

EmoDap: Dual-Agent Prompt Retrieval Framework for Zero-Shot Fine-Grained Emotion Recognition

Kun Peng^{1,2}, Cong Cao^{1*}, Hao Peng³, Xiaoyan Yu⁴, Zihao Wang⁴,
Lei Jiang¹, Yu Liu^{1,2}, Yanbing Liu^{1,2}, Philip S. Yu⁵

¹Institute of Information Engineering, Chinese Academy of Sciences, China
²School of Cyber Security, University of Chinese Academy of Sciences, China
³Beihang University, China ⁴Nanyang Technological University, Singapore
⁵University of Illinois at Chicago, USA
{pengkun, caocong}@iie.ac.cn

Abstract

Zero-shot fine-grained emotion recognition (FER) is pivotal for empathetic AI but remains constrained by two inherent limitations in retrieval-augmented in-context learning (ICL): 1) subtle emotional distinctions inject semantic noise that blurs decision boundaries, and 2) naive similarity-based retrieval often surfaces counterproductive or antonymous demonstrations. To address these, we propose **EmoDap**, a **Dual-Agent Prompt Retrieval** framework for zero-shot FER. EmoDap first constructs a high-quality demonstration pool via self-consistency pseudo-labeling and derives explicit emotion prototypes through contrastive learning. Subsequently, a demonstration retriever and an emotion predictor are jointly optimized via Group Relative Policy Optimization (GRPO): the retriever learns to select examples that maximize the predictor’s accuracy, while the predictor leverages prototype guidance to dynamically filter candidate emotions. Extensive experiments across four benchmarks demonstrate that EmoDap consistently outperforms state-of-the-art baselines, establishing a robust paradigm for zero-shot FER driven by open-source LLM backbones.

1 Introduction

Fine-grained emotion recognition (FER) is foundational to empathetic human-computer interaction (Picard, 2000; Poria et al., 2017; Liu et al., 2024c). Unlike coarse-grained recognition that targets a small set of broad, semantically distinct categories (e.g., Plutchik’s eight basic emotions (Plutchik, 2001)), FER involves a dense label space of 27-40 nuanced categories with high semantic overlap and context-dependent intensity (Rashkin et al., 2019; Demszky et al., 2020). This structural complexity amplifies decision-boundary ambiguity, making zero-shot generalization—where



Figure 1: Traditional methods struggle with emotional interference and counterproductive retrieval, while our EmoDap addresses these challenges.

models must classify unseen emotions without task-specific supervision—substantially more challenging. While retrieval-augmented in-context learning (ICL) (Wan et al., 2023; Yao et al., 2023; Liu et al., 2024b; Wang et al., 2025b) has proven effective for coarse-grained tasks by providing LLMs with relevant demonstrations (Wang et al., 2023; Tak and Gratch, 2023; Zhao et al., 2024; Tak and Gratch, 2024), it struggles to scale to fine-grained spaces due to inherent retrieval misalignments.

As illustrated in Figure 1, existing ICL retrieval strategies suffer from two critical limitations in FER: First, **Emotional Interference**: The dense, overlapping semantics of fine-grained emotions cause similarity-based retrieval to surface demonstrations of adjacent but distinct categories (e.g., *ashamed* vs. *guilty*, *annoyed* vs. *furious*). This injects semantic noise into prompts, obscuring discriminative features and confusing LLM decision boundaries. Second, **Counterproductive Retrieval**: Semantic similarity is fundamentally misaligned with emotional relevance. Naive matching frequently retrieves antonymous or contextually mismatched examples, actively misleading the

*Corresponding author.

LLM toward incorrect predictions.

To efficiently overcome these challenges and improve the effectiveness of LLMs for zero-shot FER, we propose the **Dual-Agent Prompt Retrieval (EmoDap)** framework.

Specifically, before training, the absence of golden demonstrations necessitates constructing a high-quality and diverse demonstration pool (demo-pool) in the pre-computation phase to mitigate interference caused by low-quality retrievals. Inspired by self-consistency annotating (Xie et al., 2024), we assign reliable pseudo-labels to unlabeled samples by leveraging the consistency of predictions across multiple LLM inference iterations. We then perform clustering and collect multiple category centroids, ensuring they capture the core characteristics and diversity of each emotion. Furthermore, to address the challenge of emotion interference, we learn distinct prototype representations for all given emotions based on contrastive learning. These prototypes provide explicit quantitative guidance for originally black-box LLM reasoning, serving a dual role in later stages: (i) compressing the candidate label space, and (ii) acting as the geometric basis of the reward signal during joint training. Both the demo-pool and prototypes are pre-constructed once and remain frozen during training and inference.

