Information Fusion 136 (2026) 104530

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Hinting the unknown: Effective open-set multimodal emotion recognition with a hierarchical cross-modal emotion-interactive prompting approach

Yuanyuan Liu [a], Shuyang Liu [a], Jiahao Zhang [a], Ke Wang [a], Chang Tang [b], Dapeng Tao [c,*], Zhe Chen [d,*], Wei Xiang [d]

[a] School of Computer Science, China University of Geosciences (Wuhan), Wuhan, Hubei, 430074, China
[b] School of Software Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China
[c] School of Information Science and Engineering, Yunnan University, Kunming, Yunnan, 650500, China
[d] The School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne, VIC, 3086, Australia

ARTICLE INFO

Keywords:
Open-set recognition
Multimodal emotion recognition
Prompt learning
Pretrained foundation models
Cross-modal alignment

ABSTRACT

Human emotion recognition is essential for emotionally intelligent systems in areas such as human-computer interaction, healthcare, and behavioral analysis. In real-world scenarios, emotional states are often ambiguous and fall outside predefined categories, making existing models unable to handle unseen emotion classes, thus limiting their use in sensitive domains like depression detection. To address this, open-set emotion recognition (O-ER) methods with pretrained models have been proposed to classify known emotions while detecting unknown ones. However, most focus on unimodal recognition, ignoring rich cues in multimodal data and limiting robustness. This motivates open-set multimodal emotion recognition (O-MER), which captures multimodal information to better recognize known emotions and detect unseen ones. However, directly fusing modalities often leads to semantically heterogeneous emotion representations, highlighting the need for a unified modality-interactive prompting mechanism to structurally align and fuse cross-modal affective cues. We propose Hierarchical Cross-modal Emotion-interactive Prompting (HCEP), a novel framework that adapts pretrained models for O-MER via a two-level prompt learning mechanism. HCEP includes three components: (1) Semantic-level Multimodal Emotion-aligning Prompting (SMEP), aligning and capturing multimodal emotion features via semantic emotion-unified prompts; (2) Class-level Unimodal Emotion-opposing Prompting (CUEP), refining decision boundaries per modality with positive and negative prompts; and (3) Dual-stream Prompt-driven Open-set Learning (DPOL), jointly optimizing known emotion classification and unknown emotion detection using a novel threshold-based discrimination strategy. Extensive experiments on five O-MER benchmarks demonstrate that HCEP achieves state-of-the-art results, outperforming the baseline with an average 16.79% AUROC and 28.49% OSCR relative gains, validating its strong capability to recognize known and detect unseen emotions.

1. Introduction

Emotion recognition aims to integrate emotion-related cues from video, audio, and text to accurately identify human affective states [1]. Traditional approaches treat it as a closed-set problem, assuming that all emotion categories encountered during testing are already seen during training. Although effective in controlled settings, this assumption limits real-world applicability, where emotional expressions are often ambiguous, fluid, and diverse. In practical applications such as depression detection or mental health monitoring, systems inevitably encounter rare or atypical affective states that do not fit neatly into predefined categories [2,3]. These may include deliberately masked emotions, culturally specific expressions, subtle micro-expressions, or mixed and transitional emotions [4]. Forcing such novel states into rigid categories can obscure diagnostic cues and increase the risk of misinterpretation.

To tackle this issue, open-set emotion recognition (O-ER) has emerged as a paradigm that not only classifies known emotions but also detects previously unseen ones [5]. In particular, an increasing number of O-ER approaches have been built upon pretrained foundation models, which are highly valued for their powerful representation capabilities and strong generalization [3,6]. For instance, UQ-SER [6] employs the pretrained audio model wav2vec2.0 [7] and various uncertainty quantification methods to enhance the model's ability to detect unknown emotions in audio data. HESP [3] effectively adapts the foundation model CLIP [8] for open-set facial expression recognition by introducing expression-sensitive prompts, enhancing its performance in

* Corresponding authors.
E-mail addresses: dptao@ynu.edu.cn (D. Tao), zhe.chen@latrobe.edu.au (Z. Chen).

https://doi.org/10.1016/j.inffus.2026.104530
Received 31 October 2025; Received in revised form 24 April 2026; Accepted 3 June 2026
Available online 4 June 2026

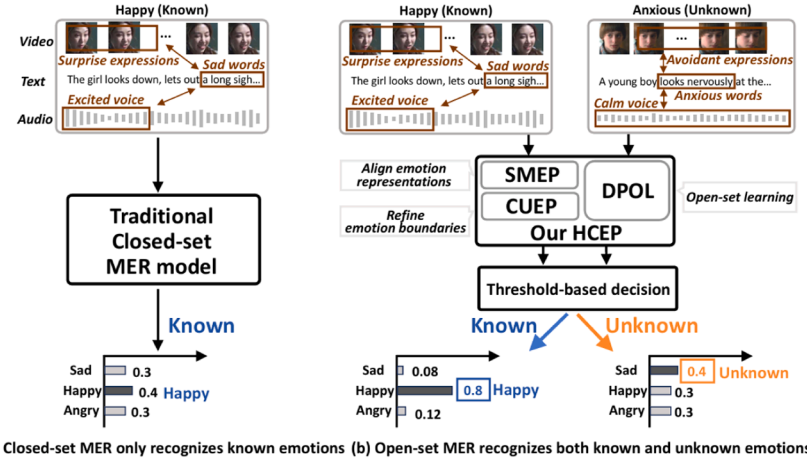

Fig. 1. Comparison of the inference process between previous methods and our method. (a) Traditional closed-set MER models can only fuse multimodal information to classify samples into predefined expression categories. (b) For the open-set MER task, the proposed HCEP employs a hierarchical cross-level prompting mechanism combined with a threshold-based decision strategy to algin and refine subtle multimmodal emotion cues, enabling effectively identifying both known and previously unseen emotions.

video-based O-ER. In addition, LMC [9] leverages pretrained models (*e.g.*, WordNet [10], DALL-E [11], CLIP [8], and DINO [12]) to effectively recognize both known and unknown classes in open-set tasks like O-ER.

However, most existing O-ER studies remain unimodal, focusing primarily on visual cues [2,3]. In real-world interactions, emotions are inherently multimodal, conveyed simultaneously through facial expressions, tone, and language [1]. Ignoring this cross-modal information risks incomplete or biased recognition, especially under open-set conditions with subtle or ambiguous emotions. To this end, ***in this paper, we extend the existing open-set emotion recognition task to the open-set multimodal emotion recognition (O-MER), aiming to capture multimodal emotional cues more comprehensively and further enhance recognition performance in open-set scenarios.*** Rather than introducing an entirely new task, our work builds upon the existing O-ER task by attempting to incorporate multimodal reasoning capabilities.

Despite the promising progress achieved by these foundational model-based methods in unimodal O-ER tasks, we found that the effectiveness of applying foundational models to the O-MER task could be significantly compromised due to ***multimodal emotion heterogeneity***. More specifically, as shown in Fig. 1, fine-grained emotions may manifest differently across modalities (*e.g.*, expressive face but neutral voice), making it difficult for the pretrained models like CLIP [8] or wav2vec2.0 [7] to learn unified emotion-relevant features. Therefore, a primitive combination of foundational models would not deliver appropriate multimodal emotion representations that could easily separate unknown emotions from known ones for O-MER, ultimately ***leading to blurred emotion boundaries for misclassification in O-MER***.

To address these challenges, our key insight is *to develop a **H**ierarchical **C**ross-modal **E**motion-interactive **P**rompting learning (**HCEP**) framework that adapts the pretrained foundation models across modalities for fine-grained alignment and unified affective representations, thereby enabling the effective discovery of unknown emotions in open-set scenarios.* Specifically, **HCEP** integrates three emotion-focused modules: (1) ***Semantic-level Multimodal Emotion-aligning Prompting (SMEP)***, which aligns and captures multimodal emotion features through unified semantic prompts, ensuring all modalities attend to consistent emotional cues; (2) ***The Class-level Unimodal Emotion-opposing Prompting (CUEP)***, which refines decision boundaries by generating contrastive positive and negative emotion prompt vectors, enhancing discriminative feature learning for both known and unknown emotion categories; and (3) ***Dual-stream Prompt-driven Open-set Learning (DPOL)***, employs a dual-stream open-set recognition scheme to jointly improve known emotion classification and unknown emotion detection. Finally, this unified design allows HCEP to effectively separate known and unknown multimodal emotion spaces and significantly outperforms previous work by introducing a threshold-based decision strategy. Unlike prior works that apply prompting or open-set learning independently or in a loosely coupled manner, HCEP introduces a hierarchical cross-modal prompting paradigm that jointly models semantic alignment and class-level opposition under a unified dual-stream open-set framework.

Our main contributions are summarized as follows:

- We analyze the challenges inherent in the O-MER task and propose **HCEP**, the first framework specifically designed for this task, enabling robust open-set performance across modalities.
- HCEP introduces a two-level cross-modal emotion prompting strategy that enhances the pretrained models' ability to capture unified and discriminative multimodal emotion representations. Specifically, the **SMEP** aligns cross-modal semantic cues for emotion-unified representations, while the **CUEP** refines class-level emotion boundaries through contrastive positive and negative prompts.
- Furthermore, the **DPOL** proposes a dual-stream learning scheme with a novel distance-based emotion discrimination mechanism that fuses semantic- and class-level prompt information, yielding finer decision boundaries and jointly improving known-class recognition and unknown-class detection.
- We establish five O-MER task settings for comprehensive evaluation. HCEP consistently outperforms baseline methods, achieving average relative improvements of **16.79%** in AUROC and **28.49%** in OSCR across all settings.

