89	69.35	67.48	50.72	57.60	42.75	46.02	45.01	56.07
VideoChat [23]	65.13	63.61	69.52	72.14	48.73	57.30	41.11	48.38	44.52	56.71
Chat-UniVi [19]	54.53	63.18	68.15	66.36	57.62	65.67	45.61	52.37	48.00	57.94
mPLUG-Owl [67]	72.40	72.91	72.13	75.00	56.86	59.89	49.11	55.54	48.18	62.45
AffectGPT [30]	**82.39**	**81.57**	**87.20**	**86.29**	74.58	75.29	57.63	62.19	61.65	74.31
AffectGPT-RL (Ours)	78.78	79.07	85.91	85.85	**77.72**	**85.29**	**61.09**	**67.42**	**62.42**	**75.95**
Input Modality: Audio, Video, Text										
PandaGPT [59]	61.92	67.61	68.38	67.23	40.21	51.89	37.88	44.04	37.12	52.92
R1-Omni [75]	58.02	56.48	71.82	68.58	64.17	67.43	43.20	51.58	55.24	59.61
Emotion-LLaMA [6]	66.13	67.66	78.32	77.23	59.38	73.62	46.76	55.47	52.97	64.17
AffectGPT [30]	**81.30**	**80.90**	**88.49**	**86.18**	78.54	78.80	55.65	60.54	62.52	74.77
AffectGPT-RL (Ours)	79.39	79.24	88.25	84.97	**84.32**	**93.75**	**63.12**	**74.26**	**68.05**	**79.48**

4.5 Ablation Study

Data Amount in Cold-start Training. AffectGPT-RL consists of two stages: cold-start training (denoted as s_1) and reinforcement learning (denoted as s_2). In this section, we examine the impact of the amount of cold-start data on both s_1 and s_2. Experimental results shown in Figure 4 (left)

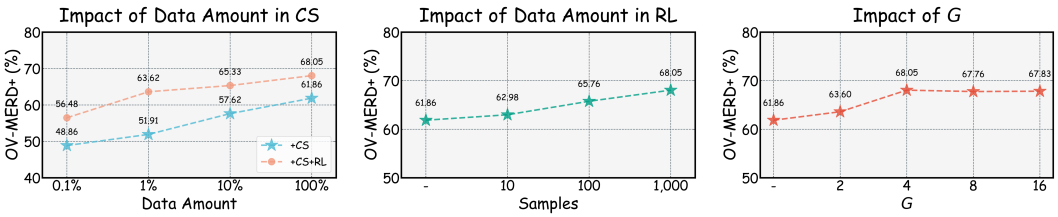


Figure 4: **Left: Impact of data amount in cold-start training.** This figure shows the performance of models trained with varying amounts of cold-start data and analyzes their influence on subsequent reinforcement learning. **Mid: Impact of data amount in reinforcement learning.** This figure shows that using a larger amount of data in reinforcement learning leads to better performance. **Right: Impact of G.** "–" denotes the baseline model without reinforcement learning.

indicate that increasing the amount of cold-start data improves performance in s_1. Furthermore, the performance after reinforcement learning strongly correlates with the performance of the base model from s_1, and additional cold-start data also benefits s_2. This is primarily because reinforcement learning mainly reinforces correct responses rather than introducing new knowledge [44]. Therefore, a high-quality base model is crucial for maximizing the effectiveness of reinforcement learning.

Data Amount in Reinforcement Learning. This section investigates the impact of data quantity on reinforcement learning. Starting from the same model obtained via cold-start training (i.e., the model trained on 100% of the data, as shown in the left part of Figure 4), we evaluate its performance using varying amounts of reinforcement learning data. Experimental results are presented in Figure 4 (mid). We observe that utilizing just 10 samples for reinforcement learning yields performance improvements, confirming its effectiveness in addressing OV-MER. Furthermore, increasing the amount of training data leads to progressively greater performance gains. These trends suggest that AffectGPT-RL has not yet reached its performance upper bound. Therefore, future work will focus on collecting more data to further enhance the model's capabilities.

Impact of G. GRPO requires sampling G outputs for group normalization. Figure 4 (right) reveals the impact of G. Due to GPU memory limitations, we restrict our comparisons to values of G between 2 and 16 and do not evaluate larger values. Experimental results show that, regardless of the specific value of G, the model trained with reinforcement learning consistently outperforms the one without it, confirming the effectiveness of our approach. Meanwhile, the selection of G also matters. A small G does not fully exploit the benefits of GRPO, mainly because limited sampling reduces the chances of identifying optimal outcomes. In contrast, a larger G (e.g., $G \geq 4$) ensures the effectiveness of GRPO, but further increasing G yields no significant improvement. Since a larger G also increases computational cost, we set $G = 4$ as the default configuration.

Stability Test. This section presents a stability test. Specifically, following the experiments in Section 4.3, we report the statistical results for single-run and multi-run settings. For the multi-run setting, we conduct the experiments three times and report both the mean score and the standard deviation. Experimental results are presented in Table 4. From this table, we observe that AffectGPT-RL exhibits good stability across different runs. These results confirm that our conclusions are not artifacts of random chance, demonstrating the reliability of our experimental analysis.