Subsequently, we design and train a novel prompt retrieval framework comprising two collaborative agents: a Demonstration Retriever ($\text{Agent}_{\text{demo}}$) that retrieves the most suitable examples from the demo-pool for the input sentence, and an Emotion Predictor ($\text{Agent}_{\text{emo}}$) that constructs prompts using these examples and incorporates prototype guidance to make accurate emotion predictions. These two agents are specifically designed to address the two mentioned challenges. For **Emotional Interference (Challenge 1)**, $\text{Agent}_{\text{emo}}$ first performs coarse filtering based on similarity matching with emotion prototypes. This approach provides dual benefits: the quantitatively learned emotion prototypes offer reliable knowledge enhancement to $\text{Agent}_{\text{emo}}$, while the reduced label space after filtering enables more precise probability predictions. For **Counterproductive Retrieval (Challenge 2)**, we jointly optimize $\text{Agent}_{\text{demo}}$ and $\text{Agent}_{\text{emo}}$ via Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which samples a group of retrieval trajectories, evaluates each with a fine-grained, prototype-aware reward, and updates both agents using relative group advantages. This en-

ables the retriever to dynamically learn demonstration selection policies that directly maximize the predictor’s calibrated confidence, effectively aligning retrieval with downstream success.

Our key contributions are threefold:

- We introduce a prototype-guided filtering mechanism that quantitatively compresses the candidate emotion space, effectively neutralizing semantic interference from closely related categories.
- We propose a GRPO-optimized dual-agent framework with a dense prototype-aware reward to jointly optimize retrieval and prediction, replacing brittle similarity heuristics with learnable, preference-driven policies.
- Extensive experiments on four benchmark datasets demonstrate that EmoDap significantly outperforms state-of-the-art baselines, establishing a new standard for zero-shot FER.

2 Related Work

2.1 LLMs for Emotion Recognition

While LLMs exhibit strong capabilities in coarse-grained emotion recognition (Tak and Gratch, 2024; Zhao et al., 2023; Liu et al., 2024d), their direct application to zero-shot fine-grained FER suffers from significant performance degradation (Zhang et al., 2024) and systematic over-prediction biases (Niu and Provost, 2025). Recent efforts to mitigate these limitations via knowledge distillation (Niu and Provost (2025)) or synthetic data augmentation (Shvets (2025)) still underperform on complex emotion taxonomies due to their reliance on supervised fine-tuning. EICL (Ren et al., 2025) replaces semantic-similarity retrieval with emotion-similarity matching, yet its dependence on an external auxiliary emotion model severely restricts zero-shot generalization.

2.2 Demonstration Retrieval in ICL

Retrieval-augmented ICL has largely superseded manual prompt engineering by automatically constructing informative demonstrations from corpora (Wan et al., 2023; Yao et al., 2023; Liu et al., 2024b; Wang et al., 2025b). Conventional retrievers prioritize surface-level features such as semantic similarity (Rubin et al., 2022; Li and Qiu, 2023; Li et al., 2023), structural patterns (Levy et al., 2023; Cai et al., 2023), or task-specific heuristics (Gonen et al., 2023; Wang et al., 2025a). However, in fine-grained FER, semantic proximity frequently misaligns with emotional relevance, causing retrievers