2. Related work

2.1. Open-set facial expression recognition

This task focuses on classifying known facial expressions while detecting unknown ones. Owing to the subtle differences between expression classes, standard open-set methods often struggle to achieve satisfactory performance. To address this, researchers have proposed various specialized solutions. For instance, Open-set FER [2] differentiates between known and unknown emotions by evaluating the consistency of attention maps before and after image augmentation. HESP [3] introduces a human expression-sensitive prompting mechanism to enhance CLIP's [8] ability for capture fine-grained emotional cues, thereby improving the discrimination between known and unknown emotion categories. OpenFE [13] leverages attention mechanisms to emphasize critical facial regions and employs reconstruction to extract low-dimensional

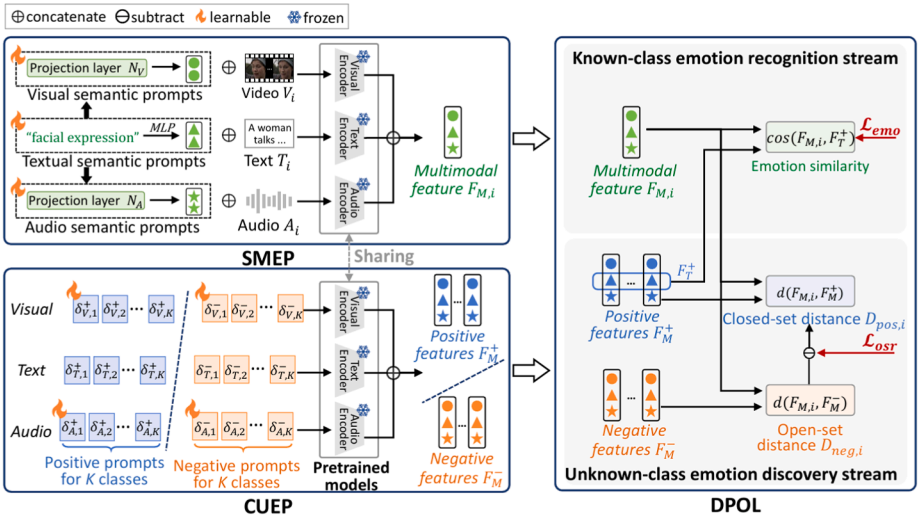

Fig. 2. HCEP first uses SMEP to align multimodal emotional features via emotion-unified prompts. CUEP then aligand refines decision boundaries with two contrastive emotion prompts, and DPOL enables known emotion classification and unknown detection through dual-stream open-set recognition.

latent features, thereby enriching known-class representations and significantly enhancing the discrimination between known and unknown facial categories. However, these vision-specific methods are not directly applicable to the more complex O-MER setting, which requires effective cross-modal alignment and interaction.

2.2. Prompt learning

Prompt learning has attracted significant research attention due to its ability to leverage the latent knowledge of pretrained foundation models and adapt them to downstream tasks [14–16]. For instance, PA-SAM [17] mitigates the segmentation quality limitations of the foundation model SAM [18] through a novel prompt-driven adapter architecture that enhances mask prediction accuracy. By explicitly modeling motion cues and generating dynamic prompts, Wang et al. [19] adapt the CLIP model to action recognition tasks, significantly enhancing both efficiency and generalization capability. MVLPT [20] integrates cross task knowledge into prompt tuning to enhance CLIP's adaptability to multiple downstream tasks. However, current prompt learning methods mainly focus on general visual understanding tasks and lack specialized methods designed for multimodal emotion related tasks.

2.3. Multimodal emotion recognition

Multimodal emotion recognition aims to identify human emotions by integrating cues from text, vision, and speech, with cross-modal alignment being one of the central challenges [21]. Wu et al. [22] designed a sophisticated fusion network based on multi-head attention mechanisms to dynamically assign weights to features from different modalities, thereby effectively enhancing recognition performance. MPT-HCL [23] introduced multimodal prompts and hybrid contrastive learning to handle noisy inputs and few-shot labels. GADN [24] addresses multimodal emotion recognition via a Multimodal Interactive Gate and graph-based distillation, improving the integration of heterogeneous multimodal features. Although existing methods are effective for closed-set recognition, they cannot detect unknown emotion classes encountered in O-MER tasks, inevitably leading to misclassification.

3. Method

3.1. Problem definition

We begin by formally introducing the O-MER problem. Following the existing formulation of open-set emotion recognition (O-ER) [3,5], each training sample is a multimodal tuple $(T_i, V_i, A_i, y_i) \in D_{\text{train}}$, where T_i, V_i, A_i, and y_i represent the text, visual frames, audio data, and emotion label of the i-th sample, respectively. The label y_i belongs to a set of known classes, denoted as $y_i \in Y_{\text{known}} = \{1, 2, \ldots, K\}$, where K is the number of known emotion categories. The test stage includes novel emotion categories that were unseen during training. We define all these unseen categories as unknown emotions, collectively denoted as $Y_{\text{unknown}} = \{K + 1\}$. Consequently, the test set is $D_{\text{test}} = \{(T_j, V_j, A_j, y_j)\}$, where $y_j \in Y_{\text{known}} \cup Y_{\text{unknown}}$, and j indexes a test sample. The goal of this task is to design a model capable of accurately identifying the category of each multimodal test sample from the combined set $Y_{\text{known}} \cup Y_{\text{unknown}}$.

3.2. Overview of HCEP

Fig. 2 illustrates the HCEP framework, comprising three modules: Semantic-level Multimodal Emotion-aligning Prompting (SMEP), Class-level Unimodal Emotion-opposing Prompting (CUEP), and Dual-stream Prompt-driven Open-set Learning (DPOL).

Firstly, to tackle emotion heterogeneity in multimodal data, the SMEP module introduces three learnable, emotion-unified, cross-modal prompts (textual, visual, and audio), aligning multimodal emotion semantic cues for emotion-unified multimodal representation. ***Secondly***, to address blurred class decision boundaries between fine-grained emotions, the CUEP module introduces two opposing cross-modal emotion prompting strategies: positive and negative emotion prompts to refine class-specific emotion cues for each modality. The "positive" prompts reinforce relevant emotion characteristics, and the "negative" prompts help to identify non-matching (*not* belong to a class) for contrasting with "positive" prompts, thereby improving boundary modeling for emotion classes. ***Finally***, the DPOL module encourages the interaction of semantic-level (SMEP) and class-level (CUEP) prompts, optimizing known-class recognition and unknown-class discovery via a dual-stream learning strategy with a novel distance-based emotion discrimination method. These modules are not independent components but are jointly optimized within a unified hierarchical prompting framework, enabling coordinated modeling of cross-modal semantic alignment and class-level discrimination.

3.3. SMEP for emotion-unified learning

O-MER is challenging due to subtle and heterogeneous cross-modal emotion semantics, even with powerful pretrained models. To address

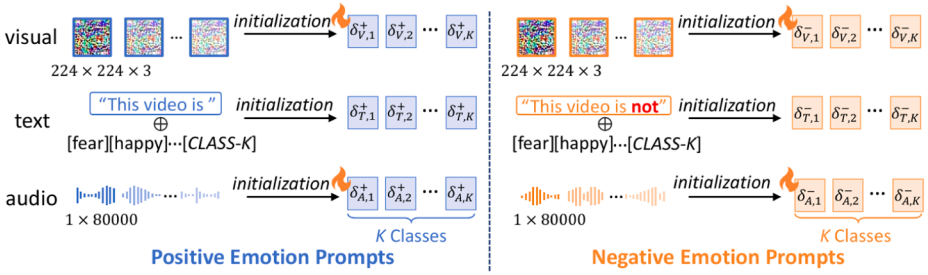

Fig. 3. Detailed design of positive and negative emotion prompts in the CUEP module.

this issue, we draw inspiration from cognitive science, where language is often regarded as a high-level semantic anchor in human emotion understanding. Compared to visual and auditory signals, which are typically implicit and susceptible to noise, textual modality can convey more direct and explicit affective semantics. Motivated by this observation, we introduce the Semantic-level Multimodal Emotion-aligning Prompting (SMEP) module, which initializes ***textual emotion semantic prompts*** using emotion-related words (*e.g.*, "facial expression") and leverages them to explicitly guide the pretrained models toward emotion-unified features. Specifically, these prompts are projected into visual and audio modalities via cross-modal prompt projection, forming emotion-unified ***visual and audio emotion semantic prompts***. By concatenating them with raw inputs and feeding them into frozen pretrained encoders, SMEP enhances cross-modal alignment for the obtained multimodal features, enabling robust adaptation to O-MER tasks.

Textual emotion semantic prompts. Since textual prompts offer direct and structured emotion semantics, making them ideal for cross-modal guidance, we first initialize these prompts with explicit emotion-related phrases "*facial expression*". During training, these prompts are dynamically updated to guide the pretrained model in capturing emotion-related semantics from the textual modality, formulated as:

$$\delta_T = MLP\big(\mathcal{T}\big(\text{"facial expression"}\big)\big) = \{t_1, t_2, \ldots, t_m\}, \tag{1}$$

where $t_1, t_2, \ldots, t_m \in \mathbb{R}^{1\times 512}$ are learnable text prompt vectors and m is the pre-defined number of them. In this study, we set m to 2 and present a detailed experimental analysis in Sec 4.3.2. Following CLIP's [8] encoding paradigm, $\mathcal{T}$ represents the tokenization operation that converts input words into discrete tokens, while MLP projects these tokens into continuous word embedding vectors.

Visual and audio emotion semantic prompts. We introduce two learnable cross-modal prompt projection layers, *i.e.*, the visual projection layer N_V and the audio projection layer N_A, to embed the textual semantic emotion prompts δ_T into visual semantic emotion prompts δ_V and audio semantic emotion prompts δ_A, respectively. This process aligns emotion semantics across all modalities, forming emotion-unified prompts. Mathematically, this process can be formulated as:

$$\begin{cases} \delta_V = \{N_V(t_1), N_V(t_2), \ldots, N_V(t_m)\}, \\ \delta_A = \{N_A(t_1), N_A(t_2), \ldots, N_A(t_m)\}, \end{cases} \tag{2}$$

where $N_V(\cdot)$ and $N_A(\cdot)$ are visual and audio projection operation, respectively. Thus, we obtain a group of emotion-unified semantic prompts as $(\delta_T, \delta_V, \delta_A)$.