Performance Gain Analysis. AffectGPT-RL consists of two training phases: cold-start training using MER-Caption+ and reinforcement learning using MER2025-OV. In Table 5, we conduct a controlled experiment where AffectGPT is trained on the same two-stage dataset used by AffectGPT-RL. Experimental results show that even when using identical training data, the performance improvement of AffectGPT is marginal and significantly lags behind AffectGPT-RL. These results indicate that the performance gains stem from reinforcement learning rather than from the quantity of training data.

Table 4: **Stability test.** Following the experiments in Section 4.3, we report both single-run and multi-run results.

Penalty	Single-run	Multi-run
$P_1(\cdot)$	67.60	67.74 ± 0.15
$P_2(\cdot)$	67.79	68.05 ± 0.33
$P_3(\cdot)$	68.05	68.00 ± 0.10

Table 5: **Performance gain analysis.** This table reveals that our performance gains stem from reinforcement learning rather than from an increase in data.

Model	Training Data	OV-MERD+
AffectGPT-RL	MER-Caption+ & MER2025-OV	68.39
AffectGPT	MER-Caption+	62.52
AffectGPT	MER-Caption+ & MER2025-OV	62.84

5 Related Works

5.1 Emotion Representation

Dimensional Emotion Theory. This theory uses continuous dimensions to capture the core attributes underlying emotions. For example, Russell [54] proposed a two-dimensional model, characterizing emotions based on *valence* (the pleasantness of an emotion) and *arousal* (the activation level of an emotion). In addition, some researchers have introduced a third dimension, *dominance*, to represent the degree of perceived control over emotions [55]. However, these dimensions are abstract and less intuitive for the general public, thus posing challenges in practical applications.

Basic Emotion Theory. This theoretical framework typically relies on fixed and limited basic labels to describe the emotional states of characters. For example, Ekman [9] argued that human emotions can be categorized into six basic labels, whereas Plutchik [48] identified eight primary emotions that form the basis for other emotional experiences. Nevertheless, human emotions are rich and diverse, extending far beyond these basic labels. Mapping such nuanced emotional expressions onto a restricted set of basic labels inevitably leads to the loss of subtle emotional distinctions.

Open-Vocabulary Emotion. To address these limitations, researchers propose open-vocabulary representations [32], which removes restrictions on the label space and enables models to predict any emotion. Unlike dimensional emotion theory, it focuses on more intuitive emotion words; Unlike basic emotion theory, it avoids restrictive label spaces, enabling more nuanced emotion representation. Therefore, the open-vocabulary paradigm opens up new opportunities for affective computing. However, this field is still in its early stages, and effective solutions remain unclear. This paper presents a pioneering effort that leverages reinforcement learning to tackle this task.

5.2 Reinforcement Learning

Policy Optimization. PPO [56] is a widely used algorithm that leverages a value model for advantage estimation and incorporates token-wise KL penalties to prevent excessive optimization. DPO [53] focuses on directly optimizing policies to align with human preferences, eliminating the need for an explicit reward model. GRPO [57] is a more efficient alternative to PPO, removing the value model and instead using a group-relative baseline computed from different outputs for the same input to guide optimization. Beyond these algorithms, some efforts have been made to enhance training efficiency and stability [16, 68]. This paper does not focus on designing new optimization algorithms, but rather on crafting optimal rewards for OV-MER.

Reward Modeling. Reward systems can be broadly categorized into rule-based and model-based rewards. The former is suitable for tasks with verifiable outcomes. For example, pass or fail unit tests in code generation [20], output language requirements in question-answering tasks [14], and IoU between predicted bounding boxes and ground-truth annotations in object detection [58]. The latter is mainly used for tasks where rewards are difficult to compute, such as safety, helpfulness, and harmlessness [72]. In prior work, how to design reward functions tailored for OV-MER remains unexplored. To this end, we introduce various rewards to supervise both thinking and answers.

6 Conclusion

This paper proposes AffectGPT-RL, the first work to reveal the role of reinforcement learning in OV-MER. Through extensive experiments, we analyze the impact of different rewards and demonstrate the benefits of EW-based rewards. Meanwhile, we identify the limitations of incorporating a thinking process and show the usefulness of different-length penalties in addressing reward hacking. Beyond open-vocabulary emotions, we reveal the impact of reinforcement learning on other emotion tasks, demonstrating benefits for basic and fine-grained emotion recognition, while noting slight performance degradation in sentiment analysis. Furthermore, we investigate the effects of various modules and hyperparameters, confirming that our performance gains stem from reinforcement learning rather than other factors. These conclusions demonstrate the potential of reinforcement learning in open-vocabulary emotions and offer valuable insights for future research. This work will advance the development of this field toward a reinforcement learning era.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62688201, 62671433, 62476146), Shanghai International Science and Technology Cooperation Project (26510713900), Open Research Fund of the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS2026003), and the Innovative Team Project in Guangdong Province (2025KCXTD040).

References

[1] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[2] Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5826–5846, 2021.

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

[4] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 131(6):1346–1366, 2023.

[5] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 5178–5193, 2023.

[6] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024.

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

[8] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, 2020.