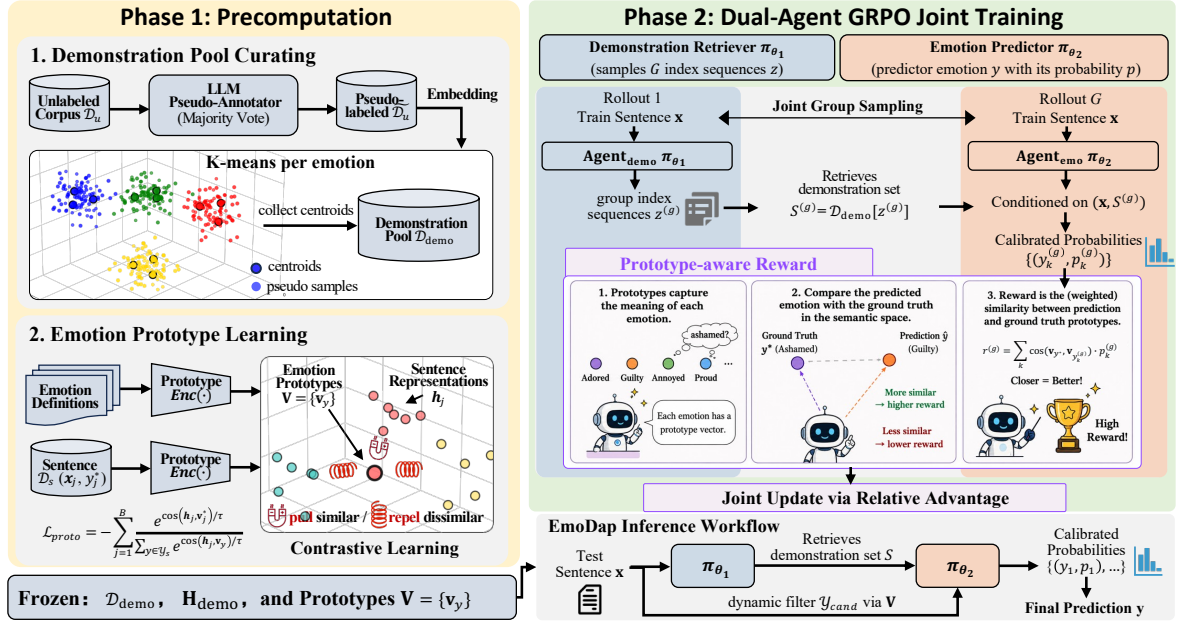


Figure 2: Overview of our EmoDap framework.

to surface antonymous or contextually confounding examples that actively mislead LLM predictions (Luo et al., 2024). This misalignment is exacerbated in zero-shot settings, where the absence of task labels prevents direct relevance filtering. Consequently, effective demonstration selection requires moving beyond static similarity metrics toward preference-driven policies that explicitly optimize retrieval for downstream prediction success.

3 Methodology

3.1 Task Formulation

We formulate zero-shot FER as a cross-label-space classification task. Let $\mathcal{D}_s$ and $\mathcal{D}_t$ denote the source and target datasets with disjoint emotion sets $\mathcal{Y}_s = \{y_1^s, \dots, y_{|\mathcal{Y}_s|}^s\}$ and $\mathcal{Y}_t = \{y_1^t, \dots, y_{|\mathcal{Y}_t|}^t\}$ ($\mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset$). We follow the transductive zero-shot setting, where unlabeled target-domain samples are accessible during training. $\mathcal{Y}_s$ is restricted to Plutchik’s eight basic emotions¹ (Plutchik, 2001) that are easier to collect, while $\mathcal{Y}_t$ comprises complex, unseen categories. Given a test sentence $x \in \mathcal{D}_t$, the goal is to predict $y \in \mathcal{Y}_t$ using an LLM-based predictor:

$$y = \mathcal{P}(T_{emo}(x, \mathcal{Y}_t, S)), \quad (1)$$

where S denotes the retrieved in-context demonstrations and T_{emo} is the prompt template. In the

¹namely, Joy, Sadness, Anger, Fear, Trust, Disgust, Surprise, and Anticipation

zero-shot setting, golden demonstrations are unavailable; thus, S must be constructed from an unlabeled corpus $\mathcal{D}_u$ for prediction.

3.2 Precomputation

The overview of our EmoDap² framework is shown in Figure 2. Prior to dual-agent training, we execute two one-time precomputation steps to establish a high-quality demonstration pool and explicit emotion prototypes.