Prompt-enhanced features based on SMEP. Using the $(\delta_T, \delta_V, \delta_A)$, we employ three frozen pretrained encoders: the CLIP text encoder ϕ_T [8], CLIP visual encoder ϕ_V [8], and wav2vec 2.0 audio encoder ϕ_A [7], to extract and fuse the emotion-unified prompts with the corresponding unimodal input (T_i, V_i, A_i), as follows:

$$F_{T,i} = \phi_T[\delta_T, T_i], \quad F_{V,i} = \phi_V[\delta_V, V_i], \quad F_{A,i} = \phi_A[\delta_A, A_i], \tag{3}$$

where $F_{T,i}$, $F_{V,i}$, $F_{A,i}$ are the prompt-enhanced textual, visual and audio features of i-th sample, respectively. $[,]$ denotes the concatenation operation. Finally, the SMEP-prompted multimodal features $F_{M,i}$ can be concatenated as: $F_{M,i} = [F_{T,i}, F_{V,i}, F_{A,i}]$.

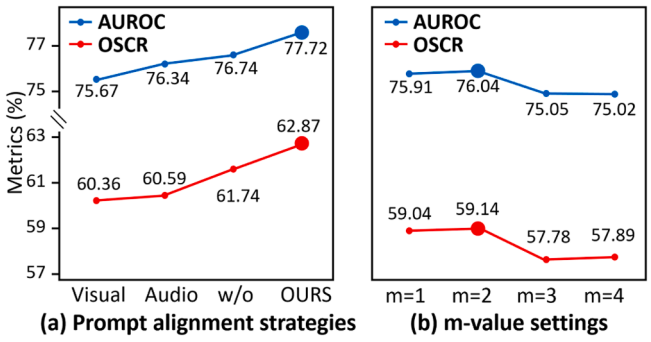

Fig. 4. Different prompt settings in SMEP.

3.4. CUEP for emotion-discriminative learning

To further refine the decision boundaries between known and unknown emotion classes in O-MER, we propose the Class-level Unimodal Emotion-opposing Prompting (CUEP) module to generate more discriminative emotion representations per modality in pretrained foundational models. Unlike prior methods such as ARPL [25] and HESP [3], which mainly emphasize negative emotion feature learning to define open-set boundaries (we refer readers to the papers [3,25] for more details), our CUEP module introduces a more comprehensive strategy to jointly exploit both positive and negative emotion features to enhance the separation of known and unknown emotions.

CUEP is composed of two contrasting prompting components: (1) ***Positive emotion prompts***, which highlight class-specific emotion patterns for known class learning. (2) ***Negative emotion prompts***, which capture emotion-irrelevant or contrasting cues to separate unknown emotional states. This dual-guidance enhances fine-grained discrimination and strengthens decision boundaries for O-MER.

Positive emotion prompts for per modality. To enhance the ability of pretrained models to capture the discriminative emotional knowledge of each known category, we design positive emotion prompts for each modality. From a cognitive science perspective, language provides clear and explicit emotion semantics, while visual and auditory signals are more implicit, noisy, and highly variable. Therefore, we use fixed natural language templates for textual prompts, and adopt learnable continuous prompts for visual and audio modalities to flexibly capture their complex patterns. Specifically, as illustrated in Fig. 3, for each known emotion category k ($k \in [1, 2, \ldots, K]$), we devise positive emotion prompts as $\delta^+_{M,k} = (\delta^+_{T,k}, \delta^+_{V,k}, \delta^+_{A,k})$. *For the text modality*, the textual positive emotion prompts $\delta^+_{T,k}$ follow a predefined template: "*This video is [CLASS-k]*", where *[CLASS-k]* represents the emotion label (*e.g.*, "happy", "angry"). *For visual and audio modalities*, we initialize two trainable tensors, visual positive prompts $\delta^+_{V,k} \in \mathbb{R}^{224\times 224\times 3}$ and audio positive prompts $\delta^+_{A,k} \in \mathbb{R}^{1\times 80000}$, with random noise, matching the dimensions of the corresponding inputs. These cross-modal positive emotion prompts are iteratively optimized during training to capture discriminative emotion patterns for the known class k.

Negative emotion prompts for per modality. Negative representations, opposite to known class representations, are particularly beneficial for discovering unknown patterns in open-set tasks [3,25]. In O-ER tasks, data with patterns can be collected from negative representations of all categories to discover unknown emotion categories. Motivated by this, we propose negative emotion prompts across modalities to help the pretrained models further explore unknown emotion patterns.

Similar to the design of positive prompts, we construct textual prompts using fixed natural language templates, while modeling visual and audio prompts as learnable continuous vectors. Formally, we introduce a set of K distinct negative emotion prompts, as $\delta_M^- = \{\delta_{M,1}^-, \delta_{M,2}^-, \dots, \delta_{M,K}^-\}$. Each represents the class-specific negative emotion prompts, $\delta_{M,k}^- = (\delta_{T,k}^-, \delta_{V,k}^-, \delta_{A,k}^-)$, which help discover cross-modal emotion patterns that do NOT belong to the specific known emotion category k. *For the text modality*, as shown in Fig. 3, we explicitly construct the textual negative emotion prompt $\delta_{T,k}^-$ as "*This video is not [CLASS-k]*", where *[CLASS-k]* enumerates all the closed-set emotional categories. *For the visual and audio modalities*, we define the visual negative emotion prompt $\delta_{V,k}^-$ and audio negative emotion prompt $\delta_{A,k}^-$, as learnable noise tensors matching the dimensions of input images and audio sequences.

Prompt-enhanced positive/negative features based on CUEP. With the positive/negative emotion prompts, we also employ the frozen pretrained models, CLIP [8] and wav2vec 2.0 [7], to encode the two contrast prompts, which can be given by

$$\begin{cases} F_{T,k}^+ = \phi_T(\delta_{T,k}^+), & F_{V,k}^+ = \phi_V(\delta_{V,k}^+), \quad F_{A,k}^+ = \phi_A(\delta_{A,k}^+), \\ F_{T,k}^- = \phi_T(\delta_{T,k}^-), & F_{V,k}^- = \phi_V(\delta_{V,k}^-), \quad F_{A,k}^- = \phi_A(\delta_{A,k}^-), \end{cases} \tag{4}$$

where $k \in K$. Next, these unimodal positive and negative emotion features are fused through a simple concatenation operation $[,]$, obtaining the prompted positive emotion features $F_{M,k}^+ = [F_{T,k}^+, F_{V,k}^+, F_{A,k}^+]$ and negative emotion features $F_{M,k}^- = [F_{T,k}^-, F_{V,k}^-, F_{A,k}^-]$, for the class k. Finally, in CUEP, we obtain two groups of prompted-enhanced positive/negative features: F_M^+ and F_M^-, for all emotion classes. These contrastive prompted features improve fine-grained emotion discrimination and sharpen known/unknown decision boundaries, thereby enabling robust O-MER performance.

3.5. DPOL for open-set recognition

Using the aforementioned prompt-enhanced features, we further introduce the Dual-stream Prompt-driven Open-set Learning (DPOL) mechanism to enhance prompt-guided O-MER. DPOL includes two collaborative streams: (1) ***a known-class emotion recognition stream*** focuses on reinforcing intra-class consistency and sharpening decision boundaries among known categories using SMEP-prompted features and CUEP-prompted positive textual features, and (2) ***an unknown-class emotion discovery stream*** that promotes divergence from known classes via SMEP-prompted multimodal features and CUEP-prompted negative and positive features, enabling the detection of emotion-irrelevant or anomalous patterns that may indicate unknown categories.

In addition, in DPOL, we propose a novel **distance-based emotion discrimination method** to unify both task streams by modeling relationships between multimodal features and positive/negative emotion features. This anchors known emotions while pushing unknowns to distinct embedding regions, enabling a discriminative feature space that balances recognition accuracy and open-set robustness.

Known-class Emotion Recognition Stream. This task stream leverages an emotion verification loss to effectively recognize known emotion classes while aligning prompted multimodal and positive features to enhance synergy between SMEP's unified and CUEP's positive emotion prompts.

With the semantic emotion-unified prompts, we would obtain a similar design with the CLIP [8], where the CUEP-prompted positive textual features F_T^+ would contain comprehensive emotion information across all K known emotion classes derived from the text prompt "*This video is [class-k]*". These features provide emotion supervisory information to guide SMEP in extracting more accurate SMEP-prompted multimodal feature $F_{M,i}$ from the i-th training sample. Using this feature, the emotion verification loss L_{emo} is constructed by computing the similarity relationships between $F_{M,i}$ and F_T^+ for all known emotion classes, where the similarity to the ground-truth class y_i is maximized through supervised learning. Formally, the loss is defined as:

$$\mathcal{L}_{emo} = \sum_i CE(softmax(cos(F_{M,i}, F_T^+)), y_i), \tag{5}$$

where $softmax$ is the softmax operation and cos is cosine similarity, y_i is the known class label for the i-th sample, and CE represents cross-entropy loss.

Unknown-class Emotion Discovery Stream. The unknown-class stream optimizes the distance relationship between SMEP-prompted multimodal features and CUEP-prompted positive/negative features to reinforce the interaction between SMEP's semantic emotion-unified prompts and CUEP's two contrast prompts. This improves distribution divergence between known and unknown emotions, boosting unknown-class emotion discovery performance.

To devise a learning stream optimized for discovering unknown classes, we first introduce a distance metric to estimate the relations between features from diverse classes and help distinguish known and unknown classes. Firstly, we devise a closed-set distance measurement $D_{pos,i}$ to calculate the distance between the i-th sample's SMEP-prompted multimodal feature $F_{M,i}$ and the CUEP-prompted positive features F_M^+ of K known classes, which is expressed as follows:

$$D_{pos,i} = d(F_{M,i}, F_M^+), \tag{6}$$

where d denotes the composite distance defined in ARPL [25] as the difference between the Euclidean distance and the dot product. Similarly, we devise an open-set distance measurement $D_{neg,i}$ to quantify the distance between $F_{M,i}$ and the CUEP-prompted negative features F_M^- of all known classes. This process can be expressed as:

$$D_{neg,i} = d(F_{M,i}, F_M^-). \tag{7}$$

Using the introduced distances, we design an open-set recognition loss L_{osr} to comprehensively optimize the decision boundaries between known and unknown classes. Formally, L_{osr} can be expressed as:

$$\mathcal{L}_{osr} = \sum_i CE(softmax(D_{neg,i} - D_{pos,i}), y_i). \tag{8}$$

The $D_{pos,i}$ measures closeness to positive features and $D_{neg,i}$ measures closeness to negative features. So, $D_{neg,i} - D_{pos,i}$ represents a contrastive distance score: higher values mean the sample is much closer to the positive feature of a specific known class and far from its negative counterpart. During training, this contrastive distance-based loss encourages the model to reduce the gap between a sample and its corresponding positive emotion representation while increasing its dissimilarity to the associated negative prompts. By applying a softmax over the distance difference vector, the model would learn to amplify discriminative cues between known and unknown emotional states, resulting in improved open-set detection.