[9] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[10] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[11] Tom Everitt, Victoria Krakovna, Laurent Orseau, and Shane Legg. Reinforcement learning with a corrupted reward channel. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4705–4713, 2017.

[12] Yiyang Fang, Jian Liang, Wenke Huang, He Li, Kehua Su, and Mang Ye. Catch your emotion: Sharpening emotion perception in multimodal large language models. In *Forty-second International Conference on Machine Learning*, 2025.

[13] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[15] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595, 2024.

[16] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

[17] Yifan Hu, Rui Liu, Yi Ren, Xiang Yin, and Haizhou Li. Chain-talker: Chain understanding and rendering for empathetic conversational speech synthesis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1988–2003, 2025.

[18] Zhijing Huang, Wen-Jue He, Baotian Hu, and Zheng Zhang. Grading-inspired complementary enhancing for multimodal sentiment analysis. *Information Fusion*, page 104174, 2026.

[19] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024.

[20] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.

[21] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[22] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2782–2800, 2022.

[23] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025.

[24] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[25] Yante Li, Jinsheng Wei, Yang Liu, Janne Kauttonen, and Guoying Zhao. Deep learning for micro-expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(4):2028–2046, 2022.

[26] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024.

[27] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640, 2023.

[28] Yunxin Li, Baotian Hu, Chen Xinyu, Yuxin Ding, Lin Ma, and Min Zhang. A multi-modal context reasoning approach for conditional inference on joint textual and visual clues. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10757–10770, 2023.

[29] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3424–3439, 2025.

[30] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Forty-second International Conference on Machine Learning*, 2025.

[31] Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, et al. Mer 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13837–13842, 2025.

[32] Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, et al. Ov-mer: Towards open-vocabulary multimodal emotion recognition. In *Forty-second International Conference on Machine Learning*, 2025.

[33] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9610–9614, 2023.

[34] Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, et al. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pages 41–48, 2024.

[35] Zheng Lian, Licai Sun, Lan Chen, Haoyu Chen, Zebang Cheng, Fan Zhang, Ziyu Jia, Ziyang Ma, Fei Ma, Xiaojiang Peng, et al. Emoprefer: Can large language models understand human emotion preferences? In *International Conference on Learning Representations*, 2026.

[36] Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[37] Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367, 2024.

[38] Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. Explainable multimodal emotion reasoning. *arXiv preprint arXiv:2306.15401*, 2023.

[39] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024.

[40] Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18698–18706, 2024.

[41] Rui Liu, Haolin Zuo, Zheng Lian, Björn W Schuller, and Haizhou Li. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*, 15(4):1856–1873, 2024.

[42] Rui Liu, Haolin Zuo, Zheng Lian, Hongyu Yuan, and Qi Fan. Hardness-aware dynamic curriculum learning for robust multimodal emotion recognition with missing modalities. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5755–5764, 2025.

[43] Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the International Conference on Multimodal Interaction*, pages 247–258, 2022.

[44] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[45] Hao Lu, Xuesong Niu, Jiyao Wang, Yin Wang, Qingyong Hu, Jiaqi Tang, Yuting Zhang, Kaishen Yuan, Bin Huang, Zitong Yu, et al. Gpt as psychologist? preliminary evaluations for gpt-4v on visual affective computing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 322–331, 2024.

[46] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024.

[47] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[48] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pages 3–33. Elsevier, 1980.

[49] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883, 2017.

[50] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 527–536, 2019.

[51] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.

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

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

[54] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

[55] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

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

[57] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[58] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[59] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants*, pages 11–23, 2023.

[60] Licai Sun, Zheng Lian, Kexin Wang, Yu He, Mingyu Xu, Haiyang Sun, Bin Liu, and Jianhua Tao. Svfap: Self-supervised video facial affect perceiver. *IEEE Transactions on Affective Computing*, 2024.

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

[62] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019.

[63] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128, 2023.

[64] Yuxiang Wu and Baotian Hu. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[65] Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26596–26605, 2024.

[66] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19323–19331, 2024.

[67] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[68] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. In *Proceedings of the Advances in Neural Information Processing Systems*, 2025.

[69] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, 2020.

[70] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[71] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[72] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, 2024.

[73] Zhicheng Zhang, Weicheng Wang, Yongjie Zhu, Wenyu Qin, Pengfei Wan, Di ZHANG, and Jufeng Yang. Videmo: Affective-tree reasoning for emotion-centric video foundation models. In *Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[74] Zhicheng Zhang, Wuyou Xia, Chenxi Zhao, Zhou Yan, Xiaoqiang Liu, Yongjie Zhu, Wenyu Qin, Pengfei Wan, Di Zhang, and Jufeng Yang. Moda: Modular duplex attention for multimodal perception, cognition, and emotion understanding. In *Forty-second International Conference on Machine Learning*, 2025.

[75] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025.