3.2.1 Demonstration Pool Curating

To mitigate retrieval of low-quality examples, we curate a diverse demonstration pool $\mathcal{D}_{demo}$ from $\mathcal{D}_u$ via self-consistency pseudo-annotation and clustering. Following Xie et al. (2024), each sample in $\mathcal{D}_u$ undergoes M independent LLM inference runs; the majority-voted label serves as the pseudo-label, yielding $\tilde{\mathcal{D}}_u$. Then, we encode pseudo-labeled demonstrations in $\tilde{\mathcal{D}}_u$ via a frozen sentence encoder (using the [CLS] token) and apply K-means per emotion. This process naturally mitigates annotation noise: centroids anchor to prototypical instances, while noisy or atypical samples are relegated to the periphery. Because clustering operates on general semantic embeddings rather than emotion-specific features, the resulting K centroids inherently preserve contextual diversity, ensuring the pool spans a broad spectrum of intra-class variations. For each centroid we take its nearest sample

²Available at <https://github.com/KunPunCN/EmoDap>

$s_{y,k}$ as the representative, and aggregate them into the final pool:

$$\mathcal{D}_{\text{demo}} = \{s_{y,k} \mid y \in \mathcal{Y}_s \cup \mathcal{Y}_t, k \in [1, K]\}. \quad (2)$$

3.2.2 Emotion Prototype Learning

To address Challenge 1 (emotional interference), we learn explicit prototype representations that anchor each emotion in a continuous space. We employ a shared prototype encoder $Enc(\cdot)$ to map emotion definitions (from the Oxford English Dictionary) to vectors $\mathbf{v}_y = Enc(y)$ ($\mathbf{v}_y \in \mathbb{R}^d$), using the [CLS] token of the last hidden layer. Similarly, given a sentence $\mathbf{x}_j \in \mathcal{D}_s$ with its ground-truth label y_j^* , we obtain its representation $\mathbf{h}_j = Enc(\mathbf{x}_j)$ ($\mathbf{h}_j \in \mathbb{R}^d$). Then, inspired by infoNCE (Oord et al., 2018), we define a contrastive loss:

$$\mathcal{L}_{\text{proto}} = -\sum_{j=1}^B \log \frac{e^{\cos(\mathbf{h}_j, \mathbf{v}_{y_j^*})/\tau}}{\sum_{y \in \mathcal{Y}_s} e^{\cos(\mathbf{h}_j, \mathbf{v}_y)/\tau}}, \quad (3)$$

where $\cos(\cdot)$ denotes cosine similarity, τ is temperature, and B is batch size. By minimizing $\mathcal{L}_{\text{proto}}$, we pull sentences closer to their ground-truth emotion prototypes while repelling representations of confusable categories, thereby yielding highly discriminative prototypes that serve as semantic anchors for each emotion.

After training the encoder, for unseen emotions, prototypes are derived solely from textual definitions without any labeled instances, preserving the strict zero-shot constraint on $\mathcal{Y}_t$. In addition to deriving the emotion prototypes $\mathbf{V} = \{\mathbf{v}_y\}$, we also map all demonstrations in $\mathcal{D}_{\text{demo}}$ into the shared prototype space: $\mathbf{H}_{\text{demo}} = Enc(\mathcal{D}_{\text{demo}})$ ($\mathbf{H}_{\text{demo}} \in \mathbb{R}^{|\mathcal{Y}_s \cup \mathcal{Y}_t| \times K \times d}$). This enables the agent to directly judge the emotional intensity across different demonstrations.

3.3 Dual-Agent FER

With $\mathcal{D}_{\text{demo}}$, $\mathbf{H}_{\text{demo}}$ and $\mathbf{V}$ frozen, we fine-tune a dual-agent architecture comprising an Emotion Predictor ($\text{Agent}_{\text{emo}}$) and a Demonstration Retriever ($\text{Agent}_{\text{demo}}$) via GRPO (Shao et al., 2024) to resolve Challenge 2 (counterproductive retrieval).

3.3.1 Agent Collaboration

Given input $\mathbf{x}$, $\text{Agent}_{\text{emo}}$ first computes its candidate emotion set via prototype similarity:

$$\mathcal{Y}_{\text{cand}} = \text{Top-}\rho(\{\cos(Enc(\mathbf{x}), \mathbf{v}_y) \mid y \in \mathcal{Y}_s \cup \mathcal{Y}_t\}). \quad (4)$$

This dynamic filtering compresses the label space, reducing decision-boundary ambiguity for downstream prediction. Then $\mathcal{Y}_{\text{cand}}$ is passed to $\text{Agent}_{\text{demo}}$, which retrieves the most relevant demonstration per candidate emotion from $\mathcal{D}_{\text{demo}}$ via a template $T_{\text{demo}}(\mathbf{x}, \mathcal{Y}_{\text{cand}}, \mathcal{D}_{\text{demo}})$ instructing it to select the most relevant sentence from $\mathcal{D}_{\text{demo}}$ for each emotion in $\mathcal{Y}_{\text{cand}}$ given $\mathbf{x}$ (full template in Appendix A.3).

