Adversarial Metric Learning for Fine-Grained Emotion Classification

Junfan Chen[1,2], **Sizhe Wu**[2], **Richong Zhang**[1,3*], **Chunming Hu**[1,2,3]
[1]CCSE, School of Computer Science and Engineering, Beihang University, Beijing, China
[2]School of Software, Beihang University, Beijing, China
[3]Zhongguancun Laboratory, Beijing, China
{chenjf, zhangrc}@act.buaa.edu.cn, {20376053, hucm}@buaa.edu.cn

Abstract

Fine-grained emotion classification (FEC) requires distinguishing subtly different emotions, where the dominant errors come from closely confusable categories. Recent progress relies on contrastive learning with hard-pair mining, implicitly assuming that a fixed similarity metric is sufficient to optimize informative pairs. We argue that this assumption is fragile because defining whether two utterances are similar becomes a problem when the label space is crowded, and hard-pair mining under a fixed metric can systematically miss the worst confusions. Thus, we treat the similarity function as a learnable component and design an adversarial metric learning (AML) framework. It follows theoretical interpretations of metric-robust representations that better separate confusable emotions. AML trains a pairwise discriminator to maximally confuse two targeted hard pair types, while training the encoder to remain discriminative under this worst-case learned metric. Our code and data are released on GitHub[1].

1 Introduction

Emotion classification is an essential task for emotion-aware NLP, supporting applications such as conversational agents (Chen and Shi, 2025; Shen et al., 2025), empathetic systems (Li et al., 2021; Yuan et al., 2025), and affective analytics (Liu et al., 2024a; Mittal et al., 2021). While coarse-grained emotion recognition focuses on a small set of basic categories (Paul, 1992; Plutchik, 1980), *fine-grained emotion classification* (FEC) aims to capture the richer and more nuanced emotional space humans express in language (Demszky et al., 2020; Rashkin et al., 2019). This shift in granularity changes the nature of the learning problem: many labels overlap semantically, and the hardest cases are precisely those where different emotions are *genuinely similar* in content and expression.

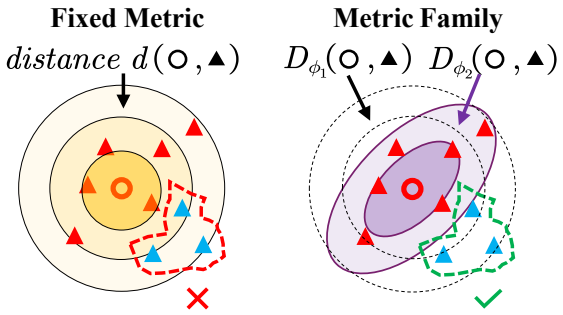

Figure 1: Training with a fixed metric (existing approaches) and learnable metrics (our AML framework).

A common strategy to tackle this challenge is to strengthen representation geometry through contrastive learning (Khosla et al., 2020; Gunel et al., 2021; Suresh and Ong, 2021). In FEC, recent methods further explore mining *strong* positives and negatives, for example using clustering to identify intra-class pairs that should be pulled together and inter-class pairs that should be pushed apart (Gong et al., 2025). These approaches embody an important principle for FEC: *aggregate intra-class instances even when they look different, and separate inter-class instances even when they look similar.*

We revisit this principle from a viewpoint that is often implicit but rarely questioned: the *choice of similarity metric*. As shown in Figure 1, most hard-pair contrastive pipelines rely on a fixed, hand-chosen metric (e.g., cosine similarity) to decide which pairs are "hard" and define the learning signal that reshapes the space. This design is convenient, but it creates a fundamental mismatch. That is, when confusions are subtle, a ***fixed metric*** can be *under-expressive* (unable to reflect task-relevant confusability) and *self-reinforcing* (training on imperfect geometry amplifies early biases), i.e., in FEC the bottleneck is not only *insufficient separation*, but also *metric mis-specification* (the objective may optimize the wrong notion of similarity).

* Corresponding author: zhangrc@act.buaa.edu.cn
[1]https://github.com/Andromeda784/AML

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45087–45099

This observation motivates a simple but consequential question: *If fine-grained confusions are difficult because similarity is subtle and context-dependent, should we assume a fixed metric when learning representations meant to resolve those confusions?* Our answer is *no* and instead the metric itself should be ***learnable*** and ***adversarial***.

In this paper, we propose an *adversarial metric learning* (AML) framework for FEC. As shown in Figure 1, instead of using a fixed similarity function, AML introduces a pairwise discriminator to model a similarity metric family. It is built upon a min-max training objective, which optimizes the encoder to maintain intra-class aggregation and inter-class separation under the worst-case learned similarity. The design principle of AML follows concrete *theoretical* rules that guarantee *worst-case metric robustness* for hard pairs on logit space.

Crucially, the discriminator is trained as an *opponent*. It actively seeks a similarity notion that maximizes confusion on the most diagnostic pair types in FEC. We construct two targeted hard-pair sets: 1) *intra-class but semantically dissimilar* pairs that should still be aggregated, and 2) *inter-class but semantically similar* pairs that must be separated. The discriminator tries to collapse these distinctions, while the encoder learns representations that preserve them despite the discriminator's strongest confusion behavior. This turns hard-pair learning problem into a *robustness problem*, i.e., we do not merely optimize separation under one metric, but learn representations that remain discriminative under a family of learned metric distortions.

This perspective yields a principled upgrade over metric-fixed hard-pair mining. Rather than relying on a single, potentially misaligned similarity function, AML trains the encoder against an adaptive adversary that continuously exposes the model's weakest confusions. As a result, the learned space is encouraged to support the desired fine-grained structure (intra-class coherence and inter-class separation) in a metric-robust way, precisely where confusable emotions overlap the most.

To summarize, we make the following contributions: 1) We revisit FEC task from a *metric robustness* perspective and reveal the similarity mismatch problem in existing fixed metric approaches. 2) We propose an AML framework with learnable similarity that theoretically guarantees nontrivial pair margin under worst-case similarity notions. 3) We provide extensive experiments and analyses to demonstrate the effectiveness of AML.

2 Related Work

Fine-grained emotion classification (FEC) aims to identify nuanced categories with subtle boundaries, and has been widely studied on datasets such as GoEmotions and EmpatheticDialogues (Demszky et al., 2020; Rashkin et al., 2019). An earlier work studies automatic emotion classification from speech under substantial inter-labeler disagreement and proposes an entropy-based evaluation measure (Steidl et al., 2005). Beyond standard fine-tuning, recent methods improve FEC by introducing additional supervision to better handle confusable labels and long-tailed distributions (Gao et al., 2023; Singh et al., 2024), label-aware objectives to emphasize confusable negatives (Suresh and Ong, 2021), hierarchy- or geometry-aware label modeling (Chen et al., 2023), and multi-view embedding frameworks that leverage instance-instance and label-label interactions (Gong et al., 2025).

Contrastive learning has become a vital tool for shaping representation geometry in both self-supervised and supervised regimes (Chen et al., 2020; Khosla et al., 2020; Gunel et al., 2021; Gao et al., 2021; Xu et al., 2023). In FEC, it often relies on hard-pair mining or reweighting strategies that focus optimization on ambiguous boundaries (e.g., hard negatives or cluster-guided strong pairs) (Suresh and Ong, 2021; Gong et al., 2025; Yang et al., 2023; Yu et al., 2024). A shared assumption in these pipelines is that hardness and confusability can be faithfully expressed under a fixed similarity metric (e.g., cosine similarity). Instead, we focus on targeted hard pairs and replace the fixed metric with a learnable, adversarial discriminator guided by worst-case confusions within a metric family.