Appendix

A Limitations and Future Work **17**

B Reproducibility Statement **17**

C Ethics Statement **17**

D Social Impact **17**

E Novelty Statement **17**

F EW-based Metric **17**

F.1 Emotion Wheels 17

F.2 Calculation Formula 18

F.3 Non-Differentiability Analysis 19

G Details of Alignment Reward **19**

H Details of Perception Reward **20**

I Baseline Details **20**

J Dataset Details **21**

J.1 Training Datasets 21

J.2 Testing Datasets 22

J.3 Data Leakage Prevention 22

J.4 Task Clarification 22

K Role of RL across Different Architectures **23**

L Impact of Training Data Scale in RL **23**

A Limitations and Future Work

While this paper represents the first work exploring the role of reinforcement learning in open-vocabulary emotions, several limitations remain, which we leave to future work. First, we observe that adding a thinking process leads to performance degradation. Future work will involve designing high-quality cold-start data and conducting an in-depth analysis to determine which samples or emotional states necessitate a thinking process. Second, reinforcement learning yields performance gains for basic and fine-grained emotions but leads to slight degradation in sentiment analysis. Future work will focus on balancing reinforcement learning performance across different tasks. Additionally, we observe that increasing the amount of training data in reinforcement learning leads to progressive performance gains. Future work will focus on collecting more data for reinforcement learning to establish the upper-bound performance of our proposed method.

B Reproducibility Statement

In this paper, we have made every effort to ensure the reproducibility of our work. Our code is available at: https://github.com/zeroQiaoba/AffectGPT/tree/master/AffectGPT-RL.

C Ethics Statement

This paper does not involve the collection of new data or the hiring of annotators. We utilize MER-Caption+ [30], MER2025-OV [31], OV-MERD+ [30], and MER-UniBench [30], with permission from the dataset owners. Furthermore, all datasets are employed in accordance with their licenses. Therefore, this paper raises no ethical concerns.

D Social Impact

Emotion plays an important role in communication, conveying human intentions and thoughts. OV-MER aims to capture subtle differences in emotion representation, serving as an important complement to existing representation methods. This paper centers on OV-MER and reveals the impact of reinforcement learning on OV-MER. Research surrounding OV-MER promises significant future social impact, particularly in enhancing the human-computer interaction experience.

E Novelty Statement

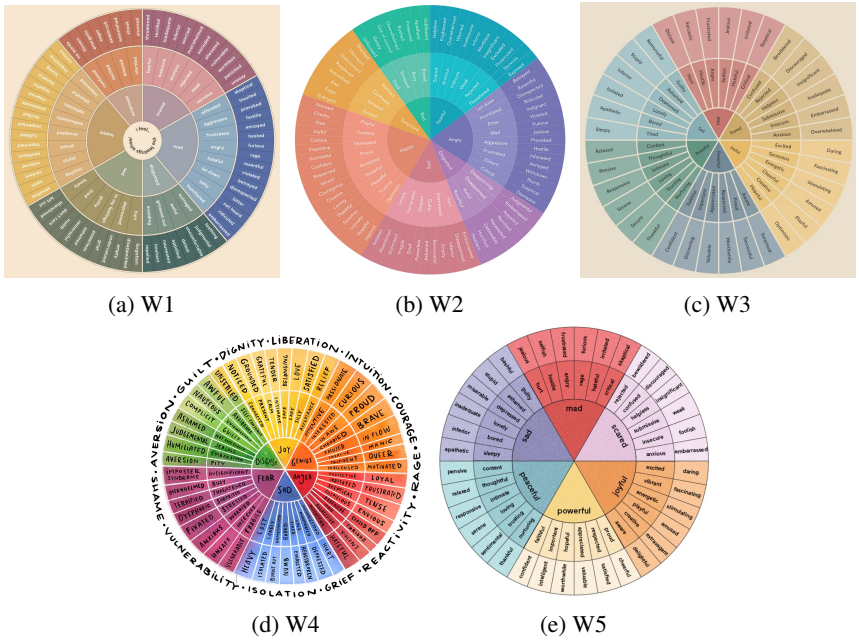
There are currently two main types of emotion tasks: *discriminative emotions* and *generative emotions*. The former primarily involves tasks with a constrained label space, such as basic or dimensional emotions [12, 60, 63]. The latter imposes no restrictions on the label space, giving rise to the emerging area of open-vocabulary emotions [32]. Recent research has shown a growing interest in shifting from *discriminative emotions* to *generative emotions*, driven by the need for more nuanced emotional representations [6, 30, 73]. However, effectively addressing *generative emotions* remains an open challenge. This paper focuses on open-vocabulary emotions, a specific subtask within *generative emotions*, and proposes leveraging reinforcement learning as a solution. During our investigation, several practical issues emerged, including the design of reward functions and strategies to mitigate reward hacking. By tackling these challenges, our experimental results demonstrate the effectiveness of reinforcement learning in handling *generative emotions*. Therefore, this paper should not be viewed merely as an application of reinforcement learning, but rather as a significant milestone in advancing *generative emotions* research. This work will inspire more researchers to consider reinforcement learning as a foundational tool for addressing *generative emotions*.

F EW-based Metric

F.1 Emotion Wheels

The emotion wheel is a structured visual tool designed to categorize and identify human emotions. It organizes emotions into core categories (the inner sections of the wheel) and their nuanced variations

(the outer sections), providing intuitive insights into emotional structure. This paper uses five emotion wheels to compute EW-based metrics, aligning with prior work [30] to ensure fair comparisons. Figure 5 illustrates these emotion wheels.



