TOWARDS ROBUST OPEN-WORLD EMOTION RECOGNITION VIA EQUIANGULAR PROTOTYPES AND PROGRESSIVE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-world emotion recognition aims to perform emotion classification using both labeled and unlabeled data, where the labeled data cover only a subset of emotion classes, while the unlabeled data encompass all classes. This setting is crucial for recognizing fine-grained unseen emotions that are difficult to collect under controlled laboratory conditions. However, existing open-world semi-supervised learning methods suffer from two major limitations: (1) they rely on similarity-based clustering for unconstrained aggregation, making it difficult to distinguish representations of different unseen classes in the absence of supervision for unseen classes; and (2) they primarily rely on the predictive confidence of a single model for class decisions, making them susceptible to distributional bias dominated by known classes and consequently prone to biased predictions. To address these issues, we propose a novel open-world emotion recognition framework, termed *ROBIN*. First, inspired by neural collapse, we introduce the simplex equiangular tight frame structure as equiangular prototypes to guide representations of different classes toward a well-separated distributional structure. Subsequently, we develop a progressive prediction strategy that first employs an auxiliary binary classifier to distinguish between known and unseen classes, and then masks class components inconsistent with the predicted base class, thereby enabling robust fine-grained class recognition. Extensive experimental results demonstrate that *ROBIN* outperforms existing state-of-the-art methods on the open-world emotion recognition.

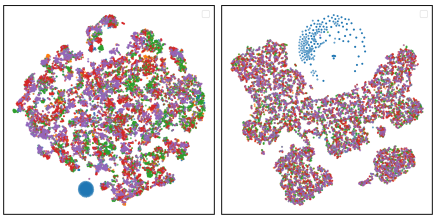
1 INTRODUCTION

Emotion recognition aims to identify an individual's affective state by leveraging multiple modalities, including electroencephalogram (EEG) signals (Li et al., 2024; Liu et al., 2024a), text (Shen et al., 2021; Tu et al., 2022), speech (Tan et al., 2022; Wagner et al., 2023), images (Zhu et al., 2023), and videos (Zhang et al., 2022b). Existing high-performance emotion recognition models typically rely heavily on large-scale datasets with precise annotations. However, recent studies suggest that human emotions can be subdivided into a variety of fine-grained classes (Demszky et al., 2020), which are often difficult to comprehensively collect and annotate under laboratory conditions.

To address this, researchers have introduced the open-world emotion recognition task to recognize unseen classes (Zhang et al., 2024; Liu et al., 2024b; 2026; Hu et al., 2026). However, existing methods generally treat different unseen classes as a single class, making them inadequate for low-supervision and fine-grained emotion recognition in open-world scenarios. Motivated by these limitations, this paper investigates a key question in open-world emotion recognition: *How can a model be trained to achieve fine-grained recognition of both known and unseen classes with limited labeled samples and incomplete class coverage?*

Early efforts introduced unseen class discovery methods (Fini et al., 2021; Han et al., 2019; 2020), which are based on the assumption that the unlabeled data contain only unseen classes. Subsequently, researchers have proposed more practical methods for open-world semi-supervised learning (Cao et al., 2022; Sun & Li, 2023; Vaze et al., 2022), which aim to identify known classes from labeled data while discovering potential unseen classes from unlabeled data.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) CGMatch (b) MOOD

Figure 1: Visualization of representation collapse produced by CGMatch (Cheng et al., 2025) and MOOD (Lu et al., 2025) on the SEED-V dataset. Blue representations correspond to multiple distinct unseen classes.

However, existing methods generally suffer from two critical limitations. **I. Representation Collapse**. Most methods rely on similarity-based (Rizve et al., 2022a; Tang et al., 2025) or contrastive learning-driven (Liu et al., 2025) clustering strategies that naively assign samples to different clusters. In addition, some approaches incorporate self-training mechanisms to generate pseudo-labels for unlabeled samples (Guo et al., 2022; Liu et al., 2023). However, due to the lack of supervision for unseen classes, these methods tend to erroneously assign samples to known classes or collapse multiple unseen classes into a single cluster. This leads to representation collapse, where different unseen classes converge to a single cluster structure in the embedding space (see Figure 1). **II. Biased Prediction**. Existing methods typically obtain reliable pseudo-labels by increasing model confidence (Cheng et al., 2025; Huang et al., 2026). However, in the absence of supervision signals for unseen classes, such single-model decision mechanisms are insufficiently robust to class-distribution bias, consequently causing the resulting predictions to favor known classes.

Recent studies have revealed a representation alignment phenomenon known as neural collapse (Papyan et al., 2020). Specifically, when a model reaches an ideal training convergence state, the representations within the same class tend to collapse toward their corresponding class mean, while the class means of different classes exhibit equal pairwise cosine similarity, thereby forming a simplex Equiangular Tight Frame (ETF) structure. Meanwhile, the weight vectors of the final classification layer tend to align directionally with their corresponding class mean vectors. Neural collapse characterizes an ideal representation geometry for classification tasks, and the high degree of symmetry exhibited by neural collapse has further motivated its application to various specialized study scenarios, including imbalanced learning (Yan et al., 2024; Shen et al., 2025), multi-task learning (Wang et al., 2026), and transfer learning (Munn et al., 2024). To alleviate representation collapse, the representation geometry revealed by neural collapse provides an important insight: *Can a fixed ETF structure be introduced to explicitly constrain the representations of different classes, thereby learning representations with strong inter-class separability?*

Against this backdrop, we propose *ROBIN*, a robust framework for open-world emotion recognition that integrates equiangular prototypes with progressive prediction. To mitigate **Limitation I**, we introduce equiangular prototypes constructed based on the ETF as classification constraints, guiding the representations of different classes to progressively approach an optimal ETF-like distribution, thereby alleviating representation collapse while enhancing discriminability. To address **Limitation II**, we further develop a progressive prediction strategy. Specifically, an auxiliary binary classifier is first employed to partition samples into known and unseen classes. Logits inconsistent with the predicted base class are then masked accordingly. Finally, fine-grained class prediction is performed over the remaining candidates, thereby improving prediction robustness. The main contributions are summarized as follows:

- **Novel Exploration:** To address the limitations of existing open-world emotion recognition methods, we propose a fine-grained emotion recognition task in a setting with limited labeled samples and incomplete class coverage. This setting more closely reflects real-world scenarios and represents a novel exploration in emotion recognition.
- **Novel Method:** We propose a novel framework for open-world semi-supervised emotion recognition, termed *ROBIN*. The proposed framework introduces equiangular prototypes to optimize the structure of representation learning, and employs a progressive prediction strategy for classification, thereby enabling robust recognition of emotion classes in open-world scenarios.
- **Extensive Empirical Validation:** We conducted comprehensive experiments on five emotion datasets. The results demonstrate the superiority of *ROBIN*. Furthermore, we evaluate the contribution of each component to the overall performance and analyze the sensitivity of the framework to key hyperparameters.

2 PRELIMINARY

2.1 PROBLEM DEFINITION

Given an emotion dataset $\mathcal{D} = \{\mathcal{D}_l, \mathcal{D}_u\}$, where $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{L}$ denotes the labeled data and $\mathcal{D}_u = \{x_i\}_{i=1}^{U}$ denotes the unlabeled data. Here, x represents a sample, and y denotes the corresponding label. Under the open-world emotion recognition setting, the labeled set $\mathcal{D}_l$ only covers a subset of emotion classes, denoted as $\mathcal{C}_{known}$, while the unlabeled set $\mathcal{D}_u$ covers the entire class space $\mathcal{C}_{all}$, where $\mathcal{C}_{known} \subset \mathcal{C}_{all}$. The unseen emotion classes are defined as $\mathcal{C}_{unseen} = \mathcal{C}_{all} \setminus \mathcal{C}_{known}$. The objective is to train a model using $\mathcal{D}$ such that it can perform fine-grained classification of both known and unseen classes. Following the settings adopted by existing open-world semi-supervised learning studies (Fini et al., 2021; Cao et al., 2022; Sun & Li, 2023), we assume that the numbers of known and total classes are both known and fixed.

2.2 THEORETICAL ANALYSIS

Proposition 1. *Let the classifier weight be $\theta \in \mathbb{R}^{d\times K}$, where d and K denote the representation dimension and the number of classes, respectively, with $d \geq K$. We fix θ to the following structure:*

$$\theta = \sqrt{\frac{K}{K-1}} O\big(I_K - \frac{J_K}{K}\big) \tag{1}$$

where I_K denotes the K-dimensional identity matrix, J_K denotes the $K \times K$ all-ones matrix, and $O \in \mathbb{R}^{d\times K}$ is a matrix with orthonormal columns. Let $\theta = [e_1, \ldots, e_K]$, where $e_i \in \mathbb{R}^d$ denotes the class prototype corresponding to the i-th class. Then, for any class prototype,

$$e_i^\top e_j = \begin{cases} 1, & i = j \\ -\frac{1}{K-1}, & i \neq j \end{cases}, \quad \forall i, j \in [1, K] \tag{2}$$

When the learning objective encourages representations from class i to approach their class prototype e_i, representations across different classes are encouraged to exhibit uniformly distributed and maximally separated angular directions. A detailed proof is provided in Appendix A.

Proposition 1 indicates that the classifier weights can be fixed as an ETF, which serves as equiangular prototypes for representations across classes. These constraints guide class-wise representations to progressively approach their class prototypes during training, thereby enhancing the separability of representations across classes.

3 METHODOLOGY

ROBIN comprises two modules, representation alignment and progressive prediction, with its overall framework illustrated in Figure 2.

3.1 REPRESENTATION ALIGNMENT

3.1.1 LOCALIZATION VIA EQUIANGULAR PROTOTYPES

To maximize the separability of representations across different classes in the embedding space, we set $\theta_{etf} \in \mathbb{R}^{d\times K}$ as non-learnable parameters and fix its weights to the equiangular structure defined in Proposition 1. To encourage the representations of different classes to approach the predefined optimal distribution, we employ the cross-entropy loss to measure the discrepancy between the logits produced by $\phi_{\theta_{etf}}$ and the corresponding ground-truth labels. Let $H_{etf(l)} \in \mathbb{R}^{|X_l|\times K}$ denote the logits output by $\phi_{\theta_{etf}}$ for the labeled samples X_l. The localization loss $\mathcal{L}_{loc}$ is defined as follows:

$$\mathcal{L}_{loc} = -\frac{1}{|X_l|}\sum_{i=1}^{|X_l|}\sum_{k=1}^{K} y_i^k \log \frac{\exp\left(h_i^k\right)}{\sum_{j=1}^{K}\exp\left(h_i^j\right)}, \quad h_i \in H_{etf(l)} = \phi_{\theta_{etf}}\left(\phi_{\theta_g}\left(X_l;\theta_g\right);\theta_{etf}\right) \tag{3}$$

where $y_i \in \{0,1\}^K$ denotes the one-hot ground-truth label of sample x_i. $\mathcal{L}_{loc}$ encourages inter-class representation separation by aligning each representation with its corresponding class prototype.

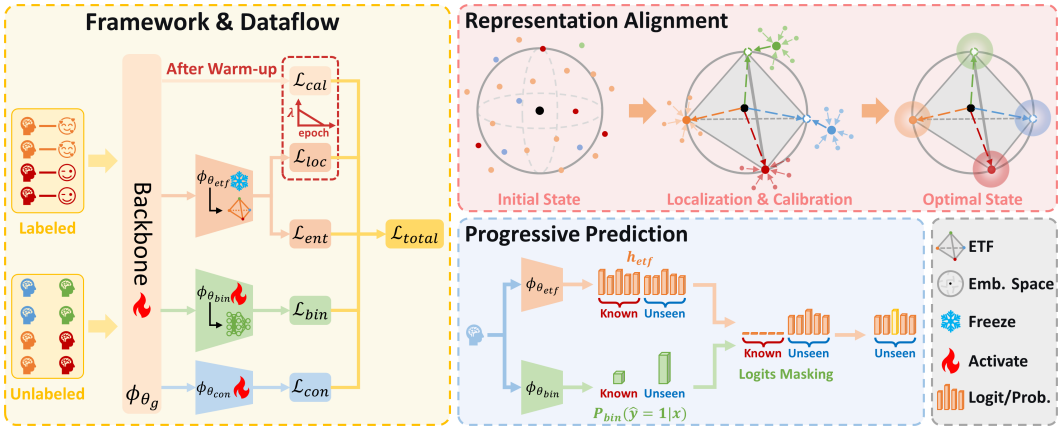


Figure 2: Overall framework of *ROBIN*. *ROBIN* consists of a backbone ϕ_{θ_g} and three functional modules for representation alignment, contrastive learning, and base-class classification. Specifically, the equiangular classifier $\phi_{\theta_{etf}}$ incorporates an ETF as equiangular prototypes to constrain the distribution of representations in the embedding space, thereby mitigating representation collapse. The binary classifier $\phi_{\theta_{bin}}$ as an auxiliary classifier first partitions samples into base classes. Based on the predicted base class, the logit of classes inconsistent with the prediction are masked, thereby narrowing the candidate class space for subsequent fine-grained classification.