Adversarial learning provides a principled way to generate challenging training signals via min-max optimization (Goodfellow et al., 2014; Miyato et al., 2017; Hu et al., 2023; Chen et al., 2024). Different from adversarial input perturbations or adversarial contrastive variants, we adversarialize the *metric itself*: the opponent learns a similarity notion that maximizes confusion on semantically diverse (similar) intra-class (inter-class) pairs. In parallel, recent LLM-based studies explore prompting or instruction tuning for emotion-related tasks, but fine-grained affect recognition remains challenging under subtle distinctions (Liu et al., 2024b; Ren et al., 2025). Our work focuses on improving discriminative representations and is compatible with different backbone encoders.

3 Methodology

3.1 FEC as Metric Learning Problem

Let $\mathcal{U} = \{(x_i, y_i)\}_{i=1}^{N}$ be a labeled emotion classification dataset, where $x_i \in \mathcal{X}$ is an utterance and $y_i \in \mathcal{Y}$ is a fine-grained emotion label. We encode utterances with an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$:

$$z_i = f_\theta(x_i).$$

A classifier head $h_\omega : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ predicts

$$p_{\theta,\omega}(y \mid x) = \mathrm{softmax}(h_\omega(f_\theta(x))).$$

The standard supervised loss is

$$\mathcal{L}_{\mathrm{cls}}(\theta, \omega) = \mathbb{E}_{(x,y)\sim\mathcal{U}}\big[-\log p_{\theta,\omega}(y \mid x)\big].$$

Metric learning hypothesis for FEC. The key property of FEC is that *dominant errors come from confusable labels* (e.g., *joy* vs. *excitement*), for which semantic overlap makes the label boundary extremely thin. A common hypothesis is that improving the geometry of z improves classification. Namely, coherent intra-class and separated inter-class representations make the classifier more reliable. Formally, for a distance $d(\cdot,\cdot)$ on $\mathbb{R}^d$, define the intra-class radius and inter-class margin

$$R_y(d) = \sup_{i,j:\, y_i=y_j=y} d(z_i, z_j), \quad \gamma_{y,y'}(d) = \inf_{i,j:\, y_i=y,\, y_j=y'} d(z_i, z_j).$$

Metric-learning-style approaches for FEC aim at minimizing $R_y(d)$ while maximizing $\gamma_{y,y'}(d)$, especially for semantically close y, y'.

Why "metric mis-specification" is structural. In coarse-grained emotion classification, utterance pairs are trivially separable, and different similarity metrics largely agree. In FEC, the label space is crowded and similarity is subtle and context-dependent, thus the metric becomes part of the problem. However, existing hard-pair contrastive methods typically assume a *fixed metric* (dot product or cosine), which can be mis-specified precisely where FEC needs the metric most. Hard-pair mining under a fixed metric may (i) fail to expose the *worst-case* confusions and (ii) amplify early *training bias* by repeatedly sampling "hard" pairs defined by an imperfect geometry.

3.2 Theoretical Motivation

We propose to treat similarity (or distance) as a learnable opponent. Instead of committing to one metric d, we consider a *family* of learnable distance functions and optimize for worst-case separation.

Adversarial similarity family. We introduce a discriminator-style similarity function to indicate whether a sampled pair is from the same class

$$D_\phi : \mathbb{R}^d \times \mathbb{R}^d \to (0, 1),$$

where the function outputs a similarity score

$$s_{ij} = D_\phi(z_i, z_j).$$

s_{ij} is interpreted as the confidence of a pair (z_i, z_j) from the same class under a learnable metric parameterized by ϕ. Let Φ denote the discriminator hypothesis class, and define the family $\mathcal{D} = \{D_\phi : \phi \in \Phi\}$. AML aims to learn representations that remain discriminative under *any* $D \in \mathcal{D}$, especially under the most adversarial similarity function D.

Targeted hard pairs. To optimize the similarity function D, FEC requires enforcing a specific geometric principle: *aggregate semantically diverse intra-class pairs, and separate semantically similar inter-class pairs.* We formalize this by defining two targeted pair sets: same-label but semantically diverse pairs $\mathcal{P}^+$ and different-label but semantically similar pairs $\mathcal{P}^-$, where target is defined as

$$t_{ij} = \begin{cases} 1, & (i,j) \in \mathcal{P}^+, \\ 0, & (i,j) \in \mathcal{P}^-. \end{cases}$$

These pairs are the diagnostic failure modes in FEC and are vital parts in worst-case separation.

Definition 1 (Hard-pair sampling distribution)**.** *Let π_+ and π_- be sampling distributions over $\mathcal{P}^+$ and $\mathcal{P}^-$ (typically uniform). Let $\pi = w_+\pi_+ + w_-\pi_-$ be their mixture with $w_+, w_- > 0$ and $w_+ + w_- = 1$.*

Definition 2 (Adversarial metric learning loss)**.** *With encoder f_θ and discriminator D_ϕ, for (i, j), $z_i = f_\theta(x_i)$ and $s_{ij} = D_\phi(z_i, z_j)$. The adversarial metric learning (AML) loss is defined as follows*

$$\mathcal{L}_{\mathrm{aml}}(\theta, \phi) = \mathbb{E}_{(i,j)\sim\pi}\big[\ell_{\mathrm{bce}}(s_{ij}, t_{ij})\big],$$
$$\ell_{\mathrm{bce}}(s, t) = -t \log s - (1 - t)\log(1 - s).$$

The AML loss $\mathcal{L}_{\mathrm{aml}}$ is a targeted pairwise objective, which encourages high similarity on $\mathcal{P}^+$ and low similarity on $\mathcal{P}^-$. Next, we will derive that a min-max game $\min_\theta \max_{\phi\in\Phi} \mathcal{L}_{\mathrm{aml}}(\theta, \phi)$ over the AML loss can be interpreted as *worst-case metric robustness* over the similarity function family $\mathcal{D}$.

Proposition 1 (Worst-case similarity robustness on targeted hard pairs)**.** *Fix $\epsilon \geq 0$. If an encoder f with parameters θ satisfies the following condition*

$$\max_{\phi\in\Phi} \mathcal{L}_{\mathrm{aml}}(\theta, \phi) \leq \epsilon,$$

then for every *discriminator D_ϕ the following hold.*

(1) Mean-score robustness.

$$\mathbb{E}_{(i,j)\sim\pi_+}[s_{ij}] \geq \exp\Big(-\frac{\epsilon}{w_+}\Big),$$
$$\mathbb{E}_{(i,j)\sim\pi_-}[s_{ij}] \leq 1-\exp\Big(-\frac{\epsilon}{w_-}\Big).$$

(2) Tail robustness. *For any* $\delta \in (0,1)$,

$$\mathbb{P}_{(i,j)\sim\pi_+}(s_{ij}\leq\delta)\leq\min\Big\{1,\frac{\epsilon}{w_+\cdot(-\log\delta)}\Big\},$$
$$\mathbb{P}_{(i,j)\sim\pi_-}(s_{ij}\geq 1-\delta)\leq\min\Big\{1,\frac{\epsilon}{w_-\cdot(-\log\delta)}\Big\}.$$

Proposition 1 implies that no $D_\phi \in \mathcal{D}$ can substantially reduce (increase) similarity on a nontrivial mass of hard positives (hard negatives) unless the worst-case loss ϵ is correspondingly large.