Rather than directly generating the selected demonstration, $\text{Agent}_{\text{demo}}$ determines the specific choices by maximizing embedding alignment:

$$z_y = \text{argmax}_{k \leq K} (\cos(\mathbf{h}_{\text{agent}}, \mathbf{h}_{\text{demo}}^{y,k})), \quad (5)$$

where $\mathbf{h}_{\text{agent}} \in \mathbb{R}^d$ is the [CLS] token of $\text{Agent}_{\text{demo}}$'s last hidden layer, and $\mathbf{h}_{\text{demo}}^{y,k} \in \mathbf{H}_{\text{demo}}$. The resulting $z = \{z_y \mid y \in \{\mathcal{Y}_s \cup \mathcal{Y}_t\}\}$ contains, for each candidate emotion, the index of its most relevant demonstration. The final retrieved demonstrations $S = \{s_1, s_2, \dots, s_{|z|}\}$ are obtained by indexing into $\mathcal{D}_{\text{demo}}$ using z : $S = \mathcal{D}_{\text{demo}}[z]$. With the retrieved S and $\mathcal{Y}_{\text{cand}}$, $\text{Agent}_{\text{emo}}$ outputs calibrated emotion probabilities $\{(y_k, p_k)\}$ via a template $T_{\text{emo}}(\mathbf{x}, \mathcal{Y}_{\text{cand}}, S)$ that asks the model to choose from $\mathcal{Y}_{\text{cand}}$ given $\mathbf{x}$ and S in the format “*emotion: probability%*” (full template in Appendix A.3). The emotion with the highest probability is selected as the final prediction y' .

3.3.2 GRPO Formalization

We formulate the joint training of $\text{Agent}_{\text{demo}}$ (θ_1) and $\text{Agent}_{\text{emo}}$ (θ_2) using GRPO, which samples a group of candidate trajectories, computes a dense prototype-aware reward for each, and optimizes both agents via relative advantages.

Joint Group Sampling. For a given input sentence $\mathbf{x}$, the retriever π_{θ_1} samples a group of G demonstration index sequences $\mathcal{Z} = \{z^{(1)}, z^{(2)}, \dots, z^{(G)}\}$. Each $z^{(g)}$ retrieves a demonstration set $S^{(g)} = \mathcal{D}_{\text{demo}}[z^{(g)}]$, which is then fed into the predictor π_{θ_2} . Conditioned on $(\mathbf{x}, S^{(g)})$, π_{θ_2} outputs a calibrated probability distribution over the candidate emotion space $\mathcal{Y}_{\text{cand}}$: $\mathcal{O}^{(g)} = \{(y_k^{(g)}, p_k^{(g)})\}_{k=1}^{|\mathcal{Y}_{\text{cand}}|}$. The complete rollout for group g is denoted as $\xi^{(g)} = (z^{(g)}, \mathcal{O}^{(g)})$.

Prototype-Aware Reward. To provide fine-grained supervision that captures subtle emotional nuances, we design a continuous reward function evaluating the semantic alignment between the predicted distribution and the ground-truth emotion

y^* . Leveraging the pre-learned emotion prototypes $\mathbf{V} = \{\mathbf{v}_y\}$ from Section 3.2.2, the reward for trajectory $\xi^{(g)}$ is computed as:

$$r^{(g)} = \sum_{k=1}^{|\mathcal{Y}_{\text{cand}}|} \cos(\mathbf{v}_{y^*}, \mathbf{v}_{y_k^{(g)}}) \cdot p_k^{(g)}. \quad (6)$$

This formulation yields a density-aware reward: it penalizes retrievals that drift toward semantically distant categories, while rewarding those that concentrate probability mass on prototypes geometrically close to y^* . With $r^{(g)}$, both agents are jointly optimized using the standard GRPO objective (details in Appendix A.3).