Overall Loss. Based on the aforementioned dual-stream learning scheme, we formulate the overall optimization objective for our O-MER model as follows:

$$\mathcal{L} = \mathcal{L}_{emo} + \mathcal{L}_{osr}. \tag{9}$$

3.6. Distance-based inference and threshold-independent evaluation

Algorithm 1 presents the inference and evaluation procedure of HCEP. Given a multimodal sample (T_j, V_j, A_j), the SMEP module first generates emotion-aligned representations $F_{M,j}$ via cross-modal prompt projection. The CUEP module then introduces positive and negative prompts to produce class-level discriminative features F_M^+ and F_M^-. Based on these features, the DPOL module computes the distance-based logits $D_{neg,j} - D_{pos,j}$, from which the predicted label $\hat{y}$ and confidence

score are obtained. For open-set performance evaluation, AUROC and OSCR, as threshold-independent metrics, provide a fair and comprehensive assessment of model performance. Specifically, we sweep over all possible decision thresholds (*e.g.*, $\theta \in (0, 1)$) and compute the corresponding TPR, FPR, and CCR at each threshold to construct the evaluation curves, where TPR denotes the proportion of known samples correctly accepted as known, FPR denotes the proportion of unknown samples misclassified as known, and CCR denotes the proportion of known samples that are both accepted and correctly classified. Importantly, this threshold sweeping is used solely for computing threshold-independent metrics (AUROC and OSCR), and does not involve selecting any operating threshold using test-set labels. The final AUROC and OSCR values are computed as the area under the TPR-FPR and CCR-FPR curves, respectively. This threshold-independent protocol avoids reliance on any specific threshold and enables a fair and comprehensive comparison across different methods.

Algorithm 1 Distance-based inference and threshold-independent evaluation.

Require: Test set $\mathcal{D}_{test}$, trained textual emotion semantic prompts δ_T, cross-modal prompt projection layers N_V and N_A, positive and negative emotion prompts $\delta_M^+ = \{\delta_{M,1}^+, \delta_{M,2}^+, \ldots, \delta_{M,K}^+\}$, $\delta_M^- = \{\delta_{M,1}^-, \delta_{M,2}^-, \ldots, \delta_{M,K}^-\}$ and pretrained foundation models ϕ_T (CLIP-text), ϕ_V (CLIP-visual), ϕ_A (wav2vec 2.0).
Ensure: AUROC and OSCR
Step 1: Compute prediction scores for all test samples
1. **for:** each sample $(T_j, V_j, A_j) \in \mathcal{D}_{test}$ **do**
2. $\delta_V = N_V(\delta_T), \delta_A = N_A(\delta_T)$
3. $F_{T,j} = \phi_T[\delta_T, T_j],\ F_{V,j} = \phi_V[\delta_V, V_j],\ F_{A,j} = \phi_A[\delta_A, A_j]$
4. $F_{M,j} = [F_{T,j}, F_{V,j}, F_{A,j}]$
5. $F_M^+ = \phi(\delta_M^+), \quad F_M^- = \phi(\delta_M^-)$
6. $D_{pos,j} = d(F_{M,j}, F_M^+),\ D_{neg,j} = d(F_{M,j}, F_M^-)$
7. $p_j = \text{softmax}(D_{neg,j} - D_{pos,j})$
8. $confidence_j = \max(p_j)$
9. $\hat{y}_j = \arg\max(p_j)$
10. **end for**
Step 2: Threshold sweeping
11. **for** $\theta \in (0, 1)$ **do**
12. **for** each sample j **do**
13. if $confidence_j \geq \theta$: $\hat{y}_j$ # classified as known
14. else: *unknown*
15. **end for**
16. Compute TPR(θ), FPR(θ), CCR(θ)
17. **end for**
Step 3: Metric computation
18. AUROC ← area under TPR-FPR curve
19. OSCR ← area under CCR-FPR curve
20. **return** AUROC, OSCR

4. Experiments and analysis

4.1. Experiment setup

4.1.1. Datasets

In this study, we evaluated our method on three traditional multimodal sentiment datasets, *i.e.*, MAFW [26], CMU-MOSEI [27] and OV-MERD [28]. **MAFW** is a large-scale, multi-modal dataset designed for dynamic facial expression recognition (FER) in real-world scenarios. It contains 10,045 video-audio clips, each annotated with one or more of 11 commonly used emotion categories: anger, disgust, fear, happiness, neutral, sadness, surprise, contempt, anxiety, helplessness, and disappointment. These annotations yield both 11 single-label emotions and 32 compound emotion classes. **CMU-MOSEI** dataset is the largest benchmark for multimodal sentiment analysis and emotion recognition. It contains 23,453 annotated sentences from 1,000+ online speakers across 250 diverse topics, with annotations covering six emotion categories: anger, happiness, sadness, surprise, fear, and disgust. **OV-MERD** is a multimodal emotion recognition dataset crafted to transcend the constraints of conventional MER frameworks limited to predefined emotion taxonomies. In addition to the six basic emotion categories, *i.e.*, anger, disgust, fear, happiness, sadness, and surprise, OV-MERD employs an open-vocabulary annotation schema encompassing 236 emotion categories, with most samples annotated with 2 to 4 labels, thereby supporting more refined discrimination across a broader and more nuanced emotional spectrum.

4.1.2. O-MER task settings

Following the experimental setup of [3], we construct multiple O-MER tasks by partitioning emotion categories into known and unknown classes. To quantify task difficulty, we adopt the *openness* metric [29]:

$$O(K : U) = 1 - \sqrt{\frac{K}{K+U}},$$

where K and U denote the numbers of known and unknown emotion classes, respectively. A higher openness indicates a more challenging setting with a larger proportion of unknown classes.

Comprehensive evaluation under open-set scenarios. To comprehensively evaluate the open-set recognition performance of our model, we design five specific experimental settings. These settings cover two key factors:

(1) *Open-set novelty*, where different numbers of unknown emotion categories are introduced within the same dataset;
(2) *Distribution shift*, where data from different datasets are combined, introducing variations in recording conditions, demographics, and dataset biases, thereby further evaluating the robustness of the proposed method.

All settings are evaluated under the open-set recognition framework. In particular, the intra-dataset partition tasks focus on open-set novelty, while the cross-dataset setting further incorporates distribution shift to evaluate the model's robustness when facing unseen emotion categories under varying data distributions. The detailed task settings are described as follows:

1) **series O-MER with 10 emotions in MAFW (intra-dataset, open-set novelty).** We define four openness settings by randomly partitioning the 10 emotion categories into known and unknown classes: $O(8:2) = 0.11$, $O(6:4) = 0.23$, $O(4:6) = 0.37$, $O(2:8) = 0.55$.
2) **series O-MER with 6 basic emotions in CMU-MOSEI (intra-dataset, open-set novelty).** We construct four partition schemes: $O(5:1) = 0.09$, $O(4:2) = 0.18$, $O(3:3) = 0.29$, $O(2:4) = 0.42$.
3) **series O-MER with 32 compound emotions in MAFW (fine-grained open-set novelty).** To evaluate the capability of handling more complex and fine-grained emotions, we treat the 10 single-label emotions as known classes and the 32 compound emotions as unknown classes, resulting in a higher openness: $O(10:32) = 0.51$.
4) **series O-MER with 236 open-world emotions in OV-MERD (large-scale open-set novelty).** In the OV-MERD dataset, we use six basic emotion categories as known classes and the 236 emotion categories as unknown classes, leading to a highly challenging setting with $O(6:236) = 0.84$.
5) **series O-MER with cross datasets (distribution shift).** To simulate distribution shift, we combine MAFW and CMU-MOSEI. Specifically, we use six shared basic emotions as known classes and treat the remaining 4 categories of MAFW as unknown, resulting in $O(6:4) = 0.23$. Unlike the above intra-dataset settings, this task evaluates the robustness of the model under feature space shift, where both data distribution and modality characteristics may vary across datasets.

Note that (1) unknown emotion categories are removed from the training set; and (2) to reduce randomness, we evaluate four different known–unknown class splits per openness level and report their average results.

Table 1
Design details of the two-level cross-modal prompting in HCEP.

Module	Prompt Modality	Construction Strategy	Dimensionality	Optimization Scheme
CEUP	Text	Fixed Natural Language Template	/	Fixed
	Vision	Learnable Continuous Vectors	(3,224,224)	Updated via Backpropagation
	Audio	Learnable Continuous Vectors	(1,80000)	Updated via Backpropagation
SMEP	Text	Text Initialization ("facial expression")	(2, 512)	Updated via Backpropagation
	Vision	Cross-modal Projection	(2, 768)	Updated via Backpropagation
	Audio	Cross-modal Projection	(2, 8000)	Updated via Backpropagation

Table 2
Results with 10 emotions in MAFW under four openesses.