Definition 3 (Logit and signed margin)**.** *For a pair* (i,j)*, define the discriminator logit as follows*

$$a_{ij} = \mathrm{logit}(s_{ij}) = \log\frac{s_{ij}}{1-s_{ij}},$$

and the signed label $\tilde{t}_{ij} = 2t_{ij} - 1 \in \{-1,+1\}$*. The signed logit margin is defined as follows*

$$m_{ij} = \tilde{t}_{ij}\, a_{ij}.$$

Proposition 2 (BCE is a margin loss in discriminator logit space)**.** *With* m_{ij} *in Definition 3, then:*

(A) Deterministic margin $\Rightarrow$ ***small loss.*** *If* $m_{ij} \geq m$ *for all* (i,j) *in a set* $\mathcal{S}$ *(with* $m \geq 0$*), then*

$$\mathbb{E}_{(i,j)\sim\mathcal{S}}\big[\ell_{\mathrm{bce}}(s_{ij},t_{ij})\big] \leq \log(1+e^{-m}).$$

(B) Small average loss $\Rightarrow$ ***few margin violations.*** *For any* $m \geq 0$ *and sampling distribution over* $\mathcal{S}$*,*

$$\mathbb{P}_{(i,j)\sim\mathcal{S}}(m_{ij}\leq m) \leq \frac{\mathbb{E}_{(i,j)\sim\mathcal{S}}[\ell_{\mathrm{bce}}(s_{ij},t_{ij})]}{\log(1+e^{-m})}.$$

Equivalently, if the average BCE loss over $\mathcal{S}$ *is at most* ε*, then at least a fraction* $1-\varepsilon/\log(1+e^{-m})$ *of hard pairs satisfy* $m_{ij} > m$*.*

Proposition 2 implies that optimizing adversarial metric learning loss with BCE is equivalent to optimizing the margin of targeted hard pairs.

Corollary 1 (Uniform margin control under adversarial similarity family)**.** *If* $\max_{\phi\in\Phi}\mathcal{L}_{\mathrm{aml}}(\theta,\phi) \leq \epsilon$*, then for every* $D_\phi \in \mathcal{D}$ *and any* $m \geq 0$*,*

$$\mathbb{P}_{(i,j)\sim\pi}(m_{ij}\leq m) \leq \frac{\epsilon}{\log(1+e^{-m})}.$$

In particular, minimizing the worst-case adversarial metric learning loss enforces a nontrivial signed logit margin on most targeted hard pairs, uniformly over the discriminator family.

Connection to AML framework. The propositions and Corollary 1 justify the min-max training objective defined in Section 3.3.3 (Proofs are provided in Appendix A). Specifically, the encoder is optimized to maintain intra-class aggregation and inter-class separation *under the worst-case learned similarity* within $\mathcal{D}$. These theoretical analyses motivate a direct design principle of the AML framework: *train representations against a worst-case, learnable similarity function on exactly the two pair types that define FEC's hardest confusions.*

3.3 Realization of AML framework

The structure of AML framework is shown in Figure 2, which is composed of a classifier, a hard-pair mining module and a discriminator.

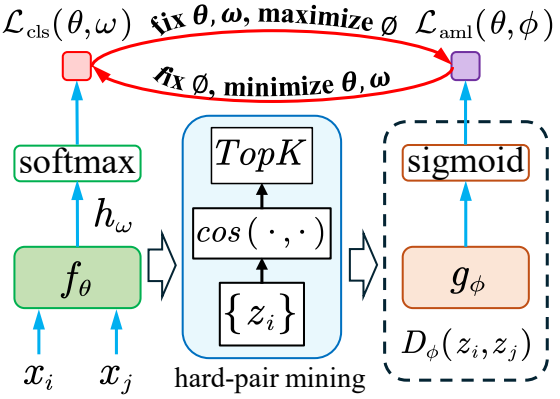

Figure 2: The structure of our AML framework.

3.3.1 Targeted Hard Pairs Construction

To construct hard pair sets $\mathcal{P}^+$ and $\mathcal{P}^-$, we need an external semantic similarity that is not trivially controlled by D_ϕ. We define an auxiliary similarity function $g : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. It provides a *stable* signal to identify "same-label but semantically diverse" and "different-label but semantically similar" pairs. Specifically, given a mini-batch $\mathcal{B}$, we compute the cosine similarity of each two utterances

$$o_{ij} = g(x_i,x_j) = \frac{f_\theta(x_i)\cdot f_\theta(x_i)}{||f_\theta(x_i)||\cdot||f_\theta(x_j)||}$$

Hard positive pairs. To construct hard positive pair set $\mathcal{P}^+$ with the same label but low similarity, for an anchor i, we choose the k_+ most semantically diverse intra-class examples as follows:

$$\mathcal{C}_i^+ = \{j \in \mathcal{B}\setminus\{i\} : y_j = y_i\},$$
$$\mathcal{H}_i^+ = \mathrm{BottomK}\Big(\{(j,o_{ij})\}_{j\in\mathcal{C}_i^+},\, k_+\Big),$$
$$\mathcal{P}^+ = \{(i,j) : i \in \mathcal{B},\, j \in \mathcal{H}_i^+\}.$$

Hard negative pairs. Similarly, to construct hard negative pair set $\mathcal{P}^-$ with different labels but high similarity, we choose the k_- most semantically similar inter-class examples as follows:

$$\mathcal{C}_i^- = \{j \in \mathcal{B} : y_j \neq y_i\},$$

$$\mathcal{H}_i^- = \mathrm{TopK}\Big(\{(j, o_{ij})\}_{j \in \mathcal{C}_i^-},\, k_-\Big),$$

$$\mathcal{P}^- = \{(i, j) : i \in \mathcal{B},\, j \in \mathcal{H}_i^-\}.$$

3.3.2 Adversarial Similarity Discriminator

We design D_ϕ to be expressive enough to represent non-trivial metric distortions, but lightweight enough to avoid overpowering the encoder.

Symmetric pair features. A standard and effective choice is an MLP over symmetric features:

$$s_{ij} = \sigma\Big(g_\phi\big([z_i; z_j; |z_i - z_j|; z_i \odot z_j]\big)\Big),$$

where $\sigma(\cdot)$ is sigmoid function. To enforce order-invariance, we optionally symmetrize:

$$D_\phi(z_i, z_j) \leftarrow \tfrac{1}{2}\big(D_\phi(z_i, z_j) + D_\phi(z_j, z_i)\big).$$

Interpretation as the learnable metric family. Unlike a fixed metric, D_ϕ can reweight dimensions and capture higher-order interactions between z_i and z_j. In our framework, this is not merely "more parameters". It is the concrete instantiation of the theory above. Namely, $\mathcal{D}$ is the similarity function family against which we require robustness.

Capacity control (robustness without collapse). Because the discriminator is adversarial (it maximizes a confusion objective), we regularize D_ϕ to remain within a plausible metric family. We use a generic regularizer $\Omega(\phi)$ (e.g., weight decay) and include it in the max-step objective. This prevents pathological solutions where D_ϕ saturates and yields vanishing gradients for the encoder.

3.3.3 Adversarial Optimization of AML

Adversarial Metric Learning Objective. After obtaining the targeted hard pair sets $\mathcal{P}^+$ and $\mathcal{P}^-$, with the well-defined encoder f_θ and adversarial similarity discriminator D_ϕ, we can sample hard pairs (i, j) from distribution π and derive the exact adversarial metric learning loss based on the expressions provided in Definition 2 as follows

$$\mathcal{L}_{\mathrm{aml}}(\theta,\phi) = \mathbb{E}_{(i,j)\sim\pi}\big[\ell_{\mathrm{bce}}(D_\phi(f_\theta(x_i), f_\theta(x_j)), t_{ij})\big].$$

Why BCE (instead of InfoNCE) fits the theory. Our theoretical statements are naturally expressed in the discriminator logit space $a_{ij} = \mathrm{logit}(s_{ij})$, where BCE directly enforces a signed margin (Proposition 2). In contrast, InfoNCE couples all negatives through a batch-softmax. While effective, InfoNCE implicitly assumes a fixed similarity to define logits. AML makes the logit itself adversarially learned and therefore aligns optimization with robustness to metric distortions.