3.1.2 Calibration via Mean Prototypes

Given the lack of supervision for unseen classes, relying solely on a small amount of labeled data for supervised learning may cause representation distribution shifts, making it difficult to satisfy the ETF property. To alleviate this issue, we introduce a mean-prototype loss during training to achieve representation calibration. Let $H_{etf(u)} \in \mathbb{R}^{|X_u| \times K}$ denote the logits output by $\phi_{\theta_{etf}}$ for the unlabeled samples X_u. Specifically, we first generate one-hot pseudo-labels $Q \in \{0,1\}^{|X_u| \times K}$ for X_u using the progressive prediction strategy (see Section 3.2), and then partition $H_{etf(u)}$ according to the pseudo-labels. Subsequently, we compute the mean representation of each class, denoted as $\{\mu_k \in \mathbb{R}^d\}_{k=1}^K$, and use these class-wise means as dynamic prototypes $\mathcal{U} \in \mathbb{R}^{K \times d}$:

$$H_{etf(u)} = \phi_{\theta_{etf}}\left(\phi_{\theta_g}\left(X_u;\theta_g\right);\theta_{etf}\right), \quad \mu_k = \text{Mean}\left(H^k_{etf(u)}\right), \quad \mathcal{U} = \{\mu_k\}_{k=1}^K$$

where $H^k_{etf(u)}$ denotes the subset of logits assigned pseudo-label k, and $\text{Mean}(\cdot)$ denotes the mean operation. Similarly, we measure the calibration loss $\mathcal{L}_{cal}$ between predictions and pseudo-labels via cross-entropy:

$$\mathcal{L}_{cal} = -\frac{1}{|X_u|}\sum_{i=1}^{|X_u|}\sum_{k=1}^{K} q_i^k \log \frac{\exp\left(z_i^\top \mu_k/\tau\right)}{\sum_{j=1}^{K}\exp\left(z_i^\top \mu_j/\tau\right)}, \quad q_i \in Q, z_i \in Z_u = \phi_{\theta_g}\left(X_u;\theta_g\right) \tag{4}$$

where τ denotes the temperature coefficient, and $Z_u \in \mathbb{R}^{|X_u| \times d}$ denotes the representations output by ϕ_{θ_g} for X_u. To ensure reliable representation calibration, we maintain $\mathcal{U}$ as a global variable. The mean prototype of each class is updated and normalized after each training iteration, i.e., $\mathcal{U}_{t+1} = \text{Norm}(\mathcal{U}_t)$, where t denotes the iteration index.

3.2 Progressive Prediction

3.2.1 Base-Class Prediction

We first introduce a binary classifier $\phi_{\theta_{bin}}$ as an auxiliary classifier to distinguish between known and unseen classes. To this end, we impose two optimization objectives on $\phi_{\theta_{bin}}$. First, we encourage $\phi_{\theta_{bin}}$ to assign a high positive-class probability to X_l while reducing that probability for X_u. Let $P_{bin}\left(\hat{y} = 1 \mid X\right)$ denote the probability of predicting the input sample X as the positive class, where

$\hat{y} \in \mathbb{R}^2$ represents the predicted base-class label. We define the corresponding objective as:

$$P_{bin}\left(\hat{y}=1 \mid X\right)=\mathrm{SoftMax}\left(\phi_{\theta_{bin}}\left(\phi_{\theta_g}\left(X;\theta_g\right);\theta_{bin}\right)\right) \tag{5}$$

$$\mathcal{L}_{pos}=\mathbb{E}\left[\log P_{bin}\left(\hat{y}=1 \mid X_u\right)\right]-\mathbb{E}\left[\log P_{bin}\left(\hat{y}=1 \mid X_l\right)\right] \tag{6}$$

However, since labeled data only cover known classes while unlabeled data contain both known and unseen classes, it is difficult to achieve effective separation between known and unseen classes using $\mathcal{L}_{pos}$ alone. To address this issue, we further introduce a consistency constraint based on Mixup (Zhang et al., 2017). Specifically, for the labeled and unlabeled samples in the current batch, we perform linear interpolation using a given coefficient to generate mixed samples $\widetilde{X}$. Meanwhile, we construct the corresponding mixed labels $\widetilde{Y}$ by combining an all-one label with $P_{bin}\left(\hat{y}=1 \mid X_u\right)$ (see Appendix D.4 for details). Based on this, the consistency loss is defined as:

$$\mathcal{L}_{mix}=\mathbb{E}\left[\left(\log \widetilde{Y}-\log P_{bin}(\hat{y}=1 \mid \widetilde{X})\right)^2\right] \tag{7}$$

This constraint helps prevent $\phi_{\theta_{bin}}$ from becoming overly biased toward known classes during training, thereby improving its discriminative ability for unseen classes. Therefore, the base-class loss $\mathcal{L}_{bin}$ is formulated as:

$$\mathcal{L}_{bin}=\mathcal{L}_{pos}+\lambda_{mix}\mathcal{L}_{mix} \tag{8}$$

where λ_{mix} denotes the weight of the consistency constraint. $\mathcal{L}_{bin}$ jointly enforces positive-class prediction and consistency regularization, enabling $\phi_{\theta_{bin}}$ to learn a more robust decision boundary during training and thereby effectively distinguish between known and unseen classes.

3.2.2 Fine-grained Classification under Base-class Constraints

After the warm-up stage, $\phi_{\theta_{bin}}$ can distinguish between known and unseen classes with relatively high accuracy. We then perform fine-grained class prediction under the base-class constraint. Specifically, we use the positive-class probability $P_{bin}\left(\hat{y}=1 \mid X_u\right)$ output by $\phi_{\theta_{bin}}$ for X_u to mask the logits in $H_{etf(u)}$ that are associated with classes predicted with low confidence. For any sample $x_i \in X_u$, we have:

$$h_i^k=\begin{cases}0, & \forall k \in \mathcal{C}_{known}, \frac{P_{bin}(\hat{y}=1|x_i)}{\epsilon}<0.5\\ 0, & \forall k \in \mathcal{C}_{unseen}, \frac{P_{bin}(\hat{y}=1|x_i)}{\epsilon}>0.5\end{cases} \tag{9}$$

where $h_i \in H_{etf(u)}$ denotes the logits of x_i, h_i^k denotes the logit corresponding to the k-th class, and $\epsilon=\max\left(P_{bin}\left(\hat{y}=1 \mid X_u\right)\right)$ denotes the maximum positive-class prediction probability output by $\phi_{\theta_{bin}}$ for unlabeled samples in the previous iteration. Therefore, the base-class prediction threshold can be adaptively adjusted as training progresses. Finally, among the remaining non-zero probability components, we select the class with the highest probability as the final prediction.

3.3 Training

Following prior work (Cao et al., 2022; Liu et al., 2023), we introduce a warm-up strategy before formal training to enhance training stability and convergence.

3.3.1 Warm-up

The model is updated solely based on the localization loss $\mathcal{L}_{loc}$ and the binary classification loss $\mathcal{L}_{bin}$ during warm-up stage. To prevent the model from converging prematurely to a suboptimal solution during the early stages of training, following previous studies (Grandvalet & Bengio, 2004; Sun et al., 2018), we further introduce an entropy regularization loss $\mathcal{L}_{ent}$ to encourage smoother probability distributions P_{etf}:

$$\mathcal{L}_{ent}=\sum_{k=1}^{K}\bar{p}^k\log\bar{p}^k, \quad \bar{p}=\frac{1}{N}\sum_{i=1}^{N}p_i, \quad p_i \in P_{etf} \tag{10}$$

where $\bar{p} \in \mathbb{R}^K$ denotes the averaged prediction probability over all classes within the current mini-batch. In addition, we further incorporate a contrastive learning loss widely adopted in unsupervised

learning (Chen et al., 2020) to encourage representation aggregation. Let $Z_{con} \in \mathbb{R}^{|X| \times d}$ denote the normalized representations output by the contrastive learning module $\phi_{\theta_{con}}$ for the samples X. The contrastive loss $\mathcal{L}_{con}$ is defined as:

$$Z_{con(w/s)} = \mathrm{Norm}\Big(\phi_{\theta_{con}} \left(\phi_{\theta_g} \left(X_{(w/s)}; \theta_g \right); \theta_{con} \right) \Big) \tag{11}$$

$$\mathcal{L}_{con} = -\log \frac{\exp\left(z_{i,w}^{\top} z_{i,s} / \tau \right)}{\sum_{j=1}^{|X|} \exp\left(z_{j,w}^{\top} z_{j,s} / \tau \right)}, \quad z_{i(w/s)}, z_{j(w/s)} \in Z_{con(w/s)} \tag{12}$$

where $z_{i,w}$ and $z_{i,s}$ represent the representations of the weakly augmented sample $x_{i,w}$ and strongly augmented sample $x_{i,s}$, respectively. Therefore, the overall loss function in the warm-up stage is defined as:

$$\mathcal{L}_{total} = \lambda_{ce}\mathcal{L}_{loc} + \lambda_{ent}\mathcal{L}_{ent} + \mathcal{L}_{bin} + \mathcal{L}_{con} \tag{13}$$

where λ_{ce} and λ_{ent} denote the corresponding weighting coefficients, respectively. After the warm-up stage, $\phi_{\theta_{etf}}$ can align the representations with their predefined positions in the embedding space, while $\phi_{\theta_{bin}}$ acquires discriminative capability and can distinguish between known and unseen classes.

3.3.2 Formal Training

In the formal training stage, to continuously maintain stable localization of labeled data in the embedding space, we further incorporate the calibration loss $\mathcal{L}_{cal}$ to preserve the compactness of the clustering structure. We combine $\mathcal{L}_{loc}$ and $\mathcal{L}_{cal}$ into an alignment loss $\mathcal{L}_{al}$, and balance their contributions through a progressively adjusted weighting strategy:

$$\mathcal{L}_{al} = (1 - \lambda)\mathcal{L}_{loc} + \lambda\mathcal{L}_{cal} \tag{14}$$

where λ is a scheduling coefficient that increases linearly from 0 to 0.5 over training iterations, enabling a dynamic trade-off between the two objectives. Therefore, the total loss is defined as:

$$\mathcal{L}_{total} = \lambda_{ce}\mathcal{L}_{al} + \lambda_{ent}\mathcal{L}_{ent} + \mathcal{L}_{bin} + \mathcal{L}_{con} \tag{15}$$

Through sequential optimization in the warm-up and main training stages, the model can achieve more robust representation learning. The complete training pseudocode is provided in Appendix B.

4 Experiments

This section introduces the datasets, baselines, evaluation metrics, and experimental settings used to evaluate the performance of ***ROBIN***.

4.1 Experimental Settings

4.1.1 Datasets

We conduct experiments on five emotion datasets. Specifically, SEED-V (Liu et al., 2022) contains 5 emotion classes, SEED-VII (Jiang et al., 2025) contains 7 emotion classes, and GoEmotions (Demszky et al., 2020) contains 28 emotion classes. To eliminate class ambiguity, we remove samples associated with multiple emotion classes, resulting in a subset covering 18 emotion classes. SE-R1510 (Chen, 2023) and SE-F1530 (Chen, 2023) each contain 3 emotion classes. Detailed descriptions of the datasets are provided in Appendix F.

4.1.2 Baselines

We select a diverse set of state-of-the-art semi-supervised learning and open-world semi-supervised learning methods for comparison, including FixMatch (Sohn et al., 2020), GCD (Vaze et al., 2022), TRSSL (Rizve et al., 2022b), OpenCon (Sun & Li, 2023), OpenNCD (Liu et al., 2023), CGMatch (Cheng et al., 2025), and MOOD (Lu et al., 2025). Detailed descriptions of these baselines are provided in Appendix G.

Table 2: Performance comparison between *ROBIN* and baselines on all classes (%).