Type	Method	AUROC					OSCR				
		O(8:2)	O(6:4)	O(4:6)	O(2:8)	Mean	O(8:2)	O(6:4)	O(4:6)	O(2:8)	Mean
O-ER	ARPL [25]	54.91	52.18	57.69	63.77	57.14	12.62	19.44	23.81	45.24	25.28
	CSSR [32]	60.35	57.38	55.51	61.60	58.71	18.34	23.07	26.37	43.78	27.89
	Open-VCLIP [33]	47.52	52.98	54.22	55.85	52.64	18.16	27.77	33.55	44.73	31.05
	HESP [3]	71.47	66.43	68.46	70.98	69.34	38.04	39.87	49.72	59.77	46.85
O-MER	CPN* [34]	63.95	68.36	70.31	70.75	68.34	46.34	51.71	58.82	67.69	56.14
	LPL* [35]	58.76	60.25	61.04	53.44	58.37	40.84	45.04	50.75	53.17	47.45
	Baseline	64.71	66.68	65.91	63.98	65.32	42.21	46.93	52.29	59.32	50.19
	Ours	**80.12**	**75.86**	**79.02**	**75.89**	**77.72**	**54.55**	**56.99**	**66.50**	**73.43**	**62.87**

4.1.3. Evaluation protocols

We adopt two widely used threshold-independent metrics in open-set recognition, namely AUROC [30] (Area Under the Receiver Operating Characteristic Curve) and OSCR [31] (Open-Set Classification Rate). AUROC measures the model's ability to distinguish known samples from unknown samples, while OSCR jointly reflects the accuracy of known-class classification and the ability to reject unknown samples. Higher values indicate better recognition performance in open-set scenarios. Notably, both metrics evaluate model performance by sweeping over all possible decision thresholds, rather than relying on a specific optimal threshold. Importantly, this evaluation protocol does not require selecting any threshold using test-set labels, ensuring a leakage-free and fair comparison across different methods. This property enables a more comprehensive and fair assessment, allowing comprehensive performance evaluation and fair comparison across different methods.

4.1.4. Implementation details

The model was implemented on Ubuntu 20.04 using PyTorch and trained on an NVIDIA GeForce RTX 3090 GPU. Training used the SGD optimizer with momentum for 150 epochs, starting from a learning rate of 0.0005 that decayed by 0.1 every 30 epochs. For fair comparison, we followed the open-set inference protocols in [3,25].

Additionally, Table 1 provides a comprehensive overview of the two-level prompting design in HCEP, detailing the construction methods, modality-specific representations, dimensional settings, and optimization strategies.

To ensure a fair comparison, all adapted baseline methods are implemented using the same pretrained multimodal encoders (CLIP for text/vision and wav2vec 2.0 for audio) as our approach. All methods are trained and evaluated on identical data splits with the same openness settings. We adopt consistent training configurations and hyperparameter tuning strategies across all methods to avoid introducing bias. In addition, all methods are evaluated using the same threshold-independent protocol (AUROC and OSCR), ensuring fair performance comparison without relying on any specific decision threshold.

4.2. Overall performance

As the first method tailored for O-MER, we evaluate HCEP against four unimodal open-set methods (ARPL [25], CSSR [32], Open-VCLIP [33], HESP [3]) and three multimodal methods (CPN* [34], LPL* [35], and our baseline). Here, CPN*, LPL*, and the baseline denote their multimodal extensions of CPN, LPL, and ARPL, respectively, where the original feature extractors are replaced with pretrained encoders. To ensure fairness, all adapted baselines use the same frozen multimodal backbone (CLIP + Wav2Vec 2.0), identical data splits, and a unified evaluation protocol based on threshold-independent metrics (AUROC and OSCR). Only the feature extraction layers are modified for multimodal inputs, while their original loss functions and architectures are preserved. Tables 2–4 report the results, where the best and second-best performances are marked in **bold** and underlined, respectively. More detailed per-category analysis is provided in Fig. 9 (Section 4.4.7).

4.2.1. Experiments on 10 emotions in MAFW

In Table 2, we compare HCEP with state-of-the-art O-ER and O-MER methods across four openness levels on the 10-emotion MAFW dataset. For the O-ER group, methods such as ARPL, CSSR, and Open-VCLIP exhibit relatively low AUROC (mostly below 60%) and OSCR (below 32%), indicating that relying solely on unimodal cues limits open-set discrimination ability. HESP achieves the best results among O-ER approaches, with a mean AUROC of 69.34% and OSCR of 46.85%, benefiting from the utilization of the pretrained model and its single-modal prompt design. For the O-MER group, both CPN* and LPL*, when adapted to the multimodal setting, show clear advantages over unimodal O-ER methods, especially in OSCR (*e.g.*, CPN* reaches a mean OSCR of 56.14%). Our reproduced multimodal baseline, which enhances ARPL with multimodal pretrained models, achieves a mean AUROC of 65.32% and OSCR of 50.19%, further validating the contribution of multimodal cues. HCEP surpasses all compared methods across all openness levels, achieving the highest AUROC (77.72%) and OSCR (62.87%), with relative gains of 18.99% and 25.26% over the multimodal baseline. These improvements are consistent across openness settings, demonstrating that HCEP maintains strong recognition of known emotions while effectively detecting unseen ones, validating the effectiveness and robustness of our design.

4.2.2. Experiments on 6 emotions in CMU-MOSEI

Table 3 shows a comparative evaluation of HCEP against other approaches on the 6-emotion CMU-MOSEI dataset under four different openness levels. For the O-ER group, ARPL, CSSR, and Open-VCLIP yield mean AUROC values below 56% and OSCR values below 34%, indicating limited capability to handle unseen classes when relying solely on

Table 3
Results with 6 basic emotions in CMU-MOSEI under four openness settings.

Type	Method	AUROC					OSCR				
		O(5:1)	O(4:2)	O(3:3)	O(2:4)	Mean	O(5:1)	O(4:2)	O(3:3)	O(2:4)	Mean
O-ER	ARPL [25]	50.66	54.80	55.41	50.36	52.80	19.30	31.49	34.96	37.93	30.92
	CSSR [32]	56.08	59.46	54.60	53.35	55.87	24.56	32.98	32.58	42.12	33.06
	Open-VCLIP [33]	50.31	54.05	55.14	52.02	52.93	20.41	28.89	34.16	38.28	30.44
	HESP [3]	61.31	57.39	57.84	53.30	57.46	30.23	35.00	38.14	41.43	36.20
O-MER	CPN* [34]	53.65	52.85	56.99	55.01	54.61	25.90	34.05	38.90	41.20	35.01
	LPL* [35]	50.74	52.32	51.85	48.09	50.75	17.81	22.83	26.31	30.17	24.28
	Baseline	56.12	53.80	52.69	54.78	54.27	26.43	31.27	33.43	40.58	32.93
	Ours	**64.94**	**63.25**	**63.84**	**60.47**	**63.13**	**35.67**	**41.52**	**44.62**	**47.74**	**42.39**

Table 4
Results on 32 compound emotions (MAFW), 236 open-world dataset (OV-MERD), and cross datasets (MAFW and CMU-MOSEI).

Type	Method	32 Compound Emotions		236 Open-world Emotions		Cross Datasets	
		AUROC	OSCR	AUROC	OSCR	AUROC	OSCR
O-ER	ARPL [25]	58.14	19.38	48.87	25.01	66.35	26.76
	CSSR [32]	56.13	19.09	50.21	29.17	67.18	25.89
	OpenVCLIP [33]	54.22	26.82	49.36	26.13	54.46	22.53
	HESP [3]	63.37	36.18	54.42	28.65	72.67	37.97
O-MER	CPN* [34]	59.74	38.45	51.40	33.34	56.49	30.88
	LPL* [35]	56.12	40.46	51.25	29.19	51.65	31.82
	Baseline	59.20	37.99	50.86	27.05	63.61	38.07
	Ours	**64.17**	**44.68**	**59.95**	**37.53**	**77.54**	**51.81**

unimodal cues. HESP performs best among O-ER methods, with a mean AUROC of 57.46% and OSCR of 36.20%, benefiting from its pretrained backbone and single-modal prompt design. For the O-MER group, the baseline achieves a balanced performance (54.27% AUROC and 32.93% OSCR), validating the effectiveness of capturing multimodal emotional cues in open-set scenarios. HCEP achieves the best results across all openness levels, with mean AUROC and OSCR of 63.13% and 42.39%, respectively, representing relative gains of 9.87% and 17.10% over the suboptimal HESP. Notably, these improvements remain consistent from low openness O(5:1) to high openness O(2:4), demonstrating that HCEP effectively recognizes known emotions and rejects unseen ones, even as the proportion of unknown classes increases. This consistent performance across openness levels underscores the robustness of HCEP in open-set multimodal emotion recognition.

4.2.3. Results on 32 compound emotions in MAFW

The compound emotion recognition setting is inherently more challenging, as samples often convey subtle differences or ambiguous affective states. As shown in Table 4, HCEP achieves the best performance in this scenario, with an AUROC of 64.17% and an OSCR of 44.68%. Compared with HESP (63.37% AUROC, 36.18% OSCR), HCEP attains a modest 0.8% gain in AUROC but a substantial 8.5% improvements in OSCR, indicating that our method not only maintains comparable detection of unknowns but also establishes significantly clearer separation between known emotion categories. Competing O-MER methods such as CPN* and LPL* are notably weaker, with AUROC around 56-59% and OSCR below 41%, reflecting their limited capacity to handle fine-grained emotion categories. By explicitly incorporating the two-level prompt design and the DPOL module, HCEP effectively captures and aligns multimodal emotional cues while refining decision boundaries, thereby enhancing its capability to address compound emotions in O-MER scenarios.

4.2.4. Results on 236 open-world emotions in OV-MERD

As summarized in Table 4, the OV-MERD [28] dataset poses a greater challenge due to its broader and more fine-grained emotion categories. In this challenging benchmark, HCEP achieves 59.95% AUROC and 37.53% OSCR, outperforming all prior approaches. Compared with HESP (54.42% AUROC, 28.65% OSCR), the improvement is 5.53% in AUROC and 8.88% in OSCR, underlining HCEP's strong capability in handling large-scale and highly diverse emotion distributions. While other methods maintain AUROC around 50% and OSCR around 30%, our approach not only preserves strong recognition of known classes but also effectively detects previously unseen classes. This suggests that the proposed two-level prompt mechanism and distance-based open-set learning strategy are highly effective in discriminating fine-grained emotion states. Overall, the results on the OV-MERD dataset further validate the effectiveness of HCEP as a robust and generalizable framework for open-set multimodal emotion recognition under real-world conditions.