Joint training with min-max optimization. To optimize the AML framework, we combine the standard supervised classification objective defined in Section 3.1 with adversarial metric learning loss under a min-max optimization procedure:

$$\min_{\theta,\omega} \max_{\phi} \; \mathcal{L}_{\mathrm{cls}}(\theta, \omega) + \lambda\, \mathcal{L}_{\mathrm{aml}}(\theta, \phi) - \beta\, \Omega(\phi),$$

where $\lambda > 0$ balances task learning and metric robustness, and $\beta \geq 0$ controls discriminator regularization. The $\max_\phi$ step seeks the most confusing similarity notion within $\mathcal{D}$ (subject to Ω) by minimizing the worst-case pairwise confusion loss, which yields representations that remain discriminative under any similarity in $\mathcal{D}$, while the $\min_{\theta,\omega}$ step enforces correct classification and robust pairwise structure under that adversary.

At inference time, AML uses the classifier head:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_{\theta,\omega}(y \mid x).$$

The discriminator is a training-time adversary that shapes a metric-robust representation space and does not affect test-time latency.

The Training Procedure of AML. Our AML framework is optimized via an alternating min–max procedure between the encoder/classifier and the learnable similarity discriminator. For each iteration, we first sample a mini-batch and construct the targeted hard positive/negative sets $(\mathcal{P}^+, \mathcal{P}^-)$ based on the auxiliary similarity g. We then compute the batch representations z and perform T_D discriminator steps to update ϕ (with z detached) so that the discriminator becomes maximally confusing on these targeted pairs, followed by T_E encoder/classifier steps to update (θ, ω) to minimize the overall objective under the current adversary. Algorithm 1 in Appendix B summarizes one full iteration of training AML framework.

4 Experiments

4.1 Experimental Setup

Datasets. Following prior work, we evaluate on two widely used FEC benchmarks: **Empathetic Dialogues (ED)** (Rashkin et al., 2019), which contains multi-turn conversations labeled with one of 32 emotions, and **GoEmotions (GE)** (Demszky et al., 2020), which contains Reddit comments annotated with fine-grained emotions. To ensure comparability with existing results, we adopt the same standard preprocessing protocol as previous works (Suresh and Ong, 2021; Chen et al., 2023; Gong et al., 2025): on ED, we construct instances from the first-turn situation description. On GE, we keep single-labeled instances and remove the `neutral` label. Dataset statistics are shown in Table 1.

Dataset	Train	Dev	Test	#Labels
ED	19,533	2,770	2,547	32
GE	23,485	2,956	2,984	27

Table 1: The statistics of the datasets. ED and GE denote Empathetic Dialogues and GoEmotions, respectively.

Baselines. We compare against the same baseline families used in recent SOTA work, covering: 1) Pre-trained models (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020)), 2) Coarse-grained emotion methods adapted to FEC (EmoBERTa (Kim and Vossen, 2021), SACL (Hu et al., 2023), TFD (Tu et al., 2023)), 3) FEC-specific methods (BERTCDP+MLM (Singh et al., 2024), LCL (Suresh and Ong, 2021), HypEMO (Chen et al., 2023)), and 4) LLM prompting / emotional LLMs (ChatGPT zero-/few-shot, EmoLLaMA-chat (Liu et al., 2024a)). We additionally include TACO (Gong et al., 2025) as a strong SOTA baseline. For fairness, we follow the original implementations and hyperparameter settings for each method when available, and otherwise tune on the dev set.

Evaluation metrics. Following prior works on FEC (Gong et al., 2025), we report Acc (top-1 accuracy), Weighted-F1 and Macro-F1. They directly capture overall correctness and frequency-weighted performance under the naturally long-tailed label distribution. The Weighted-F1 is computed by:

$$\text{Weighted-F1} = \sum_{e \in \mathcal{E}} \frac{n_e}{N} \cdot \frac{2P_e R_e}{P_e + R_e}.$$

Implementation details. Following the previous work, we use `roberta-base` as the encoder backbone for all encoder-based models. When implementing the classifier head h_ω, we follow the label-embedding match design in TACO. We optimize with AdamW (initial learning rate $2\text{e}{-5}$, weight decay $1\text{e}{-3}$). The batch size is 64. All results are averaged over 5 runs with random seeds. We report the mean and standard deviation of the results. For AML, we follow Algorithm 1 (Appendix B) and perform alternating updates per iteration. We tune the hyperparameters on the dev set. The best parameter settings are: $k_+ = k_- = 10$, $T_D = 1$, $T_E = 1$, and $\lambda = 1$, $\beta = 1$ for the ED dataset and $k_+ = k_- = 20$, $T_D = 1$, $T_E = 1$, $\lambda = 1$, $\beta = 1$ for the GE dataset. We run all experiments on one NVIDIA V100 GPU with 32GB memory.

4.2 Main Results

We compare all baselines with our AML framework. Table 2 reports the overall performance on ED and GE. AML achieves the best results in Acc and Weighted-F1, demonstrating that adversarially learning the similarity function is an effective way to address the dominant error source in FEC. In particular, AML substantially improves 8.17% Acc and 8.36% Weighted-F1 upon TACO on ED, indicating markedly better overall correctness and frequency-weighted performance under the long-tailed label distribution. The deficiency in Macro-F1 likely stems from minority classes, where per-class calibration is more fragile due to limited supervision. AML's large gains on Acc/Weighted-F1 demonstrate strong practical utility while remaining competitive on tail-sensitive evaluation.

4.3 Ablation Study

We decompose AML into several variants to isolate the effect of each design choice. The ablated models include: **AML w/o Adv** (standard supervised training with only the classification loss $\mathcal{L}_{\text{cls}}$), **AML w/o hard** (adversarial min-max training is applied, but targeted hard pairs are not mined, i.e., pairs are uniformly sampled), **AML-threshold** (hard pairs are selected by an absolute similarity threshold) and **AML-TopK** (using Bottom-K and Top-K to sample intra-class and inter-class hard pairs). Table 3 shows that each component contributes positively. Notably, AML-TopK performs best, supporting that the combination of targeted hard-pair construction and adversarially learned similarity yields more robust separation.