Method	SEED-V		SEED-VII		GoEmotions		SE-R1510		SE-F1530	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
FixMatch	28.10$_{\pm 2.14}$	21.91$_{\pm 3.24}$	28.31$_{\pm 1.44}$	27.60$_{\pm 1.15}$	23.78$_{\pm 0.20}$	9.41$_{\pm 0.79}$	47.59$_{\pm 4.57}$	25.34$_{\pm 4.81}$	25.96$_{\pm 1.12}$	19.22$_{\pm 1.15}$
GCD	30.52$_{\pm 1.98}$	27.83$_{\pm 2.75}$	31.47$_{\pm 1.30}$	31.01$_{\pm 1.09}$	25.69$_{\pm 0.42}$	9.95$_{\pm 0.28}$	62.05$_{\pm 5.06}$	37.36$_{\pm 5.15}$	26.84$_{\pm 1.84}$	25.23$_{\pm 1.27}$
TRSSL	36.44$_{\pm 1.31}$	32.72$_{\pm 3.14}$	34.24$_{\pm 1.41}$	36.64$_{\pm 1.46}$	25.01$_{\pm 0.84}$	11.35$_{\pm 0.51}$	67.37$_{\pm 5.13}$	29.63$_{\pm 4.86}$	34.51$_{\pm 1.35}$	33.70$_{\pm 1.51}$
OpenCon	39.05$_{\pm 1.13}$	36.78$_{\pm 3.61}$	36.93$_{\pm 1.97}$	40.78$_{\pm 1.83}$	28.16$_{\pm 0.77}$	10.06$_{\pm 0.99}$	70.18$_{\pm 5.20}$	40.75$_{\pm 4.75}$	30.09$_{\pm 1.19}$	21.68$_{\pm 1.36}$
OpenNCD	40.36$_{\pm 1.35}$	36.80$_{\pm 3.52}$	36.31$_{\pm 1.88}$	41.73$_{\pm 1.98}$	30.19$_{\pm 0.23}$	8.69$_{\pm 0.55}$	68.94$_{\pm 4.98}$	45.68$_{\pm 4.07}$	39.25$_{\pm 1.23}$	25.75$_{\pm 1.32}$
CGMatch	32.08$_{\pm 1.92}$	30.44$_{\pm 2.39}$	30.82$_{\pm 1.20}$	30.55$_{\pm 1.29}$	22.20$_{\pm 0.52}$	9.89$_{\pm 0.17}$	57.02$_{\pm 4.59}$	31.69$_{\pm 4.88}$	33.33$_{\pm 1.66}$	28.96$_{\pm 1.48}$
MOOD	31.41$_{\pm 1.88}$	26.69$_{\pm 3.21}$	31.37$_{\pm 1.42}$	31.99$_{\pm 1.54}$	24.64$_{\pm 0.37}$	9.34$_{\pm 0.24}$	46.85$_{\pm 5.09}$	23.98$_{\pm 4.86}$	26.55$_{\pm 1.94}$	19.49$_{\pm 1.51}$
ROBIN	**53.03**$_{\pm 1.87}$	**49.50**$_{\pm 2.01}$	**50.44**$_{\pm 1.37}$	**52.73**$_{\pm 0.99}$	**31.05**$_{\pm 0.34}$	**11.40**$_{\pm 0.19}$	**72.46**$_{\pm 5.02}$	**50.04**$_{\pm 4.39}$	**43.95**$_{\pm 1.77}$	**40.15**$_{\pm 1.76}$

Table 3: Performance comparison between *ROBIN* and baselines on known classes (%).

Method	SEED-V		SEED-VII		GoEmotions		SE-R1510		SE-F1530	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
FixMatch	83.56$_{\pm 4.40}$	82.35$_{\pm 4.69}$	53.48$_{\pm 5.16}$	50.62$_{\pm 5.22}$	16.79$_{\pm 0.14}$	13.18$_{\pm 0.67}$	21.21$_{\pm 4.23}$	17.50$_{\pm 4.78}$	21.79$_{\pm 1.13}$	17.89$_{\pm 1.15}$
GCD	86.58$_{\pm 4.62}$	86.29$_{\pm 4.71}$	67.70$_{\pm 5.07}$	67.20$_{\pm 5.03}$	18.46$_{\pm 0.94}$	14.59$_{\pm 0.24}$	56.52$_{\pm 4.35}$	36.11$_{\pm 4.77}$	20.94$_{\pm 1.74}$	17.31$_{\pm 1.71}$
TRSSL	76.15$_{\pm 5.70}$	76.07$_{\pm 5.76}$	69.45$_{\pm 5.11}$	69.58$_{\pm 5.23}$	18.55$_{\pm 0.69}$	12.57$_{\pm 0.51}$	33.33$_{\pm 4.61}$	25.80$_{\pm 4.58}$	28.63$_{\pm 1.64}$	22.26$_{\pm 1.14}$
OpenCon	70.83$_{\pm 4.22}$	70.75$_{\pm 4.41}$	71.14$_{\pm 5.29}$	70.31$_{\pm 5.14}$	16.68$_{\pm 0.32}$	12.68$_{\pm 0.98}$	45.45$_{\pm 4.69}$	31.25$_{\pm 4.71}$	30.77$_{\pm 1.40}$	23.53$_{\pm 1.32}$
OpenNCD	76.24$_{\pm 5.01}$	74.27$_{\pm 5.25}$	75.28$_{\pm 5.50}$	74.47$_{\pm 5.18}$	18.09$_{\pm 0.40}$	15.09$_{\pm 0.73}$	61.67$_{\pm 4.77}$	38.14$_{\pm 4.29}$	39.87$_{\pm 1.31}$	28.50$_{\pm 1.15}$
CGMatch	89.19$_{\pm 4.85}$	88.81$_{\pm 4.98}$	69.69$_{\pm 5.21}$	69.38$_{\pm 5.33}$	13.53$_{\pm 0.59}$	10.00$_{\pm 0.79}$	48.48$_{\pm 4.26}$	32.65$_{\pm 4.25}$	30.77$_{\pm 1.99}$	23.53$_{\pm 1.84}$
MOOD	79.62$_{\pm 4.18}$	78.94$_{\pm 5.99}$	65.73$_{\pm 5.48}$	61.88$_{\pm 5.37}$	13.56$_{\pm 0.85}$	11.51$_{\pm 0.70}$	33.33$_{\pm 4.06}$	25.49$_{\pm 3.99}$	24.79$_{\pm 1.18}$	19.86$_{\pm 1.21}$
ROBIN	**89.67**$_{\pm 5.10}$	**89.55**$_{\pm 5.59}$	**77.74**$_{\pm 5.40}$	**77.07**$_{\pm 5.18}$	**20.18**$_{\pm 0.46}$	**15.16**$_{\pm 0.70}$	**63.33**$_{\pm 4.66}$	**38.78**$_{\pm 4.04}$	**41.45**$_{\pm 1.32}$	**29.31**$_{\pm 1.85}$

Table 4: Performance comparison between *ROBIN* and baselines on unseen classes (%).

Method	SEED-V		SEED-VII		GoEmotions		SE-R1510		SE-F1530	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
FixMatch	11.65$_{\pm 2.94}$	9.31$_{\pm 3.02}$	20.64$_{\pm 1.42}$	19.52$_{\pm 1.56}$	26.53$_{\pm 0.63}$	6.47$_{\pm 0.95}$	49.65$_{\pm 3.93}$	33.18$_{\pm 2.98}$	35.24$_{\pm 1.63}$	26.06$_{\pm 1.70}$
GCD	13.89$_{\pm 2.72}$	12.80$_{\pm 3.11}$	20.44$_{\pm 1.60}$	18.51$_{\pm 1.31}$	28.54$_{\pm 0.57}$	6.69$_{\pm 0.25}$	62.65$_{\pm 4.26}$	38.52$_{\pm 2.87}$	40.00$_{\pm 1.32}$	37.25$_{\pm 1.46}$
TRSSL	14.05$_{\pm 2.85}$	13.90$_{\pm 3.75}$	23.52$_{\pm 1.47}$	22.77$_{\pm 1.42}$	29.11$_{\pm 0.55}$	9.19$_{\pm 0.86}$	67.85$_{\pm 4.58}$	40.42$_{\pm 2.61}$	47.62$_{\pm 1.87}$	46.03$_{\pm 1.94}$
OpenCon	21.13$_{\pm 2.76}$	20.46$_{\pm 3.86}$	24.81$_{\pm 1.46}$	22.50$_{\pm 1.33}$	32.68$_{\pm 0.52}$	8.07$_{\pm 0.30}$	72.10$_{\pm 4.80}$	41.90$_{\pm 2.97}$	28.57$_{\pm 1.04}$	22.22$_{\pm 1.29}$
OpenNCD	20.14$_{\pm 2.60}$	20.12$_{\pm 3.63}$	24.45$_{\pm 1.44}$	24.04$_{\pm 1.51}$	33.30$_{\pm 0.43}$	5.34$_{\pm 0.56}$	69.98$_{\pm 4.27}$	41.17$_{\pm 2.56}$	36.41$_{\pm 1.89}$	26.69$_{\pm 1.78}$
CGMatch	15.13$_{\pm 3.88}$	14.71$_{\pm 3.19}$	18.99$_{\pm 1.57}$	18.10$_{\pm 1.33}$	27.71$_{\pm 0.64}$	9.54$_{\pm 0.48}$	57.68$_{\pm 4.72}$	36.58$_{\pm 2.82}$	39.05$_{\pm 1.65}$	35.98$_{\pm 1.57}$
MOOD	17.11$_{\pm 3.83}$	12.66$_{\pm 3.57}$	20.91$_{\pm 1.59}$	18.96$_{\pm 1.34}$	27.49$_{\pm 0.57}$	7.04$_{\pm 0.43}$	47.04$_{\pm 3.80}$	31.99$_{\pm 2.57}$	30.48$_{\pm 1.92}$	23.36$_{\pm 1.95}$
ROBIN	**32.38**$_{\pm 3.80}$	**30.32**$_{\pm 3.14}$	**34.65**$_{\pm 1.47}$	**33.42**$_{\pm 1.41}$	**35.32**$_{\pm 0.44}$	**9.98**$_{\pm 0.18}$	**73.76**$_{\pm 4.57}$	**42.45**$_{\pm 2.24}$	**49.52**$_{\pm 1.52}$	**48.86**$_{\pm 1.43}$

4.1.3 Evaluation Metrics

We comprehensively evaluate the performance of all methods using four widely adopted metrics: Accuracy, Precision, Recall, and F1-score (macro). Under the open-world setting, models can only make predictions for unlabeled samples during evaluation, and the predictions cannot be directly compared with their ground-truth labels. When evaluating performance on unseen classes, we first employ the Hungarian algorithm (Kuhn, 1955) to optimally match the predicted labels with the ground-truth labels, thereby ensuring a fair comparison across different methods.

4.1.4 Implementation Details

We first split the dataset into training, validation, and test sets with a ratio of 8:1:1, and then further divide the training set into labeled and unlabeled subsets. For the training set, the proportion of known to the total number of classes is no greater than 50%. By default, the labels of the known classes range from $[0, |\mathcal{C}_{known}|)$, while those of the unseen classes range from $[|\mathcal{C}_{known}|, |\mathcal{C}_{all}|)$. We then randomly select 5% of the samples from the known classes to construct the labeled dataset, while the remaining samples from the known classes, together with all samples from the unseen classes, constitute the unlabeled dataset. Thus, labeled data account for no more than 2.5% of the entire training set. Unless otherwise specified, the random seed is fixed at 42. The hyperparameter configurations for different datasets are presented in Table 1. More implementation details are provided in Appendix D.

Table 1: Configurations of *ROBIN*.

Dataset	d	λ_{ce}	λ_{ent}	λ_{mix}	τ	K/U
SEED-V	512	0.02	0.5	0.3	0.9	[0,2)/[2,5)
SEED-VII	512	0.9	0.07	0.3	0.7	[0,3)/[3,7)
GoEmotions	128	0.5	0.5	0.3	0.9	[0,9)/[9,18)
SE-R1510	32	0.3	0.3	0.3	0.5	[0,1)/[1,3)
SE-F1530	32	0.3	0.3	0.3	0.5	[0,1)/[1,3)

4.1.5 Research Questions

This section presents the experiments conducted to evaluate the performance of *ROBIN*. In the experimental results, bold indicates the **best result**, while underlining indicates the second-best result. We organize our analysis around the following research questions.

- **Superiority** — Does *ROBIN* outperform the baselines? (Section 4.2 and 4.3)
- **Ablation** — How does each module contribute to overall performance? (Section 4.4 and C.1)
- **Sensitivity** — How do configurations affect performance? (Section 4.5 and Appendix C.2)
- **Training** — How does performance evolve during training? (Section 4.6 and Appendix C.3)
- **Case Study** — Why can *ROBIN* outperform the baselines? (Appendix C.4)

4.2 Superiority: Overall Performance Comparison

This subsection presents a comparison between *ROBIN* and the baselines across all, known, and unseen classes. All methods are evaluated using five random seeds, i.e., 0, 1, 42, 1024, and 2026, and we report the corresponding means and standard deviations. The results are presented in Tables 2, 3, and 4, respectively. We observe that: **(1)** *ROBIN* consistently outperforms all baselines across all, known, and unseen classes, demonstrating its superiority. **(2)** On known classes, *ROBIN* is close to the fully supervised results on the SEED-V dataset (see Table 7), indicating that *ROBIN* can effectively handle both known and unseen class recognition and further demonstrating its robustness in open-world scenarios. Detailed results are provided in Appendix C.5.

4.3 Superiority: Comparison with New Class Discovery Baselines

To further validate the effectiveness of *ROBIN* in base-class prediction, we compared *ROBIN* with GCD and OpenNCD on the SEED-VII dataset and evaluated their classification performance on known and unseen classes separately. Experimental results are presented in Figure 3. The corresponding accuracies of GCD, OpenNCD, and *ROBIN* are 72.19%, 77.82%, and 99.22%, respectively. *ROBIN* substantially outperforms both new class discovery baselines in terms of classification performance on both known and unseen classes. These results demonstrate that the proposed base-class classification approach, based on consistency constraints, can more effectively alleviate the model's bias toward known classes and provide a reliable foundation for subsequent robust prediction. Detailed results are provided in Appendix C.5.

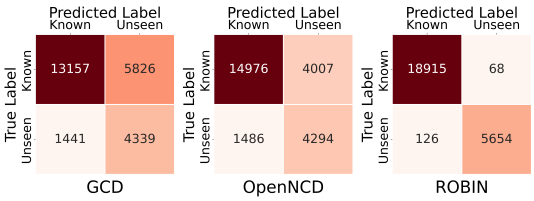


Figure 3: Classification performance comparison across known and unseen classes.