4.2.5. Results on cross datasets (MAFW and CMU-MOSEI)

As shown in Table 4, HCEP attains an AUROC of 77.54% and an OSCR of 51.81% on the cross datasets setting, surpassing all competing methods by a clear margin. In particular, compared to the strongest O-ER competitor HESP (72.67% AUROC, 37.97% OSCR), HCEP improves by 4.87% in AUROC and 13.84% in OSCR, highlighting its superior ability to reject unseen categories while maintaining accurate classification for known emotions. Against classical O-ER methods such as ARPL and CSSR, the improvements are even more pronounced, exceeding 10% in AUROC and 20% in OSCR. Furthermore, compared with multimodal methods such as CPN*, LPL*, and our baseline model, HCEP demonstrates substantial improvements in both metrics, especially in OSCR, demonstrating that HCEP achieves superior robustness and effectiveness when addressing O-MER tasks under distribution shift.

4.3. Ablation studies

4.3.1. Effects of different modules

Table 5 presents the performance improvements obtained by incrementally adding different components to the baseline for the 10-emotion O-MER task. The SMEP module captures unified cross-modal emotion features, boosting AUROC by 6.18% and OSCR by 12.35% relatively,

Fig. 5. Visualization of prediction results across different methods for known and unknown emotions.

Table 5
Effects of different modules in HCEP.

Baseline	SMEP	CUEP	DPOL		AUROC	OSCR
			L_{emo}	L_{osr}		
√					65.32	50.19
√	√				69.36 (+6.18%)	56.39 (+12.35%)
√	√		√		70.07 (+7.27%)	57.88 (+15.32%)
√		√		√	74.52 (+14.08%)	56.73 (+13.03%)
√	√	√		√	77.35 (+18.42%)	61.78 (+23.09%)
√	√	√	√	√	**77.72** (+18.98%)	**62.87** (+25.26%)

highlighting the value of cross-modal semantic alignment. Integrating the DPOL module further enhances open-set discrimination, improving rejection of unseen emotions while preserving known-class accuracy. The CUEP module contributes by sharpening class-level emotion boundaries, resulting in improvements of 14.08% in AUROC and 13.03% in OSCR. When all three modules are combined, HCEP achieves the best overall performance, with AUROC and OSCR increasing by 18.98% and 25.26%, demonstrating that these components complement each other to effectively address multimodal emotion heterogeneity and ambiguous emotion boundaries in O-MER tasks.

4.3.2. *Effects of different prompt settings in SMEP*

Fig. 4 compares different prompt designs in the SMEP module. *Firstly*, in Fig. 4(a), we compared our SMEP with three alternative prompt alignment strategies: (1) visual prompt-guided alignment (Visual), (2) audio prompt-guided alignment (Audio), and (3) without cross-modal alignment (w/o). As can be seen, the visual/audio-guided prompts lack explicit emotional semantic guidance, which leads to a significant performance degradation. In the w/o setting, where the prompt vectors of the three modalities are optimized independently without cross-modal guidance, the emotional semantics across modalities cannot be effectively aligned, resulting in performance drops of 0.98% in AUROC and 1.13% in OSCR. The proposed SMEP achieves the best results by leveraging explicit textual emotion semantics. *Secondly*, in Fig. 4(b), we investigated the impact of the number of learnable tokens m in textual semantic prompts. The best performance (AUROC: 77.04%, OSCR: 58.14%) is achieved with $m = 2$, balancing known and unknown emotion recognition.

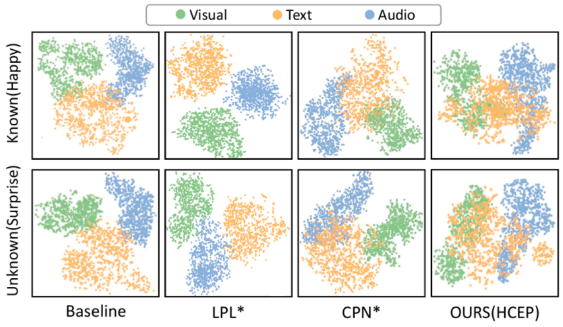

Fig. 6. Visualization on prompt-enhanced emotion-unified features in SMEP. Our method effectively address multimodal emotion heterogeneous, aligning cross-modal emotion semantic features for O-MER.

4.3.3. *Effects of different prompt designs in CUEP*

To further analyze the prompt designs behind the Class-level Unimodal Emotion-opposing Prompting (CUEP) module, we conduct ablation studies from two perspectives: (1) the construction of positive and negative features, and (2) the formulation of textual prompts. Results on the 10-class O-MER task are summarized in Table 6.

We first investigate how the construction of positive and negative features affects performance. As shown in Table 6, following ARPL [25] by using only randomly initialized negative features leads to the poorest performance (AUROC: 70.07%, OSCR: 57.88%). Adding randomly initialized positive features brings slight gains. The incorporation of our negative emotion prompts enables the pretrained models to generate more precise negative features, which results in notable gains in model performance. Ultimately, by incorporating both positive and negative emotion prompts, our HCEP yields optimal performance (AUROC: 77.72%, OSCR: 62.87%), significantly enhancing the model's capability in distinguishing between known and novel emotional categories.

We further analyze the impact of different textual prompt formulations in the CUEP module. As shown in the bottom half of Table 6, the best performance is achieved when using fixed natural language templates (*e.g.*, "This video is [Class-k]" and "This video is **not** [Class-k]") combined with explicit emotion labels. Such prompts provide clear and semantically meaningful guidance, facilitating the capture of class-level

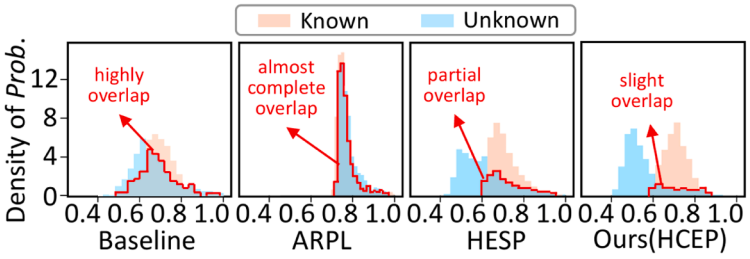

Fig. 7. Known/unknown predict probability distributions.

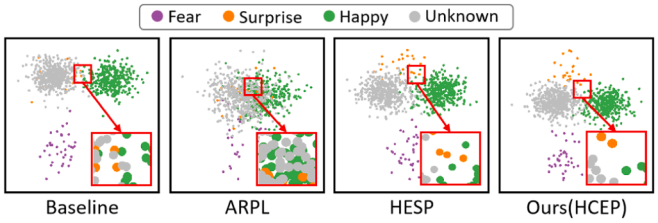

Fig. 8. Visualization of multimodal features extracted by different methods for known and unknown classes.

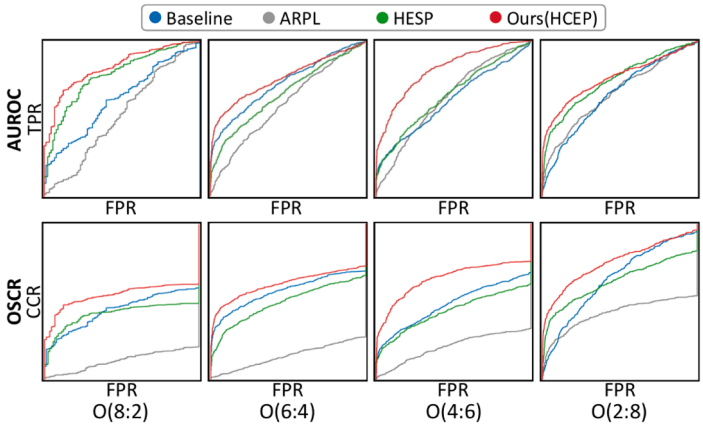

Fig. 9. Visualization of AUROC and OSCR curves under four openness settings on the MAFW dataset. Each curve is obtained by sweeping all possible thresholds..

discriminative knowledge and helping to establish sharper class boundaries in the embedding space. Text prompts based on learnable tokens combined with emotion labels achieve the second-best performance, suggesting that label supervision can partially compensate for the lack of explicit linguistic structure. In contrast, fully learnable prompts without emotion labels perform the worst, highlighting the importance of incorporating explicit emotional semantics when adapting pretrained foundation models for the O-MER task.

4.3.4. Effects of prompt-enhanced features used in emotion distances in DPOL

We adopt a hybrid feature setting in DPOL, where positive textual features F_T^+ are used in the known-class emotion recognition stream to provide stable semantic supervision, while positive and negative multimodal features F_M^+ and F_M^- are used in the unknown-class emotion discovery stream to optimize class-level emotional decision boundaries. To validate this design, we compare it with two alternatives: using only textual features or only multimodal features for all emotion distance computations. As shown in Table 7, the textual-only setting lacks multimodal emotional cues and weakens emotion boundary optimization. In contrast, the multimodal-only setting provides richer representations but suffers from instability due to the dynamic and noisy nature of learned features. Our hybrid setting effectively balances semantic consistency and discriminative adaptability, leading to the best overall performance.

4.3.5. Model complexity and efficiency analysis

To evaluate the computational complexity and efficiency of HCEP, we present comparative experiments in Table 8, covering both unimodal and multimodal methods under different open-set settings. As shown in the table, the proposed HCEP achieves significant performance improvements by introducing a two-level prompt learning mechanism to guide frozen pretrained models in distinguishing between known and unknown emotion categories. Although this design introduces a certain number of additional learnable parameters and slightly increases inference time, HCEP consistently achieves the best performance across all openness settings, which we consider a reasonable trade-off. Moreover, while both the parameter size and inference time increase as the number of known emotion categories grows, the growth remains moderate and within an acceptable range. Finally, when more advanced multimodal models, such as ImageBind and LanguageBind, are incorporated to enhance ARPL, their performance improvements on the O-MER task remain limited, even though they have already attempted to align additional modalities with vision and language during pretraining. Moreover, this comes at the cost of substantially increased computational overhead and longer inference time. In contrast, the proposed HCEP achieves superior performance while maintaining a more efficient and scalable design.