Method	**Empathetic Dialogues** (ED)			**GoEmotions** (GE)		
	Acc↑	Weighted-F1↑	Macro-F1↑	Acc↑	Weighted-F1↑	Macro-F1↑
$\text{BERT}_{\text{base}}$	$55.79_{\pm 0.35}$	$55.13_{\pm 0.51}$	$54.85_{\pm 0.56}$	$64.32_{\pm 0.33}$	$63.84_{\pm 0.32}$	$54.27_{\pm 0.89}$
$\text{RoBERTa}_{\text{base}}$	$57.67_{\pm 0.46}$	$57.13_{\pm 0.37}$	$56.87_{\pm 0.35}$	$64.79_{\pm 0.52}$	$64.34_{\pm 0.45}$	$54.91_{\pm 0.67}$
$\text{ELECTRA}_{\text{base}}$	$57.42_{\pm 0.61}$	$56.59_{\pm 0.62}$	$56.38_{\pm 0.63}$	$64.65_{\pm 0.57}$	$63.77_{\pm 0.62}$	$52.03_{\pm 1.04}$
EmoBERTa	$57.73_{\pm 0.44}$	$57.22_{\pm 0.45}$	$56.81_{\pm 0.43}$	$64.67_{\pm 0.51}$	$64.36_{\pm 0.49}$	$54.94_{\pm 0.72}$
SACL	$58.27_{\pm 0.45}$	$57.65_{\pm 0.48}$	$57.50_{\pm 0.48}$	$64.71_{\pm 0.54}$	$64.42_{\pm 0.47}$	$54.96_{\pm 0.89}$
TFD	$58.47_{\pm 0.55}$	$57.96_{\pm 0.47}$	$58.12_{\pm 0.51}$	$64.83_{\pm 0.42}$	$64.41_{\pm 0.50}$	$55.59_{\pm 1.09}$
BERTCDP+MLM	$58.51_{\pm 0.48}$	$57.94_{\pm 0.38}$	$57.74_{\pm 0.39}$	$64.98_{\pm 0.40}$	$64.56_{\pm 0.34}$	$55.84_{\pm 0.93}$
LCL	$59.52_{\pm 0.43}$	$58.72_{\pm 0.49}$	$58.38_{\pm 0.49}$	$65.22_{\pm 0.39}$	$64.55_{\pm 0.47}$	$54.48_{\pm 1.27}$
HypEMO	$58.30_{\pm 0.50}$	$57.13_{\pm 0.42}$	$56.93_{\pm 0.48}$	$64.81_{\pm 0.46}$	$64.30_{\pm 0.39}$	$53.59_{\pm 1.14}$
TACO	$60.57_{\pm 0.36}$	$59.94_{\pm 0.42}$	$\mathbf{59.82}_{\pm 0.43}$	$65.97_{\pm 0.38}$	$65.42_{\pm 0.40}$	$\mathbf{58.23}_{\pm 0.99}$
ChatGPT (zero-shot)	48.28	48.45	46.34	34.61	35.64	29.45
ChatGPT (eight-shot)	$52.21_{\pm 0.22}$	$50.69_{\pm 0.41}$	$48.68_{\pm 0.53}$	$33.90_{\pm 0.60}$	$34.88_{\pm 0.58}$	$28.87_{\pm 0.53}$
EmoLLaMA-chat-7B	27.27	28.09	26.59	24.86	24.04	22.22
EmoLLaMA-chat-13B	38.37	39.47	38.06	28.35	27.85	23.06
AML (ours)	$\mathbf{65.52}_{\pm 0.27}$	$\mathbf{64.95}_{\pm 0.36}$	$57.34_{\pm 0.38}$	$\mathbf{66.05}_{\pm 0.43}$	$\mathbf{65.67}_{\pm 0.32}$	$57.19_{\pm 0.26}$
Δ (AML vs. best baseline)	+8.17%	+8.36%		+0.12%	+0.38%	

Table 2: Main fine-grained emotion classification results on the ED and GE datasets. We copy the results of baselines from the main table of TACO. In the table, the best results in each column are marked in **bold**.

Method	Acc↑	Weighted-F1↑
AML w/o Adv	60.89	60.32
AML w/o hard	62.18	64.47
AML-threshold	65.08	64.61
AML-TopK	**65.52**	**64.95**

Table 3: Ablation study results on the ED dataset.

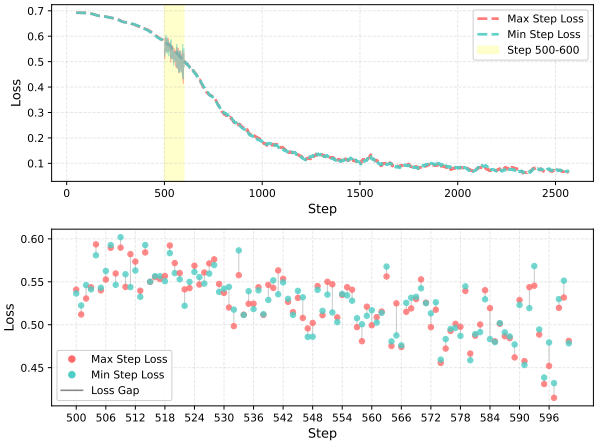

Figure 3: Loss curves and gap of max-step vs. min-step.

4.4 Adversarial Optimization Dynamics

AML is trained via alternating optimization. At each iteration, we first update the discriminator to *maximize* confusion on targeted pairs, then update the encoder/classifier to *minimize* the overall objective under the current adversary. To verify that the min-max game is being optimized stably, we track: 1) the discriminator-side loss (max-step) and 2) the encoder-side loss (min-step) throughout training. Figure 3 plots the learning curves on ED. A healthy training trajectory shows that the discriminator loss decreases during max-steps (finding harder metrics), while the encoder subsequently reduces the min-step loss, indicating that the representations adapt to withstand the adversarial similarity distortion. This empirical behavior aligns with the robustness interpretation in Section 3.

4.5 Representation Analysis

A key motivation of AML is to *aggregate dissimilar intra-class instances* while *separating similar inter-class instances*. We therefore visualize the learned utterance embeddings using t-SNE and compare AML w/o Adv with AML on the ED dataset. As shown in Figure 4, AML produces embeddings with clearer separation between confusable emotions and tighter within-class structure for many categories. This qualitative result supports our hypothesis that metric-robust learning improves the geometry where fixed-metric mining can be brittle.

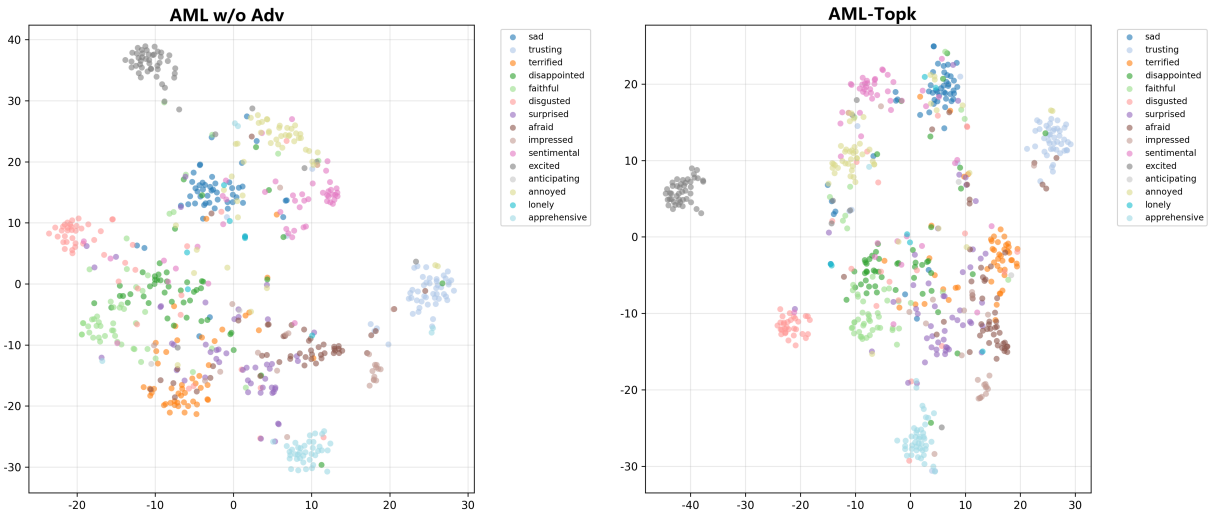

Figure 4: t-SNE visualization of ED embeddings learned by AML w/o Adv (left) and AML (right).

#	Utterance (ED situation)	AML w/o Adv		AML		Gold label confidence	
		Label	Score	Label	Score	CLS	AML
1	*"It's better to say a moment like that could truly ignite her love for the game rather than putting a bit of a damper on it."*	sad	0.35	anxious	0.41	0.37	0.41
2	*"I was teased for being a virgin when I was a 6th grader in 2005"*	trusting	0.31	impressed	0.42	0.27	0.42

Table 4: Case study on the ED dataset. "Gold label confidence" reports the probability for the true emotion.

4.6 Hyperparameter Analysis on ED

Effect of Top-K in hard-pair mining. Top-K (and Bottom-K) controls how aggressively we mine targeted pairs. If K is too small, mined pairs may be insufficiently informative; if too large, the pair sets may include easy or noisy pairs. Figure 5 (upper) shows performance as K varies.