4.4 Ablation Study: Overall Performance

We conducted comprehensive ablation experiments on the SEED-V dataset and reported classification accuracy for all, known, unseen, and base classes after removing each component. For ablating the Equiangular Prototypes (EP), we replaced them with a learnable fully connected layer. For the remaining components, we ablated them by removing their corresponding loss terms. The results are presented in Figure 4. We observe that: **(1)** removing the equiangular prototypes causes the largest accuracy drops across all, unseen, and base classes, demonstrating their critical role in improving performance. **(2)** Removing the mean prototypes (i.e., the calibration loss $\mathcal{L}_{cal}$) has the most pronounced impact on known-class performance, demonstrating the effectiveness of mean prototypes in mitigating representation distribution shifts. **(3)** Removing either the base-class loss $\mathcal{L}_{bin}$ or the consistency constraint loss $\mathcal{L}_{mix}$ leads to a decline in base-class performance, indicating that the synergy between positive-class prediction and consistency constraints is crucial for distinguishing between known and unseen classes. **(4)** The contrastive loss $\mathcal{L}_{con}$ and entropy regularization loss $\mathcal{L}_{ent}$ also contribute positively to the overall performance.

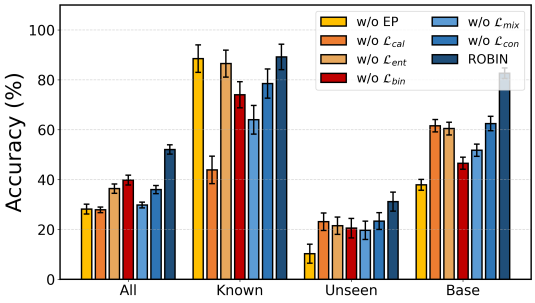


Figure 4: Contributions of individual components to classification performance.

4.5 Sensitivity Analysis: Class Proportion

We conducted sensitivity experiments on class partitioning on the SE-R1510 and SE-F1530 datasets. We report the classification accuracy of *ROBIN* on all, known, unseen, and base classes. Specifically, we sequentially designate one class as the known class and the remaining classes as unseen classes to construct different class-split settings. The experimental results are presented in Figure 5. The results show that *ROBIN* exhibits some fluctuations in accuracy on known and unseen classes, which is mainly attributed to differences in sample sizes across classes. In contrast, *ROBIN* demonstrates relatively stable performance on all and base classes, with the results across different class partitioning settings largely falling within the error range. Overall, *ROBIN* is relatively insensitive to the choice of class partitioning, and different partitioning schemes have only a minor impact on its performance. These results demonstrate the robustness of *ROBIN* under different class partitioning settings. More hyperparameter analyses are provided in Appendix C.2.

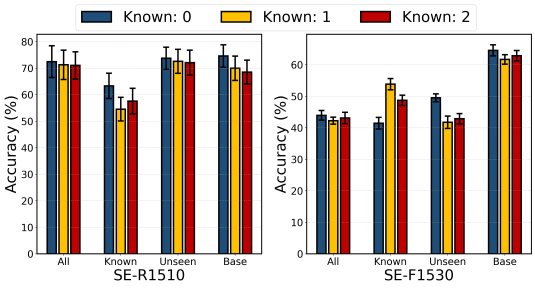


Figure 5: Sensitivity to class partitioning. Each class is sequentially set as the known class, while the other two are set as unseen classes.

4.6 Training Analysis: Evolution of Inter-Class Mean Cosine Similarity

To analyze the evolution of representation distributions during training, we record the inter-class mean cosine similarity after each training iteration on the SEED-V and SEED-VII datasets. Specifically, we record the mean representation of each class and calculate the cosine similarity for each pair of classes during training. We then average these pairwise cosine similarities as the overall inter-class mean cosine similarity. For comparison, we further report the trends of the inter-class mean cosine similarity after removing the equiangular prototypes and mean prototypes, respectively. The experimental results are shown in Figure 6. The experimental results show that the complete *ROBIN* achieves the most favorable cosine similarity among class mean representations, remaining closest to the optimal state. In contrast, removing either the equiangular prototypes or the mean prototypes causes the cosine similarity to deviate from the optimal value to some extent, indicating reduced separability among representations of different classes. These results demonstrate that *ROBIN* effectively enhances inter-class representation separability by introducing equiangular prototypes, thereby achieving more robust classification.

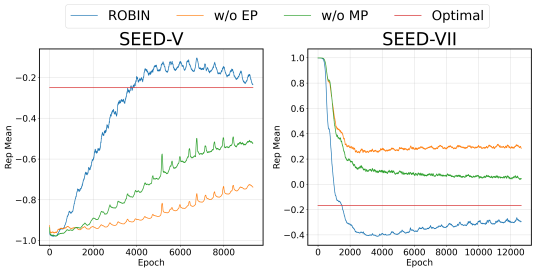


Figure 6: Trends in inter-class mean cosine similarity during training.

5 Conclusion

This study focuses on open-world semi-supervised emotion recognition. Due to the lack of supervision for unseen classes, existing methods generally suffer from two major limitations: representation collapse and biased prediction. Inspired by neural collapse, we propose *ROBIN*, an open-world semi-supervised emotion recognition framework. *ROBIN* introduces equiangular and mean prototypes to achieve representation alignment, thereby enhancing the separability of representations across different classes. Furthermore, we propose a progressive prediction strategy to achieve robust fine-grained recognition.

Limitation. To leverage equiangular prototypes for explicit localization of representations across different classes, *ROBIN* requires the total number of classes to be predetermined. Therefore, future work will further explore constructing optimally adapted equiangular prototypes when the total number of classes is unknown, so as to overcome this limitation.

AI USE STATEMENT

In this work, we used generative AI tools for literature exploration and language polishing. We have not used generative AI tools for research ideation, experimental design, data collection, data analysis, code generation, result generation, or the creation of figures, tables, or other research artifacts, and the remaining disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work and manually verified the cited literature, technical content, claims, and revised text for accuracy, originality, and consistency with our research. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=O-r8LOR-CCA.

Chengcheng Chen. How to estimate the emotions hidden behind spatio-temporal. 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Bo Cheng, Jueqing Lu, Yuan Tian, Haifeng Zhao, Yi Chang, and Lan Du. Cgmatch: A different perspective of semi-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15381–15391, June 2025.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Go emotions: A dataset of fine-grained emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4040–4054, Online, july 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372/.

P. Ekman, W. V. Freisen, and S. Ancoli. Facial signs of emotional experience. Journal of Personality and Social Psychology, 39:1125–1134, 1980. doi: 10.1037/h0077722.

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9284–9292, October 2021.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In L. Saul, Y. Weiss, and L. Bottou (eds.), Advances in Neural Information Processing Systems, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf.

Lan-Zhe Guo, Yi-Ge Zhang, Zhi-Fan Wu, Jie-Jing Shao, and Yu-Feng Li. Robust semi-supervised learning when not all classes have labels. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems, volume 35, pp. 3305–3317. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/15dce910311b9bd82ca24f634148519a-Paper-Conference.pdf.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. arXiv preprint arXiv:2002.05714, 2020.

Qintai Hu, Yifei Su, Fan Jiang, Xiaodi Huang, Qionghao Huang, and Changqin Huang. Nexpro: Multimodal negative expression prompting for open-set video-based facial expression recognition. Expert Systems with Applications, 331:133174, 2026. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2026.133174. URL https://www.sciencedirect.com/science/article/pii/S095741742602083X.

Sujia Huang, Lele Fu, Zhaoliang Chen, Tong Zhang, Xiaoli Li, and Zhen Cui. Adan: Adversarial distribution alignment network for multi-view semi-supervised classification. IEEE Transactions on Image Processing, 35:4861–4876, 2026. doi: 10.1109/TIP.2026.3689430.

Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Seed-vii: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. IEEE Transactions on Affective Computing, 16(2):969–985, 2025. doi: 10.1109/TAFFC.2024.3485057.

Bin Kang, Shuangshuang Wang, Zongyu Wang, Xin Li, Haie Dou, Lei Wang, and Zhijie Xia. Progressive masking oriented self-taught learning for occluded facial expression recognition. IEEE Transactions on Affective Computing, 16(3):1277–1289, 2025. doi: 10.1109/TAFFC.2025.3544677.

H. W. Kuhn. The hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Dongdong Li, Li Xie, Zhe Wang, and Hai Yang. Brain emotion perception inspired eeg emotion recognition with deep reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems, 35(9):12979–12992, 2024. doi: 10.1109/TNNLS.2023.3265730.

Shuzhen Li, Tong Zhang, and C. L. Philip Chen. Cyclic data distillation semi-supervised learning for multi-modal emotion recognition. IEEE Transactions on Knowledge and Data Engineering, 37(9):5078–5092, 2025a. doi: 10.1109/TKDE.2025.3581786.

Yaqi Li, Jing Jiang, Yuhang Zhang, Han Fang, Jiani Hu, and Weihong Deng. Ddl: Dynamic direction learning for semi-supervised facial expression recognition. IEEE Transactions on Affective Computing, 16(3):2274–2286, 2025b. doi: 10.1109/TAFFC.2025.3558884.

Zheng Lian, Bin Liu, and Jianhua Tao. Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition. IEEE Transactions on Affective Computing, 14(3):2415–2429, 2023. doi: 10.1109/TAFFC.2022.3141237.

Jinhao Lin, Yifei Wang, Yanwu Xu, and Qi Liu. Semi-iin: Semi-supervised intra-inter modal interaction learning network for multimodal sentiment analysis. Proceedings of the AAAI Conference on Artificial Intelligence, 39(2):1411–1419, Apr. 2025. doi: 10.1609/aaai.v39i2.32131. URL https://ojs.aaai.org/index.php/AAAI/article/view/32131.

Bingxiu Liu, Jifeng Guo, C. L. Philip Chen, Xia Wu, and Tong Zhang. Fine-grained interpretability for eeg emotion recognition: Concat-aided grad-cam and systematic brain functional network. IEEE Transactions on Affective Computing, 15(2):671–684, 2024a. doi: 10.1109/TAFFC.2023.3288885.

Jiaming Liu, Yangqiming Wang, Tongze Zhang, Yulu Fan, Qinli Yang, and Junming Shao. Open-world semi-supervised novel class discovery. IJCAI '23, 2023. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/445. URL https://doi.org/10.24963/ijcai.2023/445.

Ming-Hui Liu, Xiao-Qian Liu, Xin Luo, and Xin-Shun Xu. Data: Multi-disentanglement based contrastive learning for open-world semi-supervised deepfake attribution. IEEE Transactions on Multimedia, 27:8239–8250, 2025. doi: 10.1109/TMM.2025.3604932.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Transactions on Cognitive and Developmental Systems, 14(2):715–729, 2022. doi: 10.1109/TCDS.2021.3071170.

Yuanyuan Liu, Yuxuan Huang, Shuyang Liu, Yibing Zhan, Zijing Chen, and Zhe Chen. Open-set video-based facial expression recognition with human expression-sensitive prompting. MM '24, pp. 5722–5731, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681583. URL https://doi.org/10.1145/3664647.3681583.

Yuanyuan Liu, Shuyang Liu, Jiahao Zhang, Ke Wang, Chang Tang, Dapeng Tao, Zhe Chen, and Wei Xiang. Hinting the unknown: Effective open-set multimodal emotion recognition with a hierarchical cross-modal emotion-interactive prompting approach. Information Fusion, 136:104530, 2026. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2026.104530. URL https://www.sciencedirect.com/science/article/pii/S1566253526004094.

Yang Lu, Xiaolin Huang, Yizhou Chen, Mengke Li, Yan Yan, Chen Gong, and Hanzi Wang. Mood: Leveraging out-of-distribution data to enhance imbalanced semi-supervised learning. IEEE Transactions on Neural Networks and Learning Systems, 36(10):18545–18557, 2025. doi: 10.1109/TNNLS.2025.3573963.

Michael Munn, Benoit Dherin, and Javier Gonzalvo. The impact of geometric complexity on neural collapse in transfer learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), Advances in Neural Information Processing Systems, volume 37, pp. 66713–66742. Curran Associates, Inc., 2024. doi: 10.52202/079017-2131. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7b24015f3af598e1d9179f6e06353780-Paper-Conference.pdf.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences, 117(40): 24652–24663, 2020.

Yuxun Qu, Yongqiang Tang, Chenyang Zhang, Xiangrui Cai, Xiaojie Yuan, and Wensheng Zhang. Dual-space contrastive learning for open-world semi-supervised classification. IEEE Transactions on Neural Networks and Learning Systems, 36(8):14735–14748, 2025. doi: 10.1109/TNNLS.2025.3544405.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Mamshad Nayeem Rizve, Navid Kardan, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In European Conference on Computer Vision, pp. 382–401. Springer, 2022a.

Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. Towards realistic semi-supervised learning. In Computer Vision – ECCV 2022, pp. 437–455, Cham, 2022b. Springer Nature Switzerland. ISBN 978-3-031-19821-2.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 86–96, Berlin, Germany, aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009/.