(a) W1 (b) W2 (c) W3

(d) W4 (e) W5

Figure 5: **Emotion wheels**. This paper uses five emotion wheels to calculate EW-based metrics, ensuring a fair comparison with previous work [30, 32].

F.2 Calculation Formula

We present the calculation formula for EW-based metrics [32]. The computation method primarily consists of two components: eliminating the influence of synonyms and defining set-level metrics.

Handling Synonyms. To mitigate the impact of synonyms, we employ a three-level hierarchical grouping strategy:

(L1) We map different forms of words to their base form. For example, *happier* and *happiness* are mapped to *happy*. To achieve this, for a base word w, we first use LLM to generate various forms of w, with the output denoted as $l_1(w)$. Since different results may be produced across different calls, we repeat this process ten times and merge all the generated words, resulting in $\hat{l}_1(w)$. To map them back to their base form, we simply map $\hat{l}_1(w)$ to w, and we refer to this mapping function as $F_{l_1}(\cdot)$.

(L2) We map synonyms to a unified word. For example, *happy* and *joyful* are mapped to *happy*. To achieve this, we use LLM to generate synonyms for a given word w, repeat the process multiple times, and then merge all the generated words, obtaining $\hat{l}_2(w)$. To map them to their unified form, we map each element in $\hat{l}_2(w)$ to w, and we refer to this mapping function as $F_{l_2}(\cdot)$.

(L3) The emotion wheel offers natural grouping information for emotions [48]. Since there is no consensus on the emotion wheel, we adopt $K = 5$ distinct emotion wheels, consistent with prior work [30, 32] to ensure fair comparisons. Figure 5 visualizes all five emotion wheels. For each wheel w_k, we map all outer labels to their corresponding inner labels. This function is denoted as $F_{l_3}^{w_k}(\cdot)$.

Finally, the clustering functions can be summarized as:

$$G_{w_k}(\cdot) = F_{l_3}^{w_k}(F_{l_2}(F_{l_1}(\cdot))), k \in [1, K]. \tag{9}$$

Set-level Metrics. Since we do not restrict the number of predicted and annotated emotions, we employ set-level evaluation metrics. Specifically, suppose the dataset contains N samples. For a

sample x_i, the true labels are $\mathbf{Y}_i = \{y_i^j\}_{j=1}^{n_i}$ and the predicted labels are $\hat{\mathbf{Y}}_i = \{\hat{y}_i^j\}_{j=1}^{\hat{n}_i}$. Here, n_i and $\hat{n}_i$ are the number of predicted and annotated emotions, respectively. The evaluation metrics are defined as follows:

$$\text{Precision}_{\text{S}}^k = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|G_{w_k}(\mathbf{Y}_i) \cap G_{w_k}(\hat{\mathbf{Y}}_i)\right|}{\left|G_{w_k}(\hat{\mathbf{Y}}_i)\right|}, \tag{10}$$

$$\text{Recall}_{\text{S}}^k = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|G_{w_k}(\mathbf{Y}_i) \cap G_{w_k}(\hat{\mathbf{Y}}_i)\right|}{|G_{w_k}(\mathbf{Y}_i)|}, \tag{11}$$

$$\text{F}_{\text{S}}^k = 2 \times \frac{\text{Precision}_{\text{S}}^k \times \text{Recall}_{\text{S}}^k}{\text{Precision}_{\text{S}}^k + \text{Recall}_{\text{S}}^k}. \tag{12}$$

In this set-wise operation, we automatically remove duplicate emotion words. Finally, we compute the average F1-score across all emotion wheels as the final score:

$$\text{EW}\left(\mathbf{Y}_i, \hat{\mathbf{Y}}_i\right) = \frac{1}{K}\sum_{k=1}^{K}\text{F}_{\text{S}}^k. \tag{13}$$

F.3 Non-Differentiability Analysis

The non-differentiability of EW-based metrics stems from three compounding factors. First, the predicted label set $\hat{\mathbf{Y}}_i$ is obtained via stochastic sampling from the model's logits, which is inherently non-differentiable. Second, the three-level clustering functions $F_{l_1}(\cdot)$, $F_{l_2}(\cdot)$, and $F_{l_3}(\cdot)$ are discrete word-to-word lookup mappings, where gradients could not propagate through these discrete clustering functions. Third, set-level evaluation depends on cardinality operations. For instance, the intersection size $|G_{w_k}(\mathbf{Y}_i) \cap G_{w_k}(\hat{\mathbf{Y}}_i)|$ and the prediction-set size $|G_{w_k}(\hat{\mathbf{Y}}_i)|$ are integer-valued functions of discrete token identities, yielding zero gradient with respect to continuous model parameters. Collectively, these three obstacles render EW-based metrics non-differentiable, precluding their direct optimization via gradient backpropagation and motivating our adoption of reinforcement learning.

G Details of Alignment Reward

Emotion Word Extraction. To compute the alignment reward, we need to extract emotion words e^t from o^t. To do so, we input o^t into Qwen2.5-7B using the following prompt: *Please assume the role of an expert in the field of emotions. We provide clues related to the emotional state of a character. Based on the provided clues, please identify the character's emotional states. Separate different emotional categories with commas, and output only the clearly identifiable emotional categories in a list format. If no emotions can be identified, please return an empty list.* We select Qwen2.5-7B as the LLM backbone to ensure consistency with baselines [30, 32]. This choice mitigates potential biases arising from the use of different LLMs, thereby ensuring a fair comparison with baselines.