Effect of λ (adversarial strength). λ balances the supervised objective and adversarial metric robustness. As λ increases, AML emphasizes robustness more strongly; overly large λ may hinder fitting the label decision boundaries. Figure 5 (lower) summarizes the sensitivity trend.

4.7 Case Study

To show how AML improves decision on confusable emotions, we provide examples where AML succeeds while AML w/o Adv fails. Table 4 lists the utterance, top prediction and confidence, and the confidence assigned to the gold label. These cases suggest that AML tends to allocate higher probability mass to the correct label even when the input is ambiguous, consistent with the goal of robust separation under adversarial similarity.

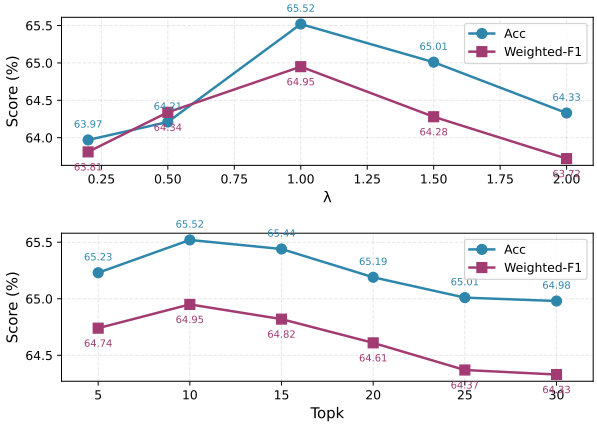

Figure 5: Hyperparameter sensitivity on the ED dataset.

5 Conclusion

In this paper, we proposed AML, an adversarial metric learning framework for fine-grained emotion classification that replaces fixed similarity metrics with a learnable metric family and trains representations to remain discriminative under worst-case similarity distortions. Extensive experiments demonstrate that AML achieves strong and consistent improvements over baselines, validating the effectiveness of metric-robust representations.

Limitations

First, the effectiveness of AML depends on hard-pair construction and related hyperparameters (e.g., K, λ, T_D and T_E). Thus, suboptimal settings or noisy mined pairs may weaken gains, especially for extremely low-resource emotion classes.

Second, our experiments follow the standard GoEmotions preprocessing used in prior fine-grained text emotion classification work, i.e., keeping only single-labeled instances and removing the neutral label. While this improves comparability, it also limits the scope of the current study: single-label supervision cannot explicitly model naturally overlapping emotions or annotator uncertainty, and removing neutral excludes an important boundary class that often absorbs mild affect or ambiguity. A natural extension is to replace binary pair targets with soft targets derived from label co-occurrence or annotator distributions, and to make hard-pair mining overlap-aware in multi-label settings.

Third, our experiments focus on English FEC benchmarks and single-utterance settings. It remains unclear how well the approach transfers to multi-turn scenarios, other languages, or larger-scale datasets without additional adaptation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. U23B2056, No. U2433212, No. 62306026), in part by the Fundamental Research Funds for the Central Universities, and in part by the State Key Laboratory of Complex & Critical Software Environment.

References

Chih-Yao Chen, Tun-Min Hung, Yi-Li Hsu, and Lun-Wei Ku. 2023. Label-aware hyperbolic embeddings for fine-grained emotion classification. In *ACL*, pages 10947–10958.

Junfan Chen, Richong Zhang, Junchi Chen, and Chunming Hu. 2024. Open-set semi-supervised text classification via adversarial disagreement maximization. In *ACL*, pages 2170–2180.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607.

Xuping Chen and Wuzhen Shi. 2025. Dynamic interactive bimodal hypergraph networks for emotion recognition in conversations. In *AAAI*, pages 1256–1264.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *ACL*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2023. The benefits of label-description training for zero-shot text classification. In *EMNLP*, pages 13823–13844.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910.

Junqing Gong, Binhan Yang, and Wei Shen. 2025. A triple-view framework for fine-grained emotion classification with clustering-guided contrastive learning. In *ACL*, pages 4970–4984.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*, pages 2672–2680.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *ICLR*.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *ACL*, pages 10835–10852.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *NeurIPS*.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *CoRR*, abs/2108.12009.

Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021. Towards an online empathetic chatbot with emotion causes. In *SIGIR*, pages 2041–2045.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024a. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *SIGKDD*, pages 5487–5496.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024b. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *SIGKDD*, pages 5487–5496.

Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. 2021. Affect2mm: Affective analysis of multimedia content using emotion causality. In *CVPR*, pages 5661–5671.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *ICLR*.

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

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, pages 5370–5381.

Zhaochun Ren, Zhou Yang, Chenglong Ye, Haizhou Sun, Chao Chen, Xiaofei Zhu, and Xiangwen Liao. 2025. Fine-grained emotion recognition via in-context learning. In *CIKM*, pages 2503–2513.

Zhiyu Shen, Yunhe Pang, Yanghui Rao, and Jianxing Yu. 2025. Coe: A clue of emotion framework for emotion recognition in conversations. In *ACL*, pages 23548–23563.

Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. 2024. Text-based fine-grained emotion prediction. *IEEE TAC*, 15(2):405–416.

Stefan Steidl, Michael Levit, Anton Batliner, Elmar Nöth, and Heinrich Niemann. 2005. "of all things the measure is man" : Automatic classification of emotions and inter-labeler consistency. In *ICASSP*, pages 317–320. IEEE.

Varsha Suresh and Desmond C. Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *EMNLP*, pages 4381–4394.

Geng Tu, Ran Jing, Bin Liang, Min Yang, Kam-Fai Wong, and Ruifeng Xu. 2023. A training-free debiasing framework with counterfactual reasoning for conversational emotion detection. In *EMNLP*, pages 15639–15650.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023. Simcse++: Improving contrastive learning for sentence embeddings from two perspectives. In *EMNLP*, pages 12028–12040.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE TAC*, 14(4):3269–3280.

Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2024. Emotion-anchored contrastive learning framework for emotion recognition in conversation. In *Findings of ACL: NAACL*, Findings of ACL, pages 4521–4534. Association for Computational Linguistics.

Jiahao Yuan, Zixiang Di, Zhiqing Cui, Guisong Yang, and Usman Naseem. 2025. Reflectdiffu: Reflect between emotion-intent contagion and mimicry for empathetic response generation via a rl-diffusion framework. In *ACL*, pages 25435–25449.

A Proofs of Theory Statements

A.1 Notation and Preliminaries

Let $\mathcal{P}^+$ and $\mathcal{P}^-$ denote the targeted hard positive/negative pair sets defined in Section 3.3.1. For any pair (i, j), let

$$z_i = f_\theta(x_i), \qquad z_j = f_\theta(x_j),$$

and the discriminator similarity score

$$s_{ij} = D_\phi(z_i, z_j) \in (0, 1).$$

We define the Bernoulli cross-entropy loss

$$\ell_{\text{bce}}(s, t) = -t \log s - (1 - t) \log(1 - s),$$

where $t \in \{0, 1\}$ is the target indicator.

For convenience, we also use the $\{\pm 1\}$ label encoding

$$\tilde{t} = 2t - 1 \in \{-1, +1\}.$$

Finally, define the logit

$$a = \text{logit}(s) = \log \frac{s}{1 - s} \quad \text{or} \quad s = \sigma(a) = \frac{1}{1 + e^{-a}}.$$

Lemma 1 (BCE as logistic loss in logit space)**.** *For $s = \sigma(a)$ and $\tilde{t} = 2t - 1$, the BCE satisfies*

$$\ell_{\text{bce}}(s, t) = \log\big(1 + \exp(-\tilde{t}a)\big).$$

Proof. We consider two cases.