4 Experiments

4.1 Settings

Datasets. We evaluate on four fine-grained emotion benchmarks: Empathetic Dialogue (ED) (Rashkin et al., 2019), GoEmotions (GE) (Demszky et al., 2020), Empathetic Intent (EI) (Welivita and Pu, 2020), and EDOS (Welivita et al., 2021). Following the zero-shot protocol, we partition labels into seen (Plutchik’s 8 basic emotions) and unseen (complex emotions), excluding neutral classes. Unseen labels are split into validation/test sets at a 1:2 ratio. Dataset statistics are summarized in Table 1.

Implementation & Metrics. We initialize the sentence and prototype encoder with *all-mpnet-base-v2* (Reimers and Gurevych, 2019), and adopt Llama3-8B (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) as LLM backbones. Pseudo-labels are generated by GPT-4o-mini (OpenAI, 2024) with $M=5$ and $K=8$. We set $\tau=0.05$, batch size $B=16$ for $\mathcal{L}_{\text{proto}}$, and prototype filtering $\rho=0.4$, GRPO group size $G=8$ for the dual-agent stage. Both agents are jointly optimized with LoRA ($r=16$, $\alpha=32$) for 3 epochs on a single A100 GPU (batch size 4). We report accuracy, macro-F1, and weighted-F1 (wF1), averaged over 3 random seeds. Detailed GRPO implementation, hyperparameter sensitivity, and component sensitivity are provided in Appendix A.3, A.4, and A.5.

Baselines. We compare three paradigms: (1) *Prototype-based classifiers*: Learn emotion-specific prototype embeddings via contrastive loss (Eq. 3), then perform nearest-neighbor classification. Implemented with **RoBERTa**_{large} (Liu et al., 2019), **MPNet** (Song et al., 2020), and

Dataset	ED		GE		EI		EDOS	
	#E	#S	#E	#S	#E	#S	#E	#S
train	8	4,891	8	7,077	8	28,744	8	1,439
dev	8	1,306	6	951	8	6,464	10	452
test	16	2,584	13	3,202	16	12,779	22	952
$\mathcal{D}_u$	/	12,911	/	16,408	/	76,926	/	5,616

Table 1: Statistics of datasets. #E and #S denote the number of emotions and samples, respectively.

TACO (Gong et al., 2025). (2) *LLM-based methods*: Use both open-source (**Llama3-8B**, **Qwen3-8B**) and commercial (**GPT-4o-mini**, **DeepSeek-V3** (Liu et al., 2024a), **Claude-3.7-sonnet** (Anthropic, 2024)) LLMs, using golden demonstrations while maintaining the same prompt template with us. More LLMs are in Appendix A.6. (3) *Automatic prompt engineering & ICL retrieval*: **APE** (Zhou et al., 2023), **OPRO** (Yang et al., 2024), **MoT** (Li and Qiu, 2023) and **EICL** (Ren et al., 2025). All of them are implemented with Qwen3-8B and construct demonstrations from the same unlabeled corpus to ensure fairness, their details are provided in Appendix A.2.

4.2 Main Results

Table 2 presents the comprehensive performance. Built on Qwen3-8B, EmoDap consistently outperforms all baselines across four datasets; it surpasses the strongest LLM baseline (Claude-3.7-sonnet) by an average of 3.26% in accuracy, 2.55% in macro-F1, and 3.35% in wF1. With Llama3-8B, EmoDap remains highly competitive against commercial LLMs. Notably, **all LLM baselines rely on golden demonstrations, whereas EmoDap operates solely on pseudo-labeled ones**, highlighting the framework’s robustness under strict zero-shot constraints. Integrating prototype guidance with dual-agent retrieval yields consistent gains over LLM-only prompting (achieving an average wF1 improvement of 5.16% with EmoDap_{Qwen3-8B} compared to using Qwen3-8B alone), confirming that quantitative emotion anchoring effectively compensates for fine-grained decision ambiguity. Compared to automatic prompt engineering methods, APE/OPRO/EICL all rely on annotated feedback while MoT lacks emotion-aware alignment, limiting their zero-shot effectiveness. EmoDap circumvents these by leveraging prototype theory to constrain LLM reasoning, achieving superior generalization without task-specific supervision.