Junao Shen, Qiyun Hu, Tian Feng, Xinyu Wang, Hui Cui, Sensen Wu, and Wei Zhang. In2nect: Inter-class and intra-class neural collapse tuning for semantic segmentation of imbalanced remote sensing images. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 6814–6822, 2025.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 35(15):13789–13797, May 2021. doi: 10.1609/aaai.v35i15.17625.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), Advances in Neural Information Processing Systems, volume 33, pp. 596–608. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf.

Haitian Sun, William W Cohen, and Lidong Bing. Semi-supervised learning with declaratively specified entropy constraints. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/12b1e42dc0746f22cf361267de07073f-Paper.pdf.

Yiyou Sun and Yixuan Li. Opencon: Open-world contrastive learning. In Transactions on Machine Learning Research, 2023. URL https://openreview.net/forum?id=2wWJxtpFer.

Liang Tan, Keping Yu, Long Lin, Xiaofan Cheng, Gautam Srivastava, Jerry Chun-Wei Lin, and Wei Wei. Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5g-enabled space-air-ground integrated intelligent transportation system. IEEE Transactions on Intelligent Transportation Systems, 23(3):2830–2842, 2022. doi: 10.1109/TITS.2021.3119921.

Peng Tang, Yitao Xu, Yutao Jiao, Maomao Zhang, Yehui Song, and Guoru Ding. Similarity-adaptive framework for semi-supervised open-world specific emitter identification. IEEE Transactions on Information Forensics and Security, 20:8266–8279, 2025. doi: 10.1109/TIFS.2025.3592529.

Geng Tu, Jintao Wen, Cheng Liu, Dazhi Jiang, and Erik Cambria. Context- and sentiment-aware networks for emotion recognition in conversation. IEEE Transactions on Artificial Intelligence, 3 (5):699–708, 2022. doi: 10.1109/TAI.2022.3149234.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7492–7501, June 2022.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9):10745–10759, 2023. doi: 10.1109/TPAMI.2023.3263585.

Youjun Wang, Boqi Li, Xin Zou, and Weiwei Liu. Neural collapse in multi-task learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), International Conference on Learning Representations, volume 2026, pp. 114780–114820, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/bab1486cec466c980b40e7d633dd4bbc-Paper-Conference.pdf.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6382–6388, Hong Kong, China, nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1670. URL https://aclanthology.org/D19-1670/.

Wuyou Xia, Guoli Jia, Sicheng Zhao, and Jufeng Yang. Seek common ground while reserving differences: Semi-supervised image-text sentiment recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 29601–29611, June 2025.

Jinheng Xie, Zhaochuan Luo, Rouyi Li, Yawen Huang, Haozhe Liu, Yuexiang Li, Yefeng Zheng, Yang Zhang, Linlin Shen, and Mike Zheng Shou. Open-world weakly-supervised object localization. Pattern Recognition, 169:111808, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.111808. URL https://www.sciencedirect.com/science/article/pii/S0031320325004686.

Yihan Xu, Hangyu Li, Mingrui Zhu, Nannan Wang, and Xinbo Gao. Boosting semi-supervised facial attribute recognition with dynamic threshold pairs. IEEE Transactions on Circuits and Systems for Video Technology, 35(7):6926–6935, 2025. doi: 10.1109/TCSVT.2025.3543408.

Hongren Yan, Yuhua Qian, Furong Peng, Jiachen Luo, Zheqing Zhu, and Feijiang Li. Neural collapse to multiple centers for imbalanced data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), Advances in Neural Information Processing Systems, volume 37, pp. 65583–65617. Curran Associates, Inc., 2024. doi: 10.52202/079017-2095. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/78a2ed8e7da1829d64fb3561adbd706b-Paper-Conference.pdf.

Bo Ye, Kai Gan, Tong Wei, and Min-Ling Zhang. Bridging the gap: Learning pace synchronization for open-world semi-supervised learning. In Kate Larson (ed.), Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, pp. 5362–5370. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/593. URL https://doi.org/10.24963/ijcai.2024/593. Main Track.

Weishan Ye, Zhiguo Zhang, Fei Teng, Min Zhang, Jianhong Wang, Dong Ni, Fali Li, Peng Xu, and Zhen Liang. Semi-supervised dual-stream self-attentive adversarial graph contrastive learning for cross-subject eeg-based emotion recognition. IEEE Transactions on Affective Computing, 16(1):290–305, 2025. doi: 10.1109/TAFFC.2024.3433470.

Guangyi Zhang, Vandad Davoodnia, and Ali Etemad. Parse: Pairwise alignment of representations in semi-supervised eeg learning for emotion recognition. IEEE Transactions on Affective Computing, 13(4):2185–2200, 2022a. doi: 10.1109/TAFFC.2022.3210441.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017.

Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. Real-time video emotion recognition based on reinforcement learning and domain knowledge. IEEE Transactions on Circuits and Systems for Video Technology, 32(3):1034–1047, 2022b. doi: 10.1109/TCSVT.2021.3072412.

Yuhang Zhang, Yue Yao, Xuannan Liu, Lixiong Qin, Wenjing Wang, and Weihong Deng. Open-set facial expression recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 646–654, 2024.

Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. Multimodal sentiment analysis with image-text interaction network. IEEE Transactions on Multimedia, 25:3375–3385, 2023. doi: 10.1109/TMM.2022.3160060.

A Proof of Proposition 1

A.1 Any Equiangular Prototypes Have Same Similarity

Proof. Let I_K denote the $K \times K$ identity matrix and J_K denote the $K \times K$ all-ones matrix. Let $O \in \mathbb{R}^{d\times K}$ be a column-orthogonal matrix satisfying $O^\top O = I_K$. Define $\theta = [e_1, \cdots, e_K] \in \mathbb{R}^{d\times K}$, where θ is constructed as:

$$\theta = \sqrt{\frac{K}{K-1}} O \left(I_K - \frac{J_K}{K} \right) \tag{16}$$

Let

$$M = \sqrt{\frac{K}{K-1}} \left(I_K - \frac{J_K}{K} \right) \tag{17}$$

then we have $\theta = OM$. The Gram matrix of θ is given by:

$$G = \theta^\top \theta = (OM)^\top OM = M^\top O^\top OM = M^\top I_K M = M^\top M \tag{18}$$

Thus, $G \in \mathbb{R}^{K\times K}$, and each element can be expressed as:

$$G = \begin{bmatrix} e_1^\top e_1 & \cdots & e_1^\top e_K \\ \vdots & \ddots & \vdots \\ e_K^\top e_1 & \cdots & e_K^\top e_K \end{bmatrix} \tag{19}$$

Let $L = I_K - \frac{J_K}{K}$, then its entries satisfy:

$$L_{i,j} = \begin{cases} \frac{K-1}{K}, & i = j \\ -\frac{1}{K}, & i \neq j \end{cases}, \qquad \forall i, j \in [1, K] \tag{20}$$

It is straightforward to verify that $L = L^\top$ and $L^2 = L$. Therefore,

$$G = M^\top M = \frac{K}{K-1} L^\top L = \frac{K}{K-1} L \tag{21}$$

Hence, the diagonal and off-diagonal elements of G are:

$$G_{i,i} = \frac{K}{K-1} \cdot \frac{K-1}{K} = 1, \qquad i = j \tag{22}$$

$$G_{i,j} = \frac{K}{K-1} \cdot \left(-\frac{1}{K} \right) = -\frac{1}{K-1}, \qquad i \neq j \tag{23}$$

Combining the above results, we obtain:

$$e_i^\top e_j = \begin{cases} 1, & i = j \\ -\frac{1}{K-1}, & i \neq j \end{cases}, \qquad \forall i, j \in [1, K] \tag{24}$$

According to the definition of cosine similarity:

$$\cos(e_i, e_j) = \frac{e_i^\top e_j}{\|e_i\| \|e_j\|} \tag{25}$$

Since $\|e_i\|^2 = e_i^\top e_i = 1$, we have $\|e_i\| = 1$. Therefore,

$$\cos(e_i, e_j) = e_i^\top e_j = \begin{cases} 1, & i = j \\ -\frac{1}{K-1}, & i \neq j \end{cases}, \qquad \forall i, j \in [1, K] \tag{26}$$

□

This demonstrates that any two equiangular prototypes share the same similarity.

A.2 Inter-Class Separation after Representation Alignment

Proof. Let μ_k denote the mean representation of class k. We assume that the learning process aligns each class mean with its corresponding equiangular prototype, such that $\mu_k = r_k e_k$, where $r_k > 0$ denotes a scaling factor. Then, for any two class mean representations μ_i and μ_j, their cosine similarity is given by:

$$\cos(\mu_i, \mu_j) = \frac{\mu_i^\top \mu_j}{\|\mu_i\|\|\mu_j\|} = \frac{r_i r_j e_i^\top e_j}{r_i r_j \|e_i\|\|e_j\|} = -\frac{1}{K-1} \tag{27}$$

□

This indicates that, as long as the learning objective aligns representations of different classes with their equiangular prototypes, the class means naturally achieve the theoretically optimal uniform inter-class angular separation, thereby alleviating representation collapse. This completes the proof.

B Algorithm Details

Algorithm 1 presents the pseudocode. N_{total} and N_{warm} are the total and warm-up iteration counts, respectively, $\mathcal{U} = \{\mu_k \in \mathbb{R}^d\}_{k=1}^K$ denotes the global mean representations that are continuously updated throughout the training for representation calibration after the warm-up, and ϵ denotes the maximum positive-class prediction probability output by $\phi_{\theta_{bin}}$ for unlabeled samples in the previous iteration, which enables dynamic adaptation during base-class prediction.

Algorithm 1: *ROBIN* Training

Input: Labeled dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^L$, unlabeled dataset $\mathcal{D}_u = \{x_i\}_{i=1}^U$, equiangular classifier $\phi_{\theta_{etf}}$, contrastive module $\phi_{\theta_{con}}$, binary classifier $\phi_{\theta_{bin}}$, and backbone ϕ_{θ_g}

// Initialize $\mathcal{U}$ and ϵ
$\mathcal{U} \leftarrow \left\{\mu_k \leftarrow \theta_{etf}^k\right\}_{k=1}^K$, $\epsilon \leftarrow 1$;
for $t \leftarrow 1$ **to** N_{total} **do**
  // Randomly sample and combine, where X denotes intermediate variables
  $X_l \leftarrow \text{Sample}(\mathcal{D}_l)$, $X_u \leftarrow \text{Sample}(\mathcal{D}_u)$, $X \leftarrow \text{Concat}(X_l, X_u)$;
  // Prediction and encoding
  $H_{etf(u)} \leftarrow \phi_{\theta_{etf}}\left(\phi_{\theta_g}\left(X_u; \theta_g\right); \theta_{etf}\right)$;
  $P_{bin}\left(\hat{y} = 1 \mid X\right) \leftarrow \text{SoftMax}\left(\phi_{\theta_{bin}}\left(\phi_{\theta_g}\left(X; \theta_g\right); \theta_{bin}\right)\right)$;
  // Calculate loss
  Calculate localization loss $\mathcal{L}_{loc}$ via Eq. 3;
  Calculate contrastive loss $\mathcal{L}_{con}$ via Eq. 11;
  Calculate base-class loss $\mathcal{L}_{bin}$ via Eq. 8;
  Calculate entropy regularization loss $\mathcal{L}_{ent}$ via Eq. 10;
  if $t \leq N_{warm}$ **then**
    Calculate total loss $\mathcal{L}_{total}$ via Eq. 13;
  else
    Calculate pseudo-labels Q via Eq. 9;
    Calculate calibration loss $\mathcal{L}_{cal}$ via Eq. 4;
    Calculate alignment loss $\mathcal{L}_{al}$ via Eq. 14;
    Calculate total loss $\mathcal{L}_{total}$ via Eq. 15;
  end
  Minimize total loss $\mathcal{L}_{total}$;
  // Update $\mathcal{U}$, ϵ
  for $k \in \mathcal{C}_{unseen}$ **do**
    $\mu_k \leftarrow \text{Mean}\left(H_{etf(u)}^k\right)$;
  end
  $\mathcal{U}_{t+1} \leftarrow \text{Norm}(\mathcal{U}_t)$, $\epsilon \leftarrow \max\left(P_{bin}\left(\hat{y} = 1 \mid X_u\right)\right)$;
end

During the initialization stage, the mean prototype of each class in $\mathcal{U}$ is initialized as its corresponding equiangular prototype θ_{etf}. During subsequent training iterations, the mean prototypes of known classes remain unchanged because once class labels are determined, the corresponding equiangular prototypes can be regarded as optimal class mean prototypes. For unseen classes, since their ground-truth labels are unavailable during training, their mean prototypes are dynamically updated using the corresponding logits $H^k_{etf(u)}$. During the warm-up stage, the localization loss $\mathcal{L}_{loc}$, contrastive loss $\mathcal{L}_{con}$, base-class loss $\mathcal{L}_{bin}$, and entropy regularization loss $\mathcal{L}_{ent}$ are jointly optimized to improve the model's base-class classification capability and pseudo-label quality. Building upon this initialization, the main training stage further introduces the calibration loss $\mathcal{L}_{cal}$ to calibrate the representations, thereby enabling more robust representation learning.