Case 1: $t = 1$ (thus $\tilde{t} = +1$). Then $\ell_{\text{bce}}(s, 1) = -\log s$. Since $s = \sigma(a) = \frac{1}{1+e^{-a}}$, we have

$$\begin{aligned} -\log s &= -\log\left(\frac{1}{1 + e^{-a}}\right) = \log(1 + e^{-a}) \\ &= \log\big(1 + \exp(-\tilde{t}a)\big). \end{aligned}$$

Case 2: $t = 0$ (thus $\tilde{t} = -1$). Then $\ell_{\text{bce}}(s, 0) = -\log(1 - s)$. Using $1 - s = 1 - \sigma(a) = \sigma(-a) = \frac{1}{1+e^{a}}$, we get

$$\begin{aligned} -\log(1 - s) &= -\log\left(\frac{1}{1 + e^{a}}\right) = \log(1 + e^{a}) \\ &= \log\big(1 + \exp(-\tilde{t}a)\big), \end{aligned}$$

because $-\tilde{t}a = a$ when $\tilde{t} = -1$. Combining the two cases proves the lemma. □

We will repeatedly use two standard inequalities.

Lemma 2 (Jensen lower bound for $-\log$)**.** *Let X be a random variable supported on $(0, 1)$. Then*

$$\mathbb{E}[-\log X] \geq -\log \mathbb{E}[X].$$

Proof. The function $g(x) = -\log x$ is convex on $(0, \infty)$, hence by Jensen's inequality

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \Rightarrow \mathbb{E}[-\log X] \geq -\log \mathbb{E}[X].$$

Lemma 2 is proved. □

Lemma 3 (Markov inequality)**.** *Let Y be a non-negative random variable and let $c > 0$. Then*

$$\mathbb{P}(Y \geq c) \leq \frac{\mathbb{E}[Y]}{c}.$$

A.2 Proposition 2 and Proof

We restate Proposition 2 in a precise form that we will prove.

Proposition 3 (Loss–margin relation in discriminator logit space)**.** *Let $a_{ij} = \text{logit}(D_\phi(z_i, z_j))$ and $\tilde{t}_{ij} = 2t_{ij} - 1$. Define the signed logit margin*

$$m_{ij} = \tilde{t}_{ij}\, a_{ij}.$$

Then the following statements hold.

(A) Deterministic margin implies an upper bound on BCE. *If $m_{ij} \geq m$ for all pairs in a set $\mathcal{S}$ (for some $m \geq 0$), then for all $(i, j) \in \mathcal{S}$,*

$$\ell_{\text{bce}}(D_\phi(z_i, z_j), t_{ij}) \leq \log(1 + e^{-m}),$$

and thus

$$\mathbb{E}_{(i,j)\sim\mathcal{S}}\big[\ell_{\text{bce}}(D_\phi(z_i, z_j), t_{ij})\big] \leq \log(1 + e^{-m}).$$

(B) Small average BCE implies that most pairs have nontrivial margin. *Fix any $m \geq 0$ and any pair distribution over $\mathcal{S}$. Then*

$$\mathbb{P}_{(i,j)\sim\mathcal{S}}(m_{ij} \leq m) \leq \frac{\mathbb{E}_{(i,j)\sim\mathcal{S}}\big[\ell_{\text{bce}}(D_\phi(z_i,z_j),t_{ij})\big]}{\log(1 + e^{-m})}.$$

Equivalently, if the average BCE is at most ϵ, then at least a $1 - \frac{\epsilon}{\log(1+e^{-m})}$ fraction of pairs satisfy $m_{ij} > m$.

Proof. By Lemma 1, for each pair we have

$$\begin{aligned} \ell_{\text{bce}}(D_\phi(z_i, z_j), t_{ij}) &= \log\big(1 + \exp(-\tilde{t}_{ij}a_{ij})\big) \\ &= \log\big(1 + \exp(-m_{ij})\big). \end{aligned}$$

Proof of (A). If $m_{ij} \geq m$, then $-m_{ij} \leq -m$, hence $\exp(-m_{ij}) \leq \exp(-m)$, and therefore

$$\begin{aligned} \log\big(1 + \exp(-m_{ij})\big) &\leq \log(1 + \exp(-m)) \\ &= \log(1 + e^{-m}). \end{aligned}$$

Taking expectation over $(i, j) \sim \mathcal{S}$ preserves the inequality.

Proof of (B). Consider the event $\mathcal{E} = \{(i,j) : m_{ij} \le m\}$. On $\mathcal{E}$ we have $-m_{ij} \ge -m$, hence $\exp(-m_{ij}) \ge \exp(-m)$, and thus

$$\ell_{\mathrm{bce}}(D_\phi(z_i, z_j), t_{ij}) = \log(1 + \exp(-m_{ij})) \ge \log(1 + \exp(-m)) = \log(1 + e^{-m}).$$

Therefore,

$$\mathbb{E}\big[\ell_{\mathrm{bce}}(D_\phi(z_i, z_j), t_{ij})\big] \ge \mathbb{E}\big[\ell_{\mathrm{bce}}(\cdot) \cdot \mathbb{I}[\mathcal{E}]\big] \ge \log(1 + e^{-m}) \cdot \mathbb{P}(\mathcal{E}).$$

Rearranging gives the claimed bound. □

A.3 Proposition 1 and Proof

As written in the main text, Proposition 1 is an informal robustness claim. Below we provide a *complete* and statement that captures the exact robustness guarantee implied by the min-max objective, and then prove it.

A.3.1 Robust objective

Define the targeted adversarial metric learning loss with explicit mixture weights. Let $w_+, w_- > 0$ with $w_+ + w_- = 1$. For any (θ, ϕ), define

$$\mathcal{L}_{\mathrm{aml}}(\theta, \phi) = w_+ \cdot \mathbb{E}_{(i,j)\sim\mathcal{P}^+}\big[-\log s_{ij}\big] + w_- \cdot \mathbb{E}_{(i,j)\sim\mathcal{P}^-}\big[-\log(1 - s_{ij})\big],$$

where $s_{ij} = D_\phi(f_\theta(x_i), f_\theta(x_j))$.

Consider the robust optimization objective (ignoring the supervised term, which is orthogonal to this proof):

$$V^\star = \min_\theta \max_{\phi\in\Phi} \mathcal{L}_{\mathrm{aml}}(\theta, \phi).$$

We will state robustness properties for any θ that achieves small worst-case loss.

A.3.2 Statement

Proposition 4 (Worst-case metric robustness for targeted hard pairs)**.** *Fix any $\epsilon \ge 0$. Suppose θ satisfies*

$$\max_{\phi\in\Phi} \mathcal{L}_{\mathrm{aml}}(\theta, \phi) \le \epsilon.$$

Then for every *discriminator $D_\phi \in \mathcal{D}$ the following hold:*

(1) Mean-score robustness on hard positives.

$$\mathbb{E}_{(i,j)\sim\mathcal{P}^+}\big[s_{ij}\big] \ge \exp\Big(-\frac{\epsilon}{w_+}\Big).$$

(2) Mean-score robustness on hard negatives.

$$\mathbb{E}_{(i,j)\sim\mathcal{P}^-}\big[s_{ij}\big] \le 1 - \exp\Big(-\frac{\epsilon}{w_-}\Big).$$

(3) Tail robustness: few hard positives can be forced to low similarity. *For any $\delta \in (0,1)$,*

$$\mathbb{P}_{(i,j)\sim\mathcal{P}^+}\big(s_{ij} \le \delta\big) \le \min\Big\{1, \frac{\epsilon}{w_+ \cdot (-\log\delta)}\Big\}.$$

(4) Tail robustness: few hard negatives can be forced to high similarity. *For any $\delta \in (0,1)$,*

$$\mathbb{P}_{(i,j)\sim\mathcal{P}^-}\big(s_{ij} \ge 1-\delta\big) \le \min\Big\{1, \frac{\epsilon}{w_- \cdot (-\log\delta)}\Big\}.$$

Consequently, no function in $\mathcal{D}$ can simultaneously push a nontrivial semantically-diverse intra-class pairs to low similarity and pull a nontrivial fraction of semantically-similar inter-class pairs to high similarity, unless ϵ is correspondingly large.