Similarity Calculation. Alignment reward involves computing the similarity between two emotion sets. In practice, we use the *Jaccard Similarity Coefficient*, which measures the similarity between two sets by comparing the size of their intersection to the size of their union. For clarity, we adopt the notation defined in Appendix F. The corresponding calculation formula is provided below:

$$\text{Similarity}_{\text{S}}^k = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|G_{w_k}(\mathbf{Y}_i) \cap G_{w_k}(\hat{\mathbf{Y}}_i)\right|}{\left|G_{w_k}(\mathbf{Y}_i) \cup G_{w_k}(\hat{\mathbf{Y}}_i)\right|}, \tag{14}$$

$$\text{is_similar}\left(\mathbf{Y}_i, \hat{\mathbf{Y}}_i\right) = \frac{1}{K}\sum_{k=1}^{K}\text{Similarity}_{\text{S}}^k. \tag{15}$$

H Details of Perception Reward

Preference of Two Descriptions. In our implementation, we adopt Qwen2.5-Omni as the MLLM, owing to its superior performance in emotion preference decoding [35]. The prompt used is as follows: *We provide two descriptions for a given input. Please determine which one is better aligned with the input content. If both descriptions are equally aligned with the input content, please output "tie". Directly output the answer without additional reasoning.*

Ranking of Multiple Descriptions. Suppose there are N descriptions $\{d_i\}_{i=1}^{N}$, where each description d_i is associated with a positive parameter θ_i quantifying its relative advantage. The probabilities that description d_i wins, loses, or ties against description d_j are defined as follows:

$$P(d_i > d_j) = \frac{\theta_i}{\theta_i + \theta_j + \beta},\ P(d_i < d_j) = \frac{\theta_j}{\theta_i + \theta_j + \beta},\ P(d_i \sim d_j) = \frac{\beta}{\theta_i + \theta_j + \beta}, \tag{16}$$

where β controls the likelihood of ties. The parameters θ can then be estimated by maximizing the log-likelihood function:

$$\max \mathcal{L}(\theta) = \prod_{i<j} \left(\frac{\theta_i}{\theta_i + \theta_j + \beta}\right)^{\text{win}_{ij}} \left(\frac{\theta_j}{\theta_i + \theta_j + \beta}\right)^{\text{lose}_{ij}} \left(\frac{\beta}{\theta_i + \theta_j + \beta}\right)^{\text{tie}_{ij}}, \tag{17}$$

where win_{ij}, lose_{ij}, and tie_{ij} are indicator variables denoting the preference between d_i and d_j, determined by the MLLM in the preceding paragraph. Finally, we rank the descriptions based on their relative strengths θ_i. A higher value indicates better alignment with the input content.

I Baseline Details

SECap [66]. This is a speech emotion captioning framework designed to describe speech emotions using natural language. Specifically, it employs HuBERT as the audio encoder to extract speech representations and then uses the Bridge-Net to obtain emotion-relevant speech features. Finally, these features are fed into LLMs, which generate natural language captions describing speech emotions.

SALMONN [61]. This model is capable of processing both speech and audio. Specifically, it employs Whisper [52] to encode speech and BEATs [5] to encode audio. The resulting features are then integrated into the LLM via a window-level Q-Former. Additionally, a LoRA adapter is applied to enhance the LLM's instruction-following and question-answering capabilities.

Qwen-Audio [7]. This model is capable of handling 30 different tasks and a variety of audio types, demonstrating universal audio understanding capabilities. Specifically, it employs an audio encoder to map input audio into hidden features. To address interference between different tasks, it adopts a multi-task input format. The output features are then passed to LLMs for answer generation.

Otter [21]. This model is built upon OpenFlamingo [1] and trained on the MIMIC-IT dataset [21], demonstrating strong capabilities in multimodal perception, reasoning, and in-context learning.

Chat-UniVi [19]. This is a unified vision-language model capable of processing both images and videos. It employs dynamic visual tokens that capture spatial details in images and temporal relationships in videos.

Video-LLaVA [39]. It aligns images and videos before projection, enabling the LLM to learn from a unified visual representation and equipping it with the ability to comprehend images and videos.

Video-ChatGPT [46]. This model averages frame-level features across temporal and spatial dimensions, extracting both spatial and temporal video features. These features are then projected into the text token space, enabling the LLM to generate responses.

VideoChat [23]. It uses video foundation models to encode videos as embeddings, then employs LLMs for response generation. After instruction fine-tuning, the model demonstrates promising capabilities across various video applications.

VideoChat2 [24]. This model leverages diverse instruction-tuning data and employs a three-stage progressive multimodal training approach, comprising vision-language alignment, vision-language connection, and instruction tuning. As a result, it outperforms leading models across multiple tasks.