C Experiments

C.1 Ablation Study: Base-Class Classification

To further analyze the effectiveness of the proposed base-class prediction method, we conducted ablation experiments on the SEED-V and SEED-VII datasets to evaluate the contributions of the consistency loss $\mathcal{L}_{mix}$, base-class loss $\mathcal{L}_{bin}$, and Equiangular Prototypes (EP) to base-class classification performance. Specifically, we report the base-class classification confusion matrices and accuracies after removing each individual loss term or module. The experimental results are shown in Figure 7.

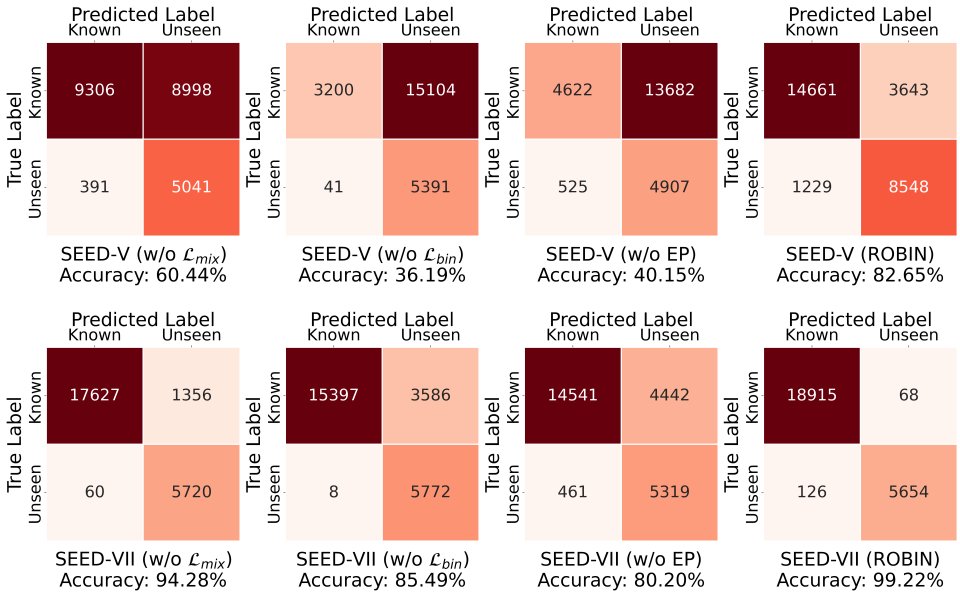


Figure 7: Contributions of the consistency loss $\mathcal{L}_{mix}$, base-class loss $\mathcal{L}_{bin}$, and equiangular prototypes to base-class classification performance.

The results demonstrate that: **(1)** the proposed equiangular prototypes and both loss terms have a significant impact on base-class classification performance. **(2)** Removing the equiangular prototypes leads to a substantial degradation in base-class classification performance, indicating that the explicit localization constraint imposed by the equiangular prototypes not only improves the overall performance of the model but also effectively enhances the separability between representations of known and unseen classes, thereby improving base-class classification performance. **(3)** Removing either $\mathcal{L}_{mix}$ or $\mathcal{L}_{bin}$ results in a decline in base-class classification performance, further validating the effectiveness of the proposed consistency-constrained base-class prediction method.

C.2 SENSITIVITY ANALYSIS: HYPERPARAMETER

To investigate the sensitivity of *ROBIN* to its hyperparameters, we conducted a hyperparameter sensitivity analysis on the SEED-VII dataset. Specifically, we adopted a controlled-variable approach, where only one hyperparameter is varied in each experiment while all other hyperparameters are kept at their default settings. Figure 8 reports the classification accuracy for all, known, unseen, and base classes under various hyperparameter configurations.

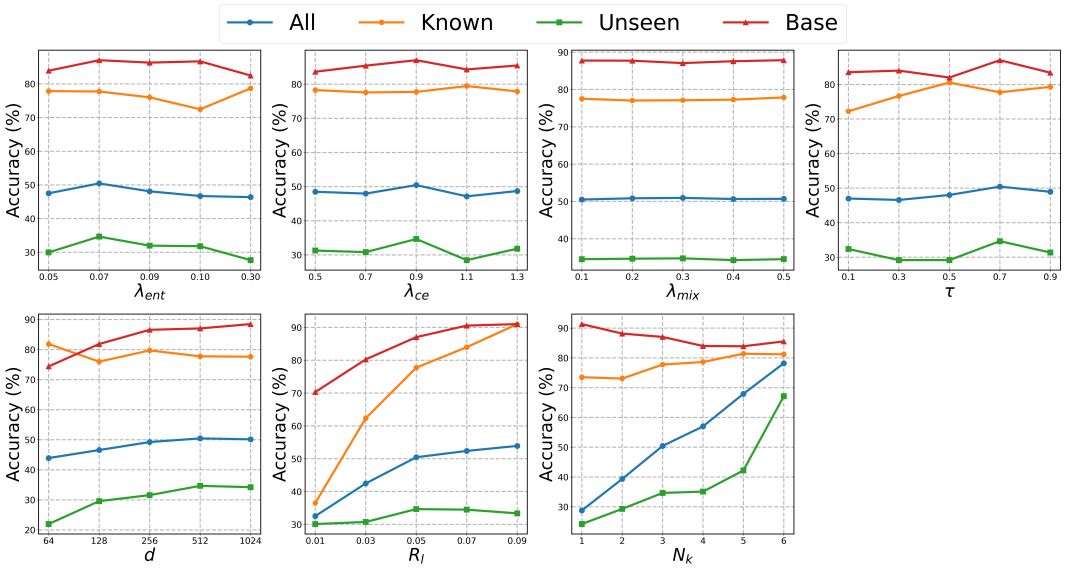


Figure 8: Sensitivity of *ROBIN* to different hyperparameters.

The experimental results indicate that: **(1)** when the weighting coefficients λ_{ent}, λ_{ce}, and λ_{mix}, and the temperature coefficient τ are set to 0.07, 0.9, 0.3, and 0.7, respectively, *ROBIN* achieves the highest accuracy across all four evaluation metrics. Among these hyperparameters, variations in λ_{ent}, λ_{ce}, and τ have relatively significant effects on the four metrics. **(2)** When λ_{mix} varies within the range of $[0.1, 0.5]$, its effect on the four metrics is relatively small, indicating that the overall performance of *ROBIN* is robust to the corresponding loss term. **(3)** When the representation dimension d is set to 512, *ROBIN* achieves the best performance on both all classes and unseen classes. For known classes, the optimal representation dimension is 64, whereas base-class classification achieves its best performance with $d = 1024$. This indicates that different modules have different requirements for the representation dimension d: an excessively small d may limit the model's representation capacity, whereas an excessively large d may lead to underfitting. **(4)** When the proportion of labeled data R_l varies within the range of $[0.01, 0.09]$, R_l exhibits an overall positive correlation with all four evaluation metrics. This demonstrates that *ROBIN* can effectively exploit labeled data, and even a modest increase in the proportion of labeled data can yield a significant improvement in the overall performance of the model. **(5)** Increasing the number of known classes N_k effectively improves the accuracy on all classes, known classes, and unseen classes, whereas the accuracy of base-class classification does not exhibit a positive correlation with N_k. This is because increasing N_k effectively changes the relative contributions of different loss terms to the total loss $\mathcal{L}_{total}$, thereby exerting different effects on the model optimization process.

In summary, with respect to R_l and N_k, the overall performance of *ROBIN* improves as either R_l or N_k increases, indicating that increasing labeled data or the number of known classes has an overall positive effect on model performance. For the weighting and temperature coefficients, appropriate combinations of hyperparameters should instead be selected according to the specific requirements of the target application.

C.3 Training Analysis: Model Performance

To analyze the evolution of model performance during training, we recorded the trends of the localization loss $\mathcal{L}_{loc}$, contrastive loss $\mathcal{L}_{con}$, base-class loss $\mathcal{L}_{bin}$, and the pseudo-label accuracy on the SEED-VII dataset. For pseudo-label accuracy, we report the results separately for all, known, unseen, and base classes to characterize the model's pseudo-label generation performance across different classes. The experimental results are presented in Figure 9. The results show that, as the number of training iterations increases, all loss terms exhibit an overall decreasing trend, indicating that model performance improves steadily throughout training. Meanwhile, the pseudo-label accuracy for all classes continuously increases, demonstrating that the model is able to progressively generate more reliable pseudo-labels and thus make more effective use of unlabeled data.

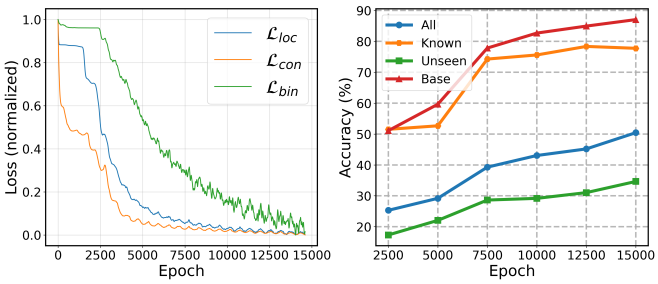
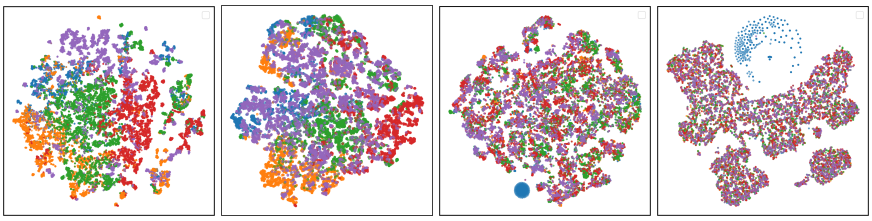


Figure 9: Performance evolution during training.

C.4 Case Study

This subsection further investigates the reasons behind *ROBIN*'s performance improvements over competing methods through a visualization analysis of the representation distributions. Specifically, we compare *ROBIN* with two semi-supervised methods, CGMatch (Cheng et al., 2025) and MOOD (Lu et al., 2025), and construct a variant of *ROBIN* without equiangular prototypes as a control model to directly examine the role of equiangular prototypes in representation alignment. For a fair comparison, all methods are trained using the same parameter settings, and t-SNE is employed to visualize the representation distributions in the embedding space.



(a) *ROBIN* (b) *ROBIN* (w/o. EP) (c) CGMatch (d) MOOD

Figure 10: Visualization comparison of representation distributions of different methods on the SEED-V dataset.

As shown in Figure 10a and Figure 10b, removing the equiangular prototypes substantially reduces the separability between representations from different classes, demonstrating that the equiangular prototypes effectively enhance inter-class separation and consequently yield a more distinct class structure. Furthermore, Figure 10c and Figure 10d reveal the limitations of existing methods in open-world scenarios. In the absence of effective supervision for unseen classes, representations of different unseen classes fail to establish clear inter-class boundaries and are therefore prone to collapsing into the same cluster, ultimately leading to degraded overall classification performance.

C.5 Detailed Performance Results

To provide a comprehensive evaluation of *ROBIN*, we further report its detailed performance across the five datasets, including classification performance for all, known, unseen, and base classes, as well as the corresponding confusion matrices. The detailed results are presented in Table 5 and Figure 11, respectively.

Table 5: Detailed classification performance of *ROBIN* (%)

Dataset	Metric	Accuracy	Precision	Recall	F1
SEED-V	All	$53.03_{\pm 1.87}$	$52.78_{\pm 2.10}$	$55.34_{\pm 0.78}$	$49.50_{\pm 2.01}$
	Known	$89.67_{\pm 5.10}$	$89.65_{\pm 3.52}$	$89.46_{\pm 5.67}$	$89.55_{\pm 5.59}$
	Unseen	$32.38_{\pm 3.80}$	$31.05_{\pm 3.02}$	$32.60_{\pm 3.48}$	$30.32_{\pm 3.14}$
	Base	$82.65_{\pm 2.12}$	$81.19_{\pm 1.03}$	$83.76_{\pm 0.69}$	$81.79_{\pm 1.59}$
SEED-VII	All	$50.44_{\pm 1.37}$	$53.61_{\pm 1.09}$	$52.09_{\pm 1.78}$	$52.73_{\pm 0.99}$
	Known	$77.74_{\pm 5.40}$	$76.99_{\pm 5.71}$	$77.20_{\pm 4.94}$	$77.07_{\pm 5.18}$
	Unseen	$34.65_{\pm 1.47}$	$33.96_{\pm 1.44}$	$33.26_{\pm 1.55}$	$33.42_{\pm 1.41}$
	Base	$99.22_{\pm 1.40}$	$99.07_{\pm 1.86}$	$98.73_{\pm 2.20}$	$98.90_{\pm 1.71}$
GoEmotions	All	$31.05_{\pm 0.34}$	$12.10_{\pm 0.20}$	$13.86_{\pm 0.31}$	$11.40_{\pm 0.19}$
	Known	$20.18_{\pm 0.46}$	$15.06_{\pm 0.48}$	$15.85_{\pm 0.79}$	$15.16_{\pm 0.70}$
	Unseen	$35.32_{\pm 0.44}$	$9.53_{\pm 0.17}$	$11.88_{\pm 0.80}$	$9.98_{\pm 0.18}$
	Base	$68.53_{\pm 0.96}$	$54.37_{\pm 0.52}$	$59.63_{\pm 0.45}$	$52.57_{\pm 0.04}$
SE-R1510	All	$72.46_{\pm 5.02}$	$48.09_{\pm 4.99}$	$53.31_{\pm 5.09}$	$50.04_{\pm 4.39}$
	Known	$63.33_{\pm 4.66}$	$50.00_{\pm 0.00}$	$31.67_{\pm 3.45}$	$38.78_{\pm 4.04}$
	Unseen	$73.76_{\pm 4.57}$	$37.86_{\pm 4.82}$	$48.30_{\pm 5.76}$	$42.45_{\pm 2.24}$
	Base	$74.59_{\pm 4.23}$	$52.61_{\pm 3.89}$	$87.12_{\pm 4.83}$	$47.56_{\pm 2.21}$
SE-F1530	All	$43.95_{\pm 1.77}$	$42.73_{\pm 1.29}$	$51.27_{\pm 1.09}$	$40.15_{\pm 1.76}$
	Known	$41.45_{\pm 1.32}$	$50.00_{\pm 0.00}$	$20.73_{\pm 1.23}$	$29.31_{\pm 1.85}$
	Unseen	$49.52_{\pm 1.52}$	$57.55_{\pm 1.47}$	$56.18_{\pm 1.84}$	$48.86_{\pm 1.43}$
	Base	$64.57_{\pm 1.88}$	$51.43_{\pm 1.17}$	$58.21_{\pm 1.08}$	$44.32_{\pm 1.26}$

D Implementation Details

D.1 Hyperparameter Configuration

ROBIN is implemented based on the PyTorch framework, and all experiments are conducted on Ubuntu 20.04. The computing platform is equipped with an AMD EPYC 7543 processor and an NVIDIA A100 GPU. The detailed hyperparameter settings of *ROBIN* are summarized in Table 6, and Adam is adopted as the optimizer for all datasets.