Method	ED			GE			EI			EDOS		
	ACC.	F1	wF1	ACC.	F1	wF1	ACC.	F1	wF1	ACC.	F1	wF1
Prototype-based												
RoBERTa _{large}	30.34	24.50	30.27	20.63	16.94	20.05	19.48	15.86	19.39	16.55	14.12	16.77
MPNet	31.78	25.22	31.45	21.56	17.99	21.11	20.70	17.29	21.43	17.66	15.08	16.91
TACO	33.23	27.65	34.02	24.41	20.14	23.97	24.21	20.18	22.70	19.24	17.53	20.80
LLM-based												
Llama3-8B	64.04	61.18	62.73	48.31	42.43	48.26	40.01	38.98	39.82	43.85	41.40	41.44
Qwen3-8B	70.12	70.39	70.94	52.65	46.50	56.05	40.58	39.96	41.33	43.12	42.89	43.08
GPT-4o-mini	69.39	65.18	69.16	54.62	42.05	55.38	38.23	39.77	40.49	42.54	40.06	41.43
Deepseek-V3	70.90	67.52	71.21	55.95	45.57	56.63	39.81	41.25	42.03	42.64	41.62	42.30
Claude-3.7-sonnet	74.33	74.21	73.98	58.15	47.79	58.47	40.91	44.85	42.57	44.82	45.27	43.63
Automatic prompt engineering												
APE	69.96	69.51	69.01	51.67	45.96	55.84	40.62	39.61	40.54	43.10	41.65	41.82
OPRO	69.39	69.31	69.60	51.09	45.82	55.37	40.05	39.46	39.09	43.53	42.61	42.36
EICL	69.85	69.64	69.52	51.56	45.01	55.29	40.51	39.67	40.02	44.02	42.81	42.28
MoT	69.03	69.32	68.86	51.17	45.60	55.14	39.75	39.44	39.10	43.74	42.90	41.70
EmoDap _{Llama3-8B}	72.14	72.41	72.53	53.85	46.88	54.31	42.79	39.52	43.02	47.27	46.67	46.29
EmoDap _{Qwen3-8B}	76.45	76.37	76.46	59.63	51.10	61.62	45.79	46.57	45.99	49.37	48.29	47.98
EmoDap-OOD _{Llama3-8B}	71.58	71.79	71.80	53.10	46.43	53.72	42.18	39.05	42.29	46.59	46.10	45.66
EmoDap-OOD _{Qwen3-8B}	<u>75.83</u>	<u>75.64</u>	<u>75.82</u>	<u>59.39</u>	<u>50.57</u>	<u>61.19</u>	<u>45.29</u>	<u>46.03</u>	<u>45.47</u>	<u>48.70</u>	<u>47.81</u>	<u>47.31</u>

Table 2: Main results on 4 datasets. The best results are in **bold**, while the second best are underlined.

Out-of-Domain Robustness. We evaluate an OOD setting where the source dataset $\mathcal{D}_s$ for prototype learning is replaced by randomly sampled data from the other three benchmarks. As shown in Table 2 (EmoDap-OOD), performance drops by only 0.4%–0.7%, while still significantly outperforming all baselines. This confirms that emotion prototypes capture generalizable semantic anchors rather than dataset-specific artifacts, demonstrating the strong generalization capability of our framework that can be readily extended to new domains.

4.3 Ablation Study

Table 3 isolates the contribution of each component. **Dual-agent decomposition** is critical: collapsing to a *single agent* reduces wF1 by 3.43%, as a unified model struggles to simultaneously optimize retrieval strategy and semantic classification. Removing *GRPO fine-tuning* degrades demonstration quality, confirming that preference alignment is necessary for learning effective selection policies. Disabling *self-consistency annotation* lowers pseudo-label reliability, while replacing *K-means* with random sampling reduces contextual diversity, both hurting performance. Removing the *prototype filter* forces the predictor to handle dense emotional interference alone, causing a 6.25% wF1 drop, validating its role in compressing the candidate space.