4.4. Visualization and analysis

4.4.1. Visualization of prediction results

Fig. 5 visualizes the emotion prediction results across different methods on multimodal inputs. As illustrated, closed-set recognition methods (*e.g.*, MoMKE [38]) are limited to identifying known emotion categories and tend to misclassify unknown samples into the most similar known classes. To address this limitation, a series of threshold-based open-set recognition methods, including LPL*, Baseline, and our proposed HCEP, have been developed. These methods distinguish between known and unknown classes by comparing the maximum prediction confidence score with a predefined threshold. Among them, LPL* and Baseline demonstrate limited open-set recognition capability. In contrast, the proposed method HCEP achieves more accurate classification of known classes and improved detection of unknown categories, exhibiting strong open-set recognition performance for both single-label emotions and compound emotion categories. Nevertheless, the current experiments are conducted under the assumption that all samples are complete and of high quality. In more challenging scenarios, such as multimodal inputs with injected temporal disorder or misalignment, different methods may suffer from more severe cross-modal inconsistency and temporal misalignment, which can further degrade recognition performance.

Table 6
Ablation results on positive/negative feature construction and text prompt formulations in CUEP. (**neg** denotes negative features, **pos** denotes positive features, **w/o** means without, and **w/** means with).

Prompt Setting		AUROC	OSCR
Feature Construction	Random neg only	70.07	57.88
	Random pos & neg	71.29 (+1.74%)	57.38 (−0.86%)
	Prompted neg only	76.67 (+9.42%)	61.74 (+6.67%)
	Prompted neg & pos	**77.72** (+10.92%)	**62.87** (+8.62%)
Text Prompt Formulations	Learnable tokens w/o labels	76.36	62.14
	Learnable tokens w/ labels	76.61 (+0.33%)	62.44 (+0.48%)
	Fixed templates w/ labels	**77.72** (+1.78%)	**62.87** (+1.17%)

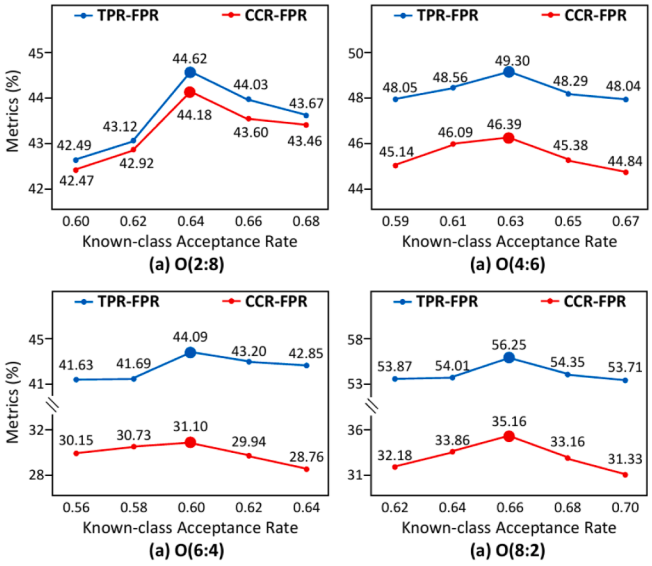

Fig. 10. Effect of threshold under different known-class acceptance rates.

Table 7
Effects of different feature settings in DPOL.

Feature Settings	AUROC	OSCR
Textual-only Features	67.52	54.31
Multimodal-only Features	76.49 (+13.28%)	61.81 (+13.81%)
Hybrid Features Setting	**77.72** (+15.11%)	**62.87** (+15.76%)

4.4.2. Visualization on SMEP-Prompted multimodal features alignment

To validate the effectiveness of the proposed SMEP module in mitigating multimodal emotion heterogeneity and aligning emotion features across modalities, Fig. 6 compares per-modality feature distributions of a known class ("happy") and an unknown class ("surprise") for the baseline, LPL* [35], CPN* [34], and our HCEP. For the baseline and LPL*, features from different modalities exhibit loose clustering and noticeable emotional discrepancies for both known and unknown classes, revealing the challenge of multimodal heterogeneity. CPN* shows slightly improved consistency but still suffers from noticeable inter-modal gaps. In contrast, HCEP achieves the most consistent and compact alignment, with visual and audio features closely anchored to their textual counterparts. This alignment reduces inter-modal distance and mitigates emotion heterogeneity, demonstrating that HCEP enables a unified multimodal emotion representation and enhances recognition performance under open-set conditions.

4.4.3. Visualization on different probability scores for known and unknown emotion categories

Fig. 7 shows the predicted probability distributions of known and unknown emotion classes for four methods: the baseline, ARPL [25], HESP [3], and our HCEP. For the baseline and ARPL, the two distributions largely overlap, indicating limited discrimination. HESP reduces this overlap by shifting unknown-class scores toward lower probabilities, though the separation remains incomplete. In contrast, HCEP produces clearly separated distributions, with known classes concentrated in high-probability regions and unknown classes in low ones. This distinct margin highlights HCEP's ability to establish precise decision boundaries and achieve superior discriminative performance in O-MER.

4.4.4. Visualization on emotion-discriminative feature distributions

Fig. 8 demonstrates emotion-discriminative features extracted by four methods–ARPL [25], the baseline, HESP [3], and our HCEP–for the 6-emotion O-MER task under openness O(3:3). Compared with other methods, HCEP shows the most distinct and compact separation between known and unknown emotions, demonstrating its ability to guide pretrained models in learning discriminative, emotion-unified multimodal features. In contrast, the baseline and ARPL exhibit substantial category overlap, indicating limited robustness in distinguishing subtle emotions. HESP improves clustering and partially restores class boundaries, yet overlaps persist between similar emotions. HCEP, however, maintains intra-class compactness, enlarges inter-class margins, and effectively isolates unknown emotions, reducing semantic ambiguity. These visualizations qualitatively confirm HCEP's superior emotion discriminability and robustness for open-set multimodal emotion recognition.

4.4.5. Visualization and analysis of AUROC and OSCR curves

To provide a more intuitive evaluation beyond scalar metrics, we visualize the AUROC (TPR-FPR) and OSCR (CCR-FPR) curves under four openness settings on MAFW (Fig. 9). These curves are obtained by sweeping all possible thresholds, reflecting performance across the full decision spectrum. From the AUROC curves, our method consistently achieves higher TPR at the same FPR, indicating stronger separability between known and unknown samples. The OSCR curves further show that our method maintains higher CCR under comparable FPR, demonstrating better joint capability of correctly classifying known samples while rejecting unknown ones. Overall, these results provide clear visual evidence that our method outperforms existing approaches across a wide range of thresholds, consistent with the reported AUROC and OSCR improvements.

4.4.6. Threshold estimation for practical inference

Although AUROC and OSCR enable fair and threshold-independent evaluation under the open-set setting, practical deployment still requires a decision threshold to distinguish known from unknown samples. To this end, we estimate the threshold θ using only known-class training data by controlling the acceptance rate of known samples. Specifically, θ is selected such that a certain proportion of known samples are correctly accepted as known, without accessing any unknown samples. We further analyze this strategy on the MAFW dataset under four openness settings. As shown in Fig. 10, increasing the acceptance rate improves TPR and CCR but also raises FPR. Both TPR - FPR and CCR - FPR exhibit a peak at

Table 8
Comparison of computational complexity and performance of different methods under varying openness levels. "Params" denotes the number of learnable parameters.

Openness	Modality	Method	AUROC	OSCR	Params(M)	Pretrained Models	Inference Time(s)
O(8:2)	Unimodal	APRL [25]	54.91	12.62	1.0	-	26.83
		CSSR [32]	60.35	18.34	9.21	-	240.91
		HESP [3]	71.47	38.04	2.21	CLIP(151.28M)	37.69
	Multimodal	LanguageBind [36] + APRL	68.12	41.36	1.97	LanguageBind(731.93M)	33.24
		ImageBind [37] + ARPL	71.36	43.26	2.37	ImageBind(1200.78MM)	60.63
		HCEP(Ours)	80.12	54.55	9.91	CLIP(151.28M) Wav2vec 2.0(94.37M)	38.80
O(6:4)	Unimodal	APRL [25]	52.18	19.44	1.0	-	25.94
		CSSR [32]	57.38	23.07	9.14	-	210.57
		HESP [3]	66.43	39.87	1.81	CLIP(151.28M)	36.28
	Multimodal	LanguageBind [36] + APRL	67.16	42.67	1.97	LanguageBind(731.93M)	34.48
		ImageBind [37] + ARPL	67.43	45.36	2.36	ImageBind(1200.78MM)	59.45
		HCEP(Ours)	75.86	56.99	8.99	CLIP(151.28M) Wav2vec 2.0(94.37M)	32.35
O(4:6)	Unimodal	APRL [25]	57.69	23.81	1.0	-	25.55
		CSSR [32]	55.51	26.37	9.08	-	153.89
		HESP [3]	68.46	49.72	1.41	CLIP(151.28M)	35.26
	Multimodal	LanguageBind [36] + APRL	67.16	52.36	1.97	LanguageBind(731.93M)	38.41
		ImageBind [37] + ARPL	68.74	53.29	2.36	ImageBind(1200.78MM)	60.64
		HCEP(Ours)	79.02	66.50	8.06	CLIP(151.28M) Wav2vec 2.0(94.37M)	36.15
O(2:8)	Unimodal	APRL [25]	63.77	45.24	1.0	-	24.30
		CSSR [32]	61.60	43.78	9.01	-	103.52
		HESP [3]	70.89	59.77	1.01	CLIP(151.28M)	34.17
	Multimodal	LanguageBind [36] + APRL	65.82	61.89	1.97	LanguageBind(731.93M)	41.52
		ImageBind [37] + ARPL	66.07	63.07	2.36	ImageBind(1200.78MM)	56.77
		HCEP(Ours)	75.89	73.43	7.13	CLIP(151.28M) Wav2vec 2.0(94.37M)	34.48

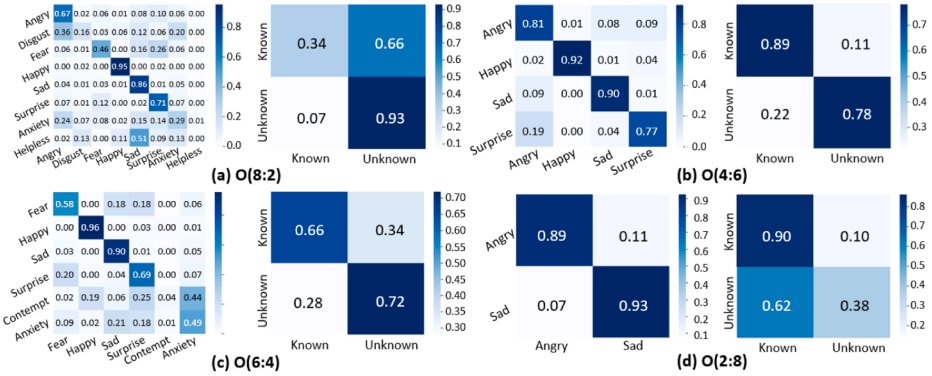

Fig. 11. Confusion matrices for the HCEP method on the O-MER task in the MAFW dataset under varying openness settings.

a moderate operating point, suggesting that overly strict or overly loose thresholds degrade performance. Empirically, a threshold in the range of 0.60-0.66 achieves near-optimal performance. These results indicate that a reasonable threshold can be approximated using only training data, while the sensitivity to θ further justifies the use of threshold-independent metrics for primary evaluation.