Table 6: Configurations of *ROBIN*.

Dataset	Wram-up/Total	LR	Batch	d	λ_{ce}	λ_{ent}	λ_{mix}	τ	Known/Unseen	ω_w/ω_s
SEED-V	30/200	1e-4	1024	512	0.02	0.5	0.3	0.9	[0,2)/[2,5)	0.01/0.02
SEED-VII	30/300	1e-4	1024	512	0.9	0.07	0.3	0.7	[0,3)/[3,7)	0.01/0.03
GoEmotions	5/15	1e-3	32	128	0.5	0.5	0.3	0.9	[0,9)/[9,18)	-
SE-R1510	300/1000	1e-3	48	32	0.3	0.3	0.3	0.5	[0,1)/[1,3)	0.05/0.1
SE-F1530	300/1000	1e-3	30	32	0.3	0.3	0.3	0.5	[0,1)/[1,3)	0.05/0.1

When partitioning the known and unseen classes, for each dataset, we designate the first $|\mathcal{C}_{known}|$ classes as known classes and the remaining $|\mathcal{C}_{all}| - |\mathcal{C}_{known}|$ classes as unseen classes, rather than adopting a random partitioning strategy. This is because random partitioning can yield $\binom{|\mathcal{C}_{known}|}{|\mathcal{C}_{all}|}$ possible class combinations, thereby introducing additional uncertainty into the experimental evaluation. To ensure fairness and stability, we therefore adopt a fixed class partitioning scheme.

D.2 Backbones and Classifiers

For each dataset, we adopt the corresponding backbone for representation extraction, as provided by the respective dataset creators. Specifically, a convolutional neural network is used for SEED-V,

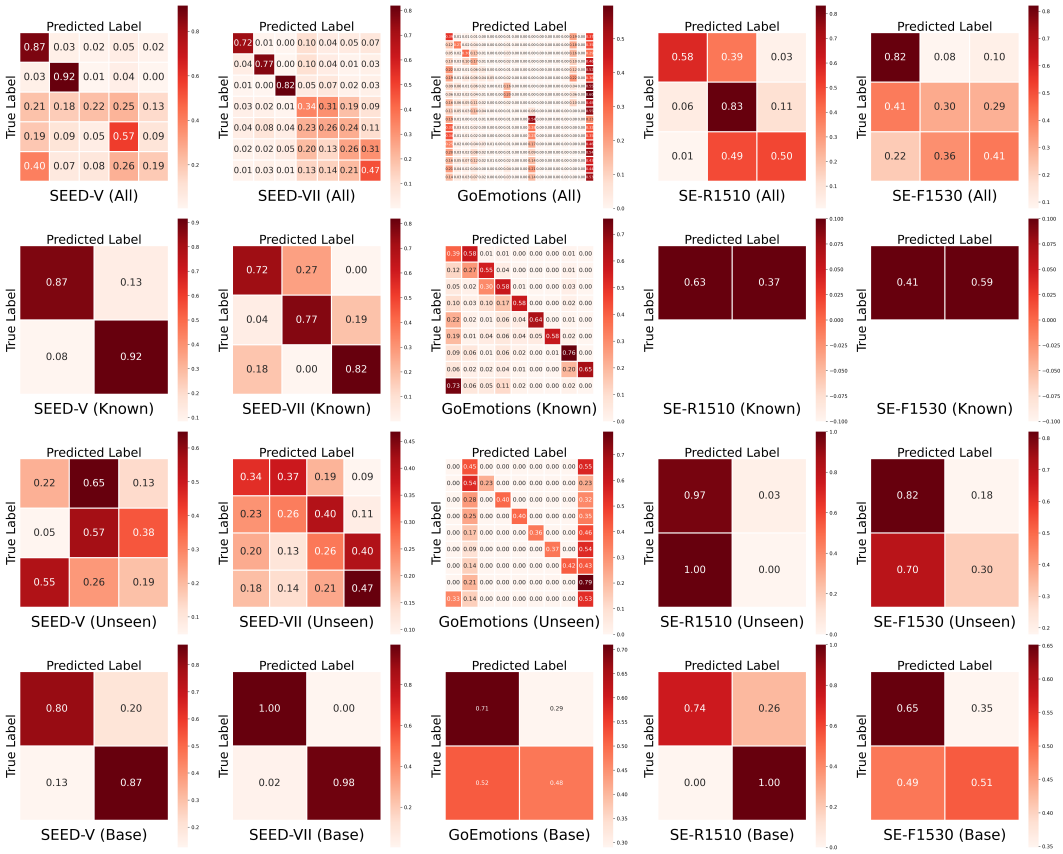


Figure 11: Confusion matrices of *ROBIN*.

SEED-VII is processed using MEAT (Jiang et al., 2025), GoEmotions is encoded using DistilBERT (Sanh et al., 2019), and both SE-R1510 and SE-F1530 are modeled using SADS (Chen, 2023). To ensure a fair comparison, all baseline methods use exactly the same backbone. The backbone hyperparameters are not additionally tuned or optimized for any specific method; instead, they are uniformly determined based on the classification performance reported in Table 7 under the fully supervised setting. In the fully supervised experiments, a fully connected layer serves as the classifier for all datasets.

Table 7: Performance of the backbone under the fully supervised setting (%).

Dataset	Accuracy	Precision	Recall	F1
SEED-V	$92.88_{\pm 6.27}$	$94.11_{\pm 5.00}$	$92.27_{\pm 6.41}$	$93.01_{\pm 6.19}$
SEED-VII	$85.97_{\pm 5.05}$	$86.63_{\pm 1.40}$	$87.36_{\pm 4.50}$	$85.71_{\pm 5.12}$
GoEmotions	$43.43_{\pm 0.23}$	$32.70_{\pm 0.88}$	$17.20_{\pm 0.30}$	$18.34_{\pm 0.24}$
SE-R1510	$73.06_{\pm 4.30}$	$67.42_{\pm 3.76}$	$63.83_{\pm 5.20}$	$65.42_{\pm 4.95}$
SE-F1530	$57.64_{\pm 1.58}$	$52.19_{\pm 1.91}$	$52.17_{\pm 1.97}$	$52.16_{\pm 1.99}$

ROBIN further introduces three sub-modules for representation alignment, contrastive learning, and base-class partitioning, respectively. The detailed architecture of these components is presented in Table 8. Furthermore, to improve numerical stability during training, we normalize the representations output by the backbone ϕ_{θ_g} and the contrastive learning module $\phi_{\theta_{con}}$, respectively.

D.3 Data Augmentation

To fully exploit unlabeled data, we incorporate contrastive learning, which is widely used in unsupervised training, into the training process. For the SEED-V, SEED-VII, SE-R1510, and SE-F1530, data augmentation is performed by adding random Gaussian noise $\mathcal{N}(0, \omega \cdot \sigma^2)$ to the original sam-

Table 8: Architecture of each module.

Module	Layer	Input	Ouput	Activate
$\phi_{\theta_{etf}}$	Linear	d	K	-
$\phi_{\theta_{con}}$	Linear	d	d	ReLU
	Linear	d	d	-
$\phi_{\theta_{bin}}$	Linear	d	$2d$	ReLU
	Linear	$2d$	$2d$	ReLU
	Linear	$2d$	2	-

ples. The weak noise strength ω_w and strong noise strength ω_s for different datasets are specified in Table 1. For the text dataset GoEmotions, we adopt back-translation (Sennrich et al., 2016) and synonym replacement (Wei & Zou, 2019) to generate augmented samples, thereby constructing weakly and strongly augmented views, respectively.

D.4 Mixup Implementation

To effectively distinguish between known and unseen classes, we introduce a Mixup-based consistency constraint to mitigate the biased predictions of $\phi_{\theta_{bin}}$ towards known classes. Specifically, we sample a mixing coefficient ρ from a Beta distribution with parameter $\alpha = 0.3$, i.e., $\rho \sim \mathrm{Beta}(0.3, 0.3)$. Subsequently, for the labeled samples X_l and unlabeled samples X_u in the current batch, we perform linear interpolation using ρ to construct the mixed samples $\widetilde{X}$ and corresponding mixed labels $\widetilde{Y}$:

$$\widetilde{X} = \rho \cdot X_u + (1 - \rho) \cdot X_l \tag{28}$$

$$\widetilde{Y} = \rho \cdot P_{bin}\left(\hat{y} = 1 \mid X_u\right) + (1 - \rho) \cdot \mathbf{1} \tag{29}$$

Since $\phi_{\theta_{bin}}$ is trained to predict labeled samples as the positive class as accurately as possible (i.e., with the output probability approaching 1) while predicting unlabeled samples as the negative class, when constructing mixed labels, the positive-class label for labeled samples is represented by the one-hot vector $\mathbf{1} = [0, 1]^\top$, whereas unlabeled samples are assigned the positive-class probability $P_{bin}\left(\hat{y} = 1 \mid X_u\right)$ predicted by $\phi_{\theta_{bin}}$.

E Discussion and Comparison

E.1 Why Does Base-Class Prediction Work?

The proposed base-class prediction first determines whether an input sample belongs to the labeled or unlabeled set. However, since the unlabeled data contain samples from both known and unseen classes, directly treating the negative class as the unseen class would result in a large number of misclassifications. To address this issue, we further introduce a Mixup-based consistency constraint that constructs continuous mixed samples and corresponding soft labels between labeled known-class samples and unlabeled samples, while constraining the binary classifier's log-probability predictions to be consistent with those of the mixed labels. This constraint encourages the model to learn varying degrees of "knownness" for known-class samples, thereby smoothing the decision boundary between known and unseen classes and reducing the interference of known-class samples in the unlabeled data on unseen-class recognition. Consequently, the binary classifier no longer simply equates unlabeled samples with unseen classes. Furthermore, the ablation study demonstrates that introducing equiangular prototypes to enhance representation separability across classes can also substantially improve base-class prediction performance (see Appendix C.1). These two components jointly enhance the separability between known and unseen classes, thereby enabling effective base-class prediction in the open-world setting.

E.2 Comparison with Existing Methods

Open-world emotion recognition has recently received increasing attention due to its ability to recognize unseen emotion classes in open-world scenarios (Zhang et al., 2024; Liu et al., 2024b; 2026; Hu et al., 2026). Although these studies are related to our work, our method differs substantially

from existing approaches. First, the method in (Zhang et al., 2024) primarily focuses on scenarios where only a few unseen classes are missing from the training set. Methods (Liu et al., 2026; Hu et al., 2026) treat all unseen classes as a single class and therefore lack the ability to perform fine-grained classification among unseen emotion classes. More importantly, all the aforementioned methods rely on high-quality sample annotations, while methods (Liu et al., 2024b; 2026; Hu et al., 2026) additionally require textual prompts. These limitations constrain the practical applicability of existing methods in open-world settings.

In contrast, the proposed *ROBIN* offers the following advantages. First, during training, *ROBIN* requires that known classes account for no more than 50% of all classes, while the labeled data ratio is no more than 5% of the samples from known classes, enabling model training under substantially weaker supervision. Second, during inference, *ROBIN* not only distinguishes between known and unseen classes but also further performs fine-grained recognition of different unseen classes. Compared with methods that simply group all unseen classes into a single class, *ROBIN* thus provides more challenging and comprehensive classification capabilities. Overall, *ROBIN* reduces the reliance on supervisory information during training while enabling more fine-grained and comprehensive open-world emotion recognition, making it more suitable for real-world applications.