LLaMA-VID [26]. This model represents each image with two tokens: a context token and a content token. The context token encodes the global image information based on user input, while the content token captures visual details within the image. By condensing each image into just a few tokens, the model can efficiently understand and process long videos.

mPLUG-Owl [67]. This approach equips LLMs with vision-language capabilities through a two-stage training paradigm, enabling strong performance on instruction-following, visual understanding, and reasoning tasks.

PandaGPT [59]. It uses ImageBind [13] to map multiple modalities into a shared embedding space, then employs a projector to transform these features into the LLM's input space, and finally leverages the LLM to generate responses.

OneLLM [15]. This model aligns eight modalities with language through a unified framework. Its architecture consists of a universal encoder, a universal projection module, and modality-specific tokens. Additionally, it employs a progressive multimodal alignment pipeline, enabling it to comprehend eight input types and generate responses based on user queries.

R1-Omni [75]. This framework also leverages reinforcement learning for emotion recognition. Unlike AffectGPT-RL, which focuses on open-vocabulary emotions, this work targets discriminative emotions. Due to the different objectives, the reward functions used in R1-Omni differ from those employed in AffectGPT-RL. Appendix E provides a more detailed comparison.

Emotion-LLaMA [6]. This framework follows the architecture of traditional MLLMs and consists of three key components: encoders, projectors, and LLMs. Specifically, it inputs text, audio, peak frames, and three levels of visual cues (i.e., local, temporal, and global representations) and employs modality-specific encoders to convert these inputs into corresponding hidden features. These features are then aligned into a shared dimensional space with text embeddings through a linear projection mechanism. Finally, the LLM integrates information from all modalities for emotion recognition and reasoning. Experimental results on benchmark datasets demonstrate that this framework achieves improved performance over previous solutions.

AffectGPT [30]. This framework differs from previous MLLMs, which typically delegate the entire multimodal fusion process to the language model. Instead, it relocates cross-modal interaction outside the language model and incorporates a pre-fusion operation to enhance multimodal integration, emphasizing the inherently multimodal nature of human emotions. Specifically, the framework introduces two types of pre-fusion mechanisms: Q-Former and attention mechanisms. Both modules are computationally efficient and introduce only a small number of trainable parameters. Experimental results demonstrate the effectiveness of AffectGPT in generalized emotion understanding, including sentiment analysis, basic emotion recognition, and fine-grained emotion detection.

J Dataset Details

J.1 Training Datasets

MER-Caption+ [30]. It is a large-scale descriptive emotion dataset comprising approximately 31K samples. To reduce annotation costs, it introduces a novel *model-led human-assisted* strategy that leverages human priors to guide both description generation and sample filtering. Specifically, in the description generation stage, the pre-trained models are carefully selected to maximize performance on fine-grained emotion detection. During sample filtering, a two-level filtering process is employed to remove samples with mismatched audio and video, abnormally long descriptions, and those that perform poorly on sentiment analysis and basic emotion recognition.

MER2025-OV [31]. It is specifically designed for OV-MER. Unlike the automatically annotated MER-Caption+, MER2025-OV is a *high-quality* dataset created through a purely manual annotation process. It employs a multi-round annotation strategy. Initially, a pool of candidate emotions is provided, and four experts in affective computing are asked to select the correct labels and add any missing ones. All selected labels are combined to form the candidate set for the next round. In the second round, four annotators perform another round of labeling, and only labels selected by at least two annotators are retained. This two-stage approach ensures both completeness, by accepting labels chosen by any annotator in the first round, and accuracy, by filtering for consensus in the second. Through this rigorous process, MER2025-OV guarantees the high quality of its annotated labels.

J.2 Testing Datasets

OV-MERD+ [30]. OV-MERD+ is a dataset specifically designed for OV-MER, which is an extended version of OV-MERD [32]. The original OV-MERD dataset consists of 332 samples, all of which are fully annotated by humans for open-vocabulary emotions. Following a similar annotation process, OV-MERD+ expands the dataset to a total of 532 samples. In this study, we employ OV-MERD+ as the test set to evaluate the performance of different models on open-vocabulary emotion detection.

MER-UniBench [30]. MER-UniBench is a benchmark that encompasses three representative tasks: sentiment analysis, basic emotion recognition, and fine-grained emotion detection. Specifically, sentiment analysis focuses on binary classification into positive and negative sentiments. Basic emotion recognition restricts the label space to a predefined set of basic emotion categories. In contrast, fine-grained emotion detection removes these constraints, allowing predictions across any category. This benchmark is designed to assess generalized emotion understanding capabilities. Table 6 summarizes the data composition for each task.

Table 6: **MER-UniBench.** This table presents the data components for sentiment analysis, basic emotion recognition, and fine-grained emotion detection. Totally, it includes 12,799 test samples.