Retrieval Policy Ablation. To further analyze the contribution of $\text{Agent}_{\text{demo}}$, we keep the trained $\text{Agent}_{\text{emo}}$ frozen and replace only the demonstration-selection rule, while still using the

Method	ACC.	F1	wF1
EmoDap_{Qwen3-8B}	57.81	55.58	58.01
w/ single-agent	55.19	51.39	54.58
w/o GRPO fine-tuning	55.44	52.67	54.71
w/o self-consistency	55.01	51.55	54.36
w/o K-means	55.47	52.43	54.44
w/o prototype	51.88	48.80	51.76
<i>Retrieval policy variants ($\text{Agent}_{\text{emo}}$ frozen):</i>			
$\text{Agent}_{\text{demo}} \leftarrow \text{adversarial}^\dagger$	50.06	48.10	50.21
$\text{Agent}_{\text{demo}} \leftarrow \text{random}$	55.30	52.03	54.62
$\text{Agent}_{\text{demo}} \leftarrow \text{centroid}$	55.31	52.53	54.63
$\text{Agent}_{\text{demo}} \leftarrow \text{nearest-NN}$	55.56	52.73	54.66
$\text{Agent}_{\text{demo}} \leftarrow \text{oracle}^*$	59.33	56.97	59.51

Table 3: Averaged ablation results. **Random**: uniform sampling; **Centroid/Nearest-NN**: closest to class centroid or query (standard similarity retrieval); **Adversarial[†]**: most query-similar demonstration *outside* the candidate class (worst-case lower bound); **Oracle^{*}**: exhaustive search for the same-class demonstration maximizing predictor confidence (empirical upper bound). Per-dataset results in Appendix Table 11.

same $\mathcal{D}_{\text{demo}}$. We compare five strategies that span a wide spectrum of retrieval quality (Table 3, lower part). Three observations follow. (i) Similarity is not enough: random/centroid/nearest-NN retrievers cluster within 0.04% wF1 of one another, confirming that surface-similar demonstrations carry little discriminative signal in fine-grained settings. (ii) Counterproductive retrieval is real: the adversarial retriever drops 4.4% wF1 below random, directly evidencing Challenge 2, since a similarity-maximizing but emotion-blind policy actively misleads the predictor. (iii) Closing the oracle gap: the trained $\text{Agent}_{\text{demo}}$ recovers

Reward variant	ACC.	F1	wF1
Confidence-only (p_{y^*})	55.62	53.18	55.83
Cosine-only (top-1 weighted)	53.91	51.48	54.07
Prototype-aware (ours)	57.81	55.58	58.01

Table 4: Reward ablation, averaged over four datasets.



Figure 3: Shared-reward training dynamics on ED. Blue: joint prototype-aware reward $r^{(g)}$ (Eq. 6); orange (right axis): KL to the reference policy. Lines: mean over 3 seeds; bands: ± 1 std.

$\sim 69\%$ of the gap between nearest-NN (54.66%) and the oracle (59.51%), demonstrating that joint GRPO optimization learns substantially more than what cosine similarity can express. Substitution-robustness is deferred to Appendix A.7.

Reward Function Ablation. The GRPO reward in Eq. 6 couples predictor confidence $p_k^{(g)}$ with prototype similarity $\cos(\mathbf{v}_{y^*}, \mathbf{v}_{\hat{y}_k^{(g)}})$. To verify that this coupling, rather than either ingredient alone, drives the gain, we compare three variants while keeping every other component fixed: **Confidence-only** ($r^{(g)} = p_{y^*}^{(g)}$, hard-label likelihood); **Cosine-only** ($r^{(g)} = \cos(\mathbf{v}_{y^*}, \mathbf{v}_{\hat{y}_k^{(g)}})$ with $\hat{y}_k^{(g)} = \arg \max_k p_k^{(g)}$, scoring only the top-1 prediction); and our full **Prototype-aware** form (Eq. 6).

Table 4 shows that confidence-only loses 2.18% wF1, since a hard-label reward cannot distinguish “confidently wrong on a near-synonym” from “confidently wrong on an unrelated emotion”. Cosine-only is even worse (-3.94%), because collapsing the distribution onto its top-1 discards the very information GRPO uses to differentiate trajectories within a group. The full prototype-aware reward consistently dominates across all three metrics, confirming that the soft, distribution-weighted coupling is what turns prototype geometry into an effective training signal.

4.4 Training Dynamics

We analyze the joint optimization from two complementary angles: (i) whether a single shared reward suffices to coordinate both agents, and (ii) how the retriever’s selection criterion evolves.