E.3 Novelty

We are the first to extend semi-supervised emotion recognition to the open-world setting. Among existing methods, contrastive learning-based approaches for unseen class discovery have become a mainstream paradigm (Sun & Li, 2023; Vaze et al., 2022; Liu et al., 2023). However, due to the inherent semantic similarity among emotional samples, representations from different classes may not exhibit sufficiently distinct boundaries, which can limit the effectiveness of such soft-separation approaches. In contrast, *ROBIN* introduces an ETF as equiangular prototypes and incorporates mean prototypes for representation alignment, thereby guiding the representations of different classes to form a more clearly separated structure.

Existing new class discovery methods typically assume that labeled samples contain only known classes, while unlabeled samples contain only unseen classes (Fini et al., 2021; Han et al., 2019; 2020). However, in the open-world setting, labeled samples contain only known classes, whereas unlabeled samples span all classes. To address this issue, we design a consistency-constrained base-class prediction method and employ it as an auxiliary classifier to first effectively distinguish between known and unseen classes in the open-world setting, thereby enhancing the robustness of subsequent fine-grained class prediction.

F Dataset Details

To validate the performance of *ROBIN*, we conduct experiments on five emotion datasets spanning four different emotional modalities, all of which have been widely adopted in emotion recognition research. Specifically, SEED-V (Liu et al., 2022) is an EEG dataset represented as three-dimensional tensors and contains five emotion classes. SEED-VII (Jiang et al., 2025) is a multimodal dataset comprising EEG and eye-tracking signals, with both modalities represented as high-dimensional vectors, and contains seven emotion classes. GoEmotions (Demszky et al., 2020) is a comment dataset consisting of English-language text and originally contains 28 emotion classes. To eliminate class ambiguity, following the recommendations of the dataset collectors, we further remove samples associated with multiple emotion classes, resulting in a subset covering 18 emotion classes. SE-R1510 (Chen, 2023) and SE-F1530 (Chen, 2023) are social behavior datasets represented as multivariate time series, each containing three emotion classes. Detailed information of each dataset is summarized in Table 9.

- **SEED-V** (Liu et al., 2022): This dataset records EEG signals from 16 subjects while they watch film clips. After each session, subjects are required to report their emotional states. It contains five emotion classes: happiness, sadness, neutral, disgust, and fear. The dataset is available at: `https://bcmi.sjtu.edu.cn/home/seed/seed-v_cn.html`.
- **SEED-VII** (Jiang et al., 2025): This dataset collects both EEG and eye-tracking signals from 20 subjects during film viewing. After each session, subjects provide self-reported emotion labels. It includes seven emotion classes: happiness, sadness, neutral, disgust, fear, surprise, and

Table 9: Detailed information of the datasets. SEED-V is an EEG dataset represented as 3D tensors. SEED-VII is a multimodal dataset consisting of EEG and eye-tracking signals, represented as high-dimensional vectors. GoEmotions is a comment-based dataset represented in English text. SE-R1510 and SE-F1530 are societal behavior datasets represented as multivariate time series.

Dataset	#Items	#Participants	#Labels	Modality
SEED-V	29,168	16	5	EEG signal
SEED-VII	69,742	20	7	EEG and EYE signal
GoEmotions	58,009	-	18	English text
SE-R1510	714,587	15	3	Societal behavior
SE-F1530	1,815,531	15	3	Societal behavior

anger. The dataset can be downloaded from: `https://bcmi.sjtu.edu.cn/home/seed/seed-vii_cn.html`.

- **GoEmotions** (Demszky et al., 2020): This dataset consists of comments extracted from popular English subreddits on Reddit and annotated manually. It contains 28 emotion classes, including 12 positive emotions, 11 negative emotions, 4 ambiguous emotion classes, and 1 neutral class. It is widely used for conversational understanding tasks requiring fine-grained emotion recognition. The dataset is publicly available via the HuggingFace datasets library.
- **SE-R1510** (Chen, 2023): This dataset contains spatiotemporal trajectories of 15 participants recorded over 10 to 12 consecutive days. During the data collection period (6:00-24:00 daily), participants performed at least three self-reported emotion evaluations each day. Throughout this period, all participants' events were restricted to within a school campus to investigate the impact of limited mobility on emotions.
- **SE-F1530** (Chen, 2023): This dataset includes spatiotemporal trajectories of 15 participants recorded over 29 to 30 consecutive days. During the data collection period (6:00-24:00 daily), participants performed at least three self-reported emotion evaluations each day. In contrast to SE-R1510, participants' events were unrestricted, allowing for the study of the effects of diverse societal events under normal conditions on emotional states.

G Baseline Details

The selected state-of-the-art baselines cover both representative semi-supervised learning approaches and open-world semi-supervised learning methods. Among these baselines, FixMatch, CGMatch, and MOOD are representative semi-supervised learning methods. GCD (clustering-based), TRSSL (out-of-distribution-based), OpenCon (contrastive learning-based), and OpenNCD (contrastive learning-based) are representative open-world semi-supervised learning approaches. Detailed descriptions are provided as follows:

- **FixMatch** (Sohn et al., 2020): A simple yet representative semi-supervised learning method. Its core idea is to generate pseudo-labels based on predictions of weakly augmented unlabeled samples during training. Only high-confidence predictions are retained as pseudo-labels. The model is then trained to predict these pseudo-labels on strongly augmented versions of the same samples. The official implementation is available at: `https://github.com/google-research/fixmatch`.
- **GCD** (Vaze et al., 2022): A classic open-world semi-supervised learning method that adapts unseen class discovery techniques to unseen class recognition. It leverages a Transformer backbone combined with contrastive representation learning to handle open-world scenarios. A simple yet effective semi-supervised K-means clustering strategy is then applied to automatically partition unlabeled data into known and unknown classes. The official code is available at: `https://www.robots.ox.ac.uk/~vgg/research/gcd`.
- **TRSSL** (Rizve et al., 2022b): An open-world semi-supervised learning method designed for class-imbalanced datasets. It leverages sample uncertainty and incorporates prior knowledge of class distribution to generate reliable and distribution-aware pseudo-labels for both known and unseen data. The official code is available at: `https://github.com/nayeemrizve/TRSSL`.

- **OpenCon** (Sun & Li, 2023): An open-world semi-supervised learning method that addresses the challenge of learning compact representations for both known and unseen classes via contrastive learning. It can be rigorously interpreted from an EM perspective, where minimizing the contrastive loss partially maximizes the likelihood by clustering similar samples in the embedding space. The official code is available at: `https://github.com/deeplearning-wisc/opencon`.
- **OpenNCD** (Liu et al., 2023): An open-world semi-supervised learning method based on progressive dual-level contrastive learning with multiple prototypes. It consists of two mutually reinforcing components: a dual-level contrastive learning strategy that preserves both pairwise prototype similarities and group-level prototype relationships for improved representation learning, and a reliable prototype similarity metric that progressively groups highly similar prototypes for known class recognition and unseen class discovery. The official code is available at: `https://github.com/LiuJMzzZ/OpenNCD`.
- **CGMatch** (Cheng et al., 2025): A semi-supervised learning method aimed at improving pseudo-label reliability. It introduces a novel metric called the count gap, which effectively identifies unlabeled samples beneficial for training. CGMatch further proposes a fine-grained dynamic selection strategy that partitions unlabeled data into three subsets—easy, ambiguous, and hard samples. By selectively filtering subsets and applying corresponding regularization, it mitigates the impact of incorrect pseudo-labels. The official code is available at: `https://github.com/BoCheng-96/CGMatch`.
- **MOOD** (Lu et al., 2025): A semi-supervised learning method designed for class-imbalanced scenarios. MOOD first filters out out-of-distribution samples from unlabeled data and integrates them with labeled data to enhance representation diversity. To prevent representation overlap with out-of-distribution samples, a push-pull loss is introduced, which pulls in-distribution samples toward their class centers while pushing away out-of-distribution samples. The official code is available at: `https://github.com/xlhuang132/MOODv2`.

H Related Work

H.1 Semi-supervised Emotion Recognition

Semi-supervised emotion recognition is an important research direction in the field of affective computing. For the widely studied EEG-based emotion recognition task, Ye et al. (2025) proposed a semi-supervised framework that integrates self-attention mechanisms with contrastive learning, aiming to alleviate the scarcity of labeled data in cross-subject EEG emotion recognition. Building upon this, Zhang et al. (2022a) introduced a representation alignment method based on data augmentation and convex combination sharpening to mitigate the potential distribution mismatch between abundant unlabeled data and limited labeled samples. In the context of multimodal semi-supervised emotion recognition, Lian et al. (2023) proposed a modality interaction integration framework that effectively exploits unlabeled data by modeling both intra-modal and cross-modal interactions to extract emotion-relevant representations. Xia et al. (2025) further decomposed multimodal representations into shared and private representations, leveraging private representations to train unimodal classifiers for enhanced modality-specific representations, and subsequently achieving adaptive multimodal fusion via a modality-selection-based attention mechanism. To further improve pseudo-label quality, Li et al. (2025a) employed multiple pre-trained unimodal teacher models combined with a confidence-enhanced pseudo-labeling strategy, thereby improving the reliability of pseudo-label predictions in multimodal ensemble models. Lin et al. (2025) adopted a self-training approach that incorporates masked attention mechanisms and gated fusion modules to model intra- and inter-modal interactions, enabling dynamic and effective information selection. For facial expression recognition, Kang et al. (2025) integrated convolutional neural networks with Transformer architectures within a multi-scale framework, achieving weakly supervised facial expression recognition under occlusion scenarios. Xu et al. (2025) addressed the issue that a single sample may correspond to multiple candidate pseudo-labels by proposing a dynamic threshold triplet mechanism based on pseudo-label prediction, which guides pseudo-label generation for unlabeled samples. In addition, Li et al. (2025b) tackled the class imbalance problem in semi-supervised facial expression recognition by dynamically adjusting self-training constraints based on the performance of a balanced validation set, resulting in a more stable and balanced training process.

In summary, semi-supervised emotion recognition has been extensively studied across various affective analysis tasks. However, these methods are predominantly developed under a closed-world assumption, where all emotion classes are presumed to be fully covered in the labeled data, thereby neglecting the presence of fine-grained emotion classes that are difficult to collect under laboratory conditions (Demszky et al., 2020). Consequently, existing approaches generally lack the capability to recognize and model unseen emotion classes in an open-class space.

H.2 Open-world Semi-supervised Learning

Open-world semi-supervised learning has recently emerged as an active research direction, aiming to address the problem that labeled data in real-world scenarios typically covers only a subset of classes. To handle the requirement of recognizing unseen classes, early studies proposed unseen class discovery methods (Fini et al., 2021; Han et al., 2019; 2020), which leverage labeled data to assist in mining new classes from unlabeled data. However, these approaches generally assume that all unlabeled data originate exclusively from unseen classes, making them unsuitable for more general settings where unlabeled data may contain both known and unknown classes. To address this limitation, early methods primarily rely on sample-level similarity for clustering (Guo et al., 2022; Ye et al., 2024), and generate pseudo-labels based on the resulting clusters to achieve class partitioning. Building upon contrastive learning, Qu et al. (2025) proposed a semi-supervised contrastive learning framework that integrates representation space and prediction space, enhancing both intra-class compactness and inter-class separability. Xie et al. (2026) further introduced contrastive representation collaborative learning, which jointly exploits labeled and unlabeled data and incorporates a generalized class activation mechanism, thereby improving overall recognition performance.

Although existing methods have achieved certain progress in unseen class discovery, they typically rely on spontaneous clustering based on sample similarity, overlooking the potential risk that representations of different classes may collapse into the same cluster in the absence of supervision for unseen classes, leading to representation collapse. Moreover, most existing approaches directly assign pseudo-labels to unlabeled samples based on representation similarity. In scenarios with significant class imbalance or bias, models tend to incorrectly assign unlabeled samples to known classes, resulting in biased predictions.

H.3 Open-World Emotion Recognition

Open-world emotion recognition has recently emerged as a new research task, aiming to recognize unseen emotion classes in open-world scenarios. Zhang et al. (2024) proposed a noise-label-based sample selection method that first uses a model trained on known classes to generate pseudo-labels for unseen-class samples. Subsequently, two collaborative teacher models iteratively select high-quality samples and perform model training, thereby effectively separating known and unseen classes. Liu et al. (2024b) proposed a CLIP-based (Radford et al., 2021) open-world video facial emotion recognition method that employs text and visual prompt modules to capture emotion-related patterns in videos and complements the original CLIP inputs with these patterns, thereby enhancing the model's representation capability for both known and unseen emotion classes. For open-world multimodal emotion recognition, Liu et al. (2026) proposed a hierarchical cross-modal emotion recognition framework that performs emotion classification and unseen emotion detection through multimodal representation alignment and similarity measurement. Hu et al. (2026) proposed a CLIP fine-tuning-based multimodal emotion recognition framework that enhances CLIP's classification capability for known emotion classes and detection capability for unseen emotion classes by designing more comprehensive and fine-grained textual prompts.

Although existing methods have explored open-world emotion recognition, they still suffer from several limitations. Most existing approaches treat all unseen emotion classes as a single class, lacking the ability to further distinguish among unseen emotion classes. More importantly, these methods typically rely on high-quality supervisory information in the training set, such as human annotations or textual prompts, which limits their applicability in semi-supervised or unsupervised settings.