Raw Dataset	Selected Subset	Number of Samples	Emotion Space
Sentiment Analysis			
CMU-MOSI [70]	Test	686	*positive, negative*
CMU-MOSEI [71]	Test	4,659	*positive, negative*
CH-SIMS [69]	Test	457	*positive, negative*
CH-SIMS v2 [43]	Test	1,034	*positive, negative*
Basic Emotion Recognition			
MER2023 [33]	MER-MULTI	411	*worry, happy, neutral, angry, surprised, sad*
MER2024 [34]	MER-SEMI	1,169	*worry, happy, neutral, angry, surprised, sad*
IEMOCAP [3]	Session5	1,241	*anger, happiness, sadness, neutral*
MELD [50]	Test	2,610	*anger, joy, sadness, neutral, disgust, fear, surprise*
Fine-grained Emotion Detection			
OV-MERD+ [30]	All	532	*unrestricted label space*

J.3 Data Leakage Prevention

To prevent data leakage, we perform a rigorous verification to ensure no overlap exists between the training and testing data. During this process, we identify 200 overlapping samples between MER2025-OV (training set) and OV-MERD+ (test set). Consequently, we remove these samples from MER2025-OV, resulting in 1,000 samples available for training (see Table 1). Apart from this instance, we observe no other overlap between the training and testing sets. This procedure ensures that no data leakage exists between the datasets used for training and testing.

J.4 Task Clarification

This paper focuses on *utterance-level* emotion recognition, as opposed to *dialogue-level* emotion recognition. The key difference between these tasks lies in whether conversational history is taken into account. For *utterance-level* emotion recognition, the input consists of a single video clip centered on one character [62]. In contrast, *dialogue-level* emotion recognition requires leveraging

additional dialogue information [49]. This distinction has led to the development of two types of benchmarks: the *utterance-level* benchmark [36] and the *dialogue-level* benchmark [51].

In MER-UniBench, although certain datasets (e.g., MELD and IEMOCAP) contain dialogue information, others do not. Furthermore, even when dialogue context is available in datasets like MELD and IEMOCAP, its inclusion is not necessary and depends on the specific experimental setup [62, 49]. Consequently, MER-UniBench is designed as an *utterance-level* benchmark, and this work adopts the same setting to ensure fair comparison with prior studies [30]. Dialogue-level emotion recognition is beyond the scope of this paper and will be explored in the future.

K Role of RL across Different Architectures

This paper adopts Qwen2.5-7B-Instruct as the default LLM. In Table 7, we replace Qwen2.5-7B-Instruct with various alternative LLMs to investigate the role of RL across different architectures. Experimental results are presented in Table 7. The second column reports the SFT baseline results, while the third column shows the results after applying RL. These results demonstrate that reinforcement learning yields consistent performance gains across different model architectures, underscoring the effectiveness of RL training for tackling open-vocabulary emotion recognition.

Table 7: **Role of RL across different architectures.** To evaluate the generalization of RL, we replace the default Qwen2.5-7B-Instruct backbone in AffectGPT-RL with various alternative LLMs. Performance is assessed on OV-MERD+ to quantify the impact on open-vocabulary emotion recognition.

LLM Backbone	SFT	RL	Δ Gain
Qwen2.5-7B-Instruct	62.52	68.39	+5.87
Qwen2-7B-Instruct	59.63	67.02	+7.39
Baichuan2-7B-Chat	61.35	66.56	+5.21
Meta-Llama-3-8B-Instruct	62.98	66.24	+3.26

L Impact of Training Data Scale in RL

In the main experiments, we use MER2025-OV for reinforcement learning (see Table 1). To investigate the impact of data scale, we additionally annotated 1,000 open-vocabulary emotion samples, termed OV-New, and combined them with MER2025-OV to construct an extended training set of 2,000 samples. The corresponding experimental results are presented in Table 8. From this table, we observe that increasing the amount of RL data leads to limited performance gains, indicating that the performance of RL has reached its upper bound.

Table 8: **Impact of training data scale in RL.** We investigate whether increasing the volume of RL training data yields continued performance improvements.

Dataset	# Samples	Performance on OV-MERD+
MER2025-OV [31]	1,000	68.39
MER2025-OV + OV-New	2,000	68.91

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are well-aligned with the paper's actual contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The limitation is discussed in Appendix A.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present any theoretical results requiring formal assumptions or mathematical proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In this paper, we have made every effort to ensure the reproducibility of our work. In the supplementary material, we provide our source code and command lines for cold-start training, reinforcement learning, and inference. Additionally, we include a detailed

README file. To comply with NeurIPS's anonymity requirements, we have removed all external links and owner names. After acceptance, we will release the full code on GitHub to facilitate further research.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: In the supplementary material, we provide our source code and command lines for cold-start training, reinforcement learning, and inference. After acceptance, we will release the full code on GitHub to facilitate further research.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all relevant experimental details necessary to understand and reproduce the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 4.5 reports the statistical results for single-run and multi-run settings. We observe that AffectGPT-RL exhibits good stability across different runs. These results confirm that our conclusions are not artifacts of random chance, demonstrating the reliability of our experimental analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 3 provides our compute resources.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conducted in the paper conforms to the NeurIPS Code of Ethics in all respects. Appendix C provides detailed ethical statements.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix D discusses broader impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any models or data with high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used in the paper, including pre-trained models and datasets, are properly cited with references to their original publications. The terms of use for each asset are respected, and our code is released under the MIT license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: In the supplementary material, all data and code are well-documented in a README file.

    Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or external human subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve human subjects in a manner that requires IRB or ethics board approval. No direct interaction with participants occurred, and no personally identifiable information was collected.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.