4.4.7. Comparison of per-class emotion recognition

To comprehensively evaluate our HCEP's recognition capabilities in both known-category classification and unknown-category detection, we employ two complementary confusion matrix visualizations under varying openness settings. Fig. 11 shows the confusion matrices of our method on the 10 emotion O-MER task. The closed-set matrix (left) reveals our method's discriminative power in different known emotion recognition, while the open-set matrix (right) demonstrates consistent unknown-class rejection performance. Overall, our method demonstrates relatively effective discrimination of different known emotion categories and detection of unknown emotion categories across various openness scenarios, especially for prominent emotion categories such as angry, sad, and happy.

5. Conclusion

In this paper, we propose a novel Hierarchical Cross-modal Emotion-interactive Prompting (HCEP) method to address Open-set Multimodal Emotion Recognition (O-MER) in real-world scenarios. The proposed framework comprises three key modules: Semantic-level Multimodal Emotion-aligning Prompting (SMEP) to capture unified multimodal emotion features; Class-level Unimodal Emotion-opposing Prompting (CUEP) to refine decision boundaries for fine-grained emotion discrimination; and Dual-stream Prompt-driven Open-set Learning (DPOL) to jointly enhance known emotion recognition and unknown emotion detection across modalities. Extensive experiments on five O-MER tasks demonstrate that HCEP significantly outperforms existing state-of-the-art methods in both known-class recognition and unknown-class detection.

Despite these promising results, our approach has several limitations that guide future work. First, HCEP's performance relies heavily on specific pretrained backbones (*e.g.*, CLIP and Wav2Vec 2.0). Future research will integrate more powerful foundation models, such as ImageBind [37], to further enhance cross-modal representation. Second, the current framework is limited to text, vision, and audio. To capture

the full complexity of human emotion, we aim to incorporate heterogeneous modalities like physiological signals (*e.g.*, EEG) and body language. Third, HCEP assumes temporally aligned, high-quality inputs, limiting its robustness against data misalignment and noise prevalent in real-world scenarios. We plan to address this by exploring meta-learning [39] and temporal alignment strategies. Finally, our current open-set detection treats all unseen samples as a single broad category. Moving forward, we will investigate structured modeling techniques to move beyond simple detection and achieve fine-grained distinction among different unknown emotion categories.

CRediT authorship contribution statement

Yuanyuan Liu: Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Funding acquisition; **Shuyang Liu:** Writing – review & editing, Writing – original draft, Visualization, Investigation; **Jiahao Zhang:** Visualization, Validation; **Ke Wang:** Writing – review & editing; **Chang Tang:** Supervision; **Dapeng Tao:** Formal analysis; **Zhe Chen:** Writing – review & editing, Validation, Supervision; **Wei Xiang:** Supervision.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by Yunling Scholar Talent Program of Yunnan Province under Grant K264202230207.

References

[1] A.V. Geetha, T. Mala, D. Priyanka, E. Uma, Multimodal emotion recognition with deep learning: advancements, challenges, and future directions, Inf. Fusion 105 (2024) 102218.
[2] Y. Zhang, Y. Yao, X. Liu, L. Qin, W. Wang, W. Deng, Open-set facial expression recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 646–654.
[3] Y. Liu, Y. Huang, S. Liu, Y. Zhan, Z. Chen, Z. Chen, Open-set video-based facial expression recognition with human expression-sensitive prompting, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 5722–5731.
[4] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: a tutorial and review, Inf. Fusion 59 (2020) 103–126.
[5] J. Zhu, B. Luo, A. Sun, J. Tan, X. Zhao, Y. Gao, Variance-aware Bi-attention expression transformer for open-set facial expression recognition in the wild, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 862–870.
[6] O. Schrüfer, M. Milling, F. Burkhardt, F. Eyben, B. Schuller, Are you sure? Analysing uncertainty quantification approaches for real-world speech emotion recognition, arXiv preprint arXiv:2407.01143 (2024).
[7] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Adv. Neural Inf. Process. Syst. 33 (2020) 12449–12460.
[8] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PmLR, 2021, pp. 8748–8763.
[9] H. Qu, X. Hui, Y. Cai, J. Liu, LMC: large model collaboration with cross-assessment for training-free open-set object recognition, Adv. Neural Inf. Process. Syst. 36 (2023) 46491–46504.
[10] G.A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
[11] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: International Conference on Machine Learning, PMLR, 2021, pp. 8821–8831.
[12] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.
[13] J. Shao, Z. Song, J. Wu, W. Shen, OpenFE: feature-extended openmax for open set facial expression recognition, Signal Image Video Process. 18 (2) (2024) 1355–1364.
[14] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (9) (2023) 1–35.
[15] A. Jha, In the era of prompt learning with vision-language models, arXiv preprint arXiv:2411.04892 (2024).
[16] Y. Wang, Y. Liu, S. Zhou, Y. Huang, C. Tang, W. Zhou, Z. Chen, Emotion-oriented cross-modal prompting and alignment for human-centric emotional video captioning, IEEE Trans. Multimed. (2025) 27 3766–3780.
[17] Z. Xie, B. Guan, W. Jiang, M. Yi, Y. Ding, H. Lu, L. Zhang, PA-SAM: prompt adapter sam for high-quality image segmentation, in: 2024 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2024, pp. 1–6.
[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
[19] Q. Wang, J. Du, K. Yan, S. Ding, Seeing in flowing: adapting clip for action recognition with motion prompts learning, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5339–5347.
[20] S. Shen, S. Yang, T. Zhang, B. Zhai, J.E. Gonzalez, K. Keutzer, T. Darrell, Multitask vision-language prompt tuning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 5656–5667.
[21] Y. Shou, T. Meng, W. Ai, F. Zhang, N. Yin, K. Li, Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations, Inf. Fusion 112 (2024) 102590.
[22] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, Y. Huang, Video sentiment analysis with bimodal information-augmented multi-head attention, Knowl.-Based Syst. 235 (2022) 107676.
[23] S. Zou, X. Huang, X. Shen, Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5994–6003.
[24] Y. Chen, H. Luo, J. Chen, D. Wang, Multimodal emotion recognition algorithm based on graph attention network, in: 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), 2024, pp. 814–822. https://doi.org/10.1109/AINIT61980.2024.10581429
[25] G. Chen, P. Peng, X. Wang, Y. Tian, Adversarial reciprocal points learning for open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2021) 8065–8081.
[26] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, S. Shan, MAFW: a large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 24–32.
[27] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. 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), 2018, pp. 2236–2246.
[28] Z. Lian, H. Sun, L. Sun, H. Chen, L. Chen, H. Gu, Z. Wen, S. Chen, Z. Siyuan, H. Yao, B. Liu, R. Liu, S. Liang, Y. Li, J. Yi, J. Tao, OV-MER: towards open-vocabulary multimodal emotion recognition, in: Forty-second International Conference on Machine Learning, 2025. https://openreview.net/forum?id=Y8lfuSoqQz.
[29] L. Neal, M. Olson, X. Fern, W.-K. Wong, F. Li, Open set learning with counterfactual images, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 613–628.
[30] M. McDermott, H. Zhang, L. Hansen, G. Angelotti, J. Gallifant, A closer look at auroc and auprc under class imbalance, Adv. Neural Inf. Process. Syst. 37 (2024) 44102–44163.
[31] A.R. Dhamija, M. Günther, T. Boult, Reducing network agnostophobia, Adv. Neural Inf. Process. Syst. 31 (2018).
[32] H. Huang, Y. Wang, Q. Hu, M.-M. Cheng, Class-specific semantic reconstruction for open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2022) 4214–4228.
[33] Z. Weng, X. Yang, A. Li, Z. Wu, Y.-G. Jiang, Open-VCLIP: transforming CLIP to an open-vocabulary video model via interpolated weight optimization, in: International Conference on Machine Learning, PMLR, 2023, pp. 36978–36989.
[34] Z. Cheng, X.-Y. Zhang, C.-L. Liu, Unified classification and rejection: a one-versus-all framework, Mach. Intell. Res. 21 (5) (2024) 870–887.
[35] Y. Fu, Z. Liu, Z. Wang, Logit prototype learning with active multimodal representation for robust open-set recognition, Sci. China Inf. Sci. 67 (6) (2024) 162204.
[36] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, et al., LanguageBind: extending video-language pretraining to n-modality by language-based semantic alignment, arXiv preprint arXiv:2310.01852 (2023).
[37] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K.V. Alwala, A. Joulin, I. Misra, ImageBind: one embedding space to bind them all, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15180–15190.
[38] W. Xu, H. Jiang, X. Liang, Leveraging knowledge of modality experts for incomplete multimodal learning, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 438–446.
[39] D. Yu, X. Zhang, Y. Chen, A. Liu, Y. Zhang, P.S. Yu, I. King, Recent advances of multimodal continual learning: a comprehensive survey, IEEE Trans. Neural Netw. Learn. Syst. (2026).