Proof. Fix any $\phi \in \Phi$. Let

$$L_+(\theta, \phi) = \mathbb{E}_{(i,j)\sim\mathcal{P}^+}\big[-\log s_{ij}\big],$$
$$L_-(\theta, \phi) = \mathbb{E}_{(i,j)\sim\mathcal{P}^-}\big[-\log(1 - s_{ij})\big].$$

By definition,

$$\mathcal{L}_{\mathrm{aml}}(\theta, \phi) = w_+ L_+(\theta, \phi) + w_- L_-(\theta, \phi).$$

The assumption $\max_\phi \mathcal{L}_{\mathrm{aml}}(\theta, \phi) \le \epsilon$ implies that for this fixed ϕ,

$$w_+ L_+(\theta, \phi) + w_- L_-(\theta, \phi) \le \epsilon.$$

Since $L_+(\theta, \phi) \ge 0$ and $L_-(\theta, \phi) \ge 0$, we immediately obtain the separate bounds

$$L_+(\theta, \phi) \le \frac{\epsilon}{w_+}, \qquad L_-(\theta, \phi) \le \frac{\epsilon}{w_-}.$$

Proof of (1). Let us apply Lemma 2 to $X = s_{ij} \in (0,1)$ under $(i,j) \sim \mathcal{P}^+$:

$$L_+(\theta, \phi) = \mathbb{E}[-\log s_{ij}] \ge -\log \mathbb{E}[s_{ij}].$$

Therefore,

$$-\log \mathbb{E}_{\mathcal{P}^+}[s_{ij}] \le L_+(\theta, \phi) \le \frac{\epsilon}{w_+},$$

which implies

$$\mathbb{E}_{(i,j)\sim\mathcal{P}^+}[s_{ij}] \ge \exp\Big(-\frac{\epsilon}{w_+}\Big).$$

Proof of (2). Apply Lemma 2 to $X = 1 - s_{ij} \in (0,1)$ under $(i,j) \sim \mathcal{P}^-$:

$$L_-(\theta, \phi) = \mathbb{E}[-\log(1 - s_{ij})] \ge -\log \mathbb{E}[1 - s_{ij}].$$

Thus,

$$-\log \mathbb{E}_{\mathcal{P}^-}[1 - s_{ij}] \le L_-(\theta, \phi) \le \frac{\epsilon}{w_-},$$

so

$$\mathbb{E}_{\mathcal{P}^-}[1 - s_{ij}] \ge \exp\Big(-\frac{\epsilon}{w_-}\Big).$$

Finally, using $\mathbb{E}[s_{ij}] = 1 - \mathbb{E}[1 - s_{ij}]$ yields

$$\mathbb{E}_{\mathcal{P}^-}[s_{ij}] \le 1 - \exp\Big(-\frac{\epsilon}{w_-}\Big).$$

Proof of (3). Consider the nonnegative random variable $Y = -\log s_{ij}$ under $(i,j) \sim \mathcal{P}^+$. For any $\delta \in (0,1)$, the event $\{s_{ij} \le \delta\}$ implies $-\log s_{ij} \ge -\log \delta$. Therefore,

$$\mathbb{P}(s_{ij} \le \delta) = \mathbb{P}(-\log s_{ij} \ge -\log \delta) \le \frac{\mathbb{E}[-\log s_{ij}]}{-\log \delta} = \frac{L_+(\theta,\phi)}{-\log \delta} \le \frac{\epsilon}{w_+ \cdot (-\log \delta)},$$

where we used Markov's inequality (Lemma 3) and $L_+(\theta,\phi) \le \epsilon / w_+$.

Proof of (4). Similarly, define $Y = -\log(1 - s_{ij})$ under $(i,j) \sim \mathcal{P}^-$. Event $\{s_{ij} \ge 1-\delta\}$ implies $1 - s_{ij} \le \delta$, hence $-\log(1-s_{ij}) \ge -\log \delta$. Thus,

$$\mathbb{P}(s_{ij} \ge 1-\delta) = \mathbb{P}(-\log(1-s_{ij}) \ge -\log \delta) \le \frac{\mathbb{E}[-\log(1-s_{ij})]}{-\log \delta} = \frac{L_-(\theta,\phi)}{-\log \delta} \le \frac{\epsilon}{w_- \cdot (-\log \delta)}.$$

Consequence (simultaneous robustness). All bounds above hold for an *arbitrary* fixed ϕ. Since the assumption is $\max_\phi \mathcal{L}_{\text{aml}}(\theta,\phi) \le \epsilon$, they hold *uniformly for all* $\phi \in \Phi$. Thus, within the metric family $\mathcal{D}$, no adversary can significantly degrade hard-positive similarity or inflate hard-negative similarity without forcing ϵ to be large. □

A.3.3 Connection to the min–max optimum

If $\theta^\star \in \arg\min_\theta \max_\phi \mathcal{L}_{\text{aml}}(\theta,\phi)$ achieves value $V^\star$, then Proposition 4 holds with $\epsilon = V^\star$. More generally, if θ is η-suboptimal, i.e.,

$$\max_\phi \mathcal{L}_{\text{aml}}(\theta,\phi) \le V^\star + \eta,$$

the same bounds hold with $\epsilon = V^\star + \eta$.

Remark (why this is the right robustness notion for our goal). The bounds in Proposition 4 are stated directly in terms of D_ϕ's outputs on the two targeted hard pair sets. This matches our operational objective in FEC. Namely, regardless of which similarity notion the discriminator tries to impose, the representation should (i) keep hard intra-class pairs highly similar and (ii) keep hard inter-class pairs dissimilar. The proof shows this is exactly what minimizing the *worst-case* BCE enforces, quantitatively, via both mean-score and tail guarantees.

B The algorithm of training AML.

As shown in Algorithm 1, for each iteration, a mini-batch is sampled and used to construct the targeted hard positive/negative sets $(\mathcal{P}^+, \mathcal{P}^-)$ based on the auxiliary similarity g. Then we perform T_D discriminator steps to update ϕ (with z detached) and T_E encoder/classifier steps to update (θ,ω) to minimize the overall objective under the current adversary.

Algorithm 1 AML training (one iteration)

```
Sample a mini-batch B from D.
Mine P+, P− using similarity g on B.
Compute z_i = f_θ(x_i) for i ∈ B.
for t = 1 to T_D do
    Update φ by gradient descent on L_D(φ; θ), detaching z.
end for
for t = 1 to T_E do
    Update (θ, ω) by gradient descent on loss L_E(θ, ω; φ).
end for
```

C Discussion: Relation to Hard-Pair Contrastive Learning

Existing clustering-guided contrastive approaches can be viewed as enforcing the FEC principle under a *fixed* similarity metric: clustering (or cosine-based mining) identifies hard positives/negatives and a contrastive loss reshapes the space accordingly. AML keeps the same high-level principle but changes the core mechanism:

From fixed metric to metric family. We do not assume cosine similarity is the correct notion of confusability. Instead, we learn a family $\mathcal{D}$ and optimize for the worst-case metric in that family.

From hard mining to adversarial stress-testing. The discriminator actively attempts to *collapse* hard inter-class negatives and *split* hard intra-class positives. The encoder must preserve discrimination under this stress test, yielding stronger robustness where FEC is most fragile.

From heuristic geometry to theory-guided robustness. The min-max objective is directly motivated by worst-case metric robustness (Proposition 1) and has a margin interpretation (Proposition 2), providing a principled explanation for why the learned space improves confusable emotion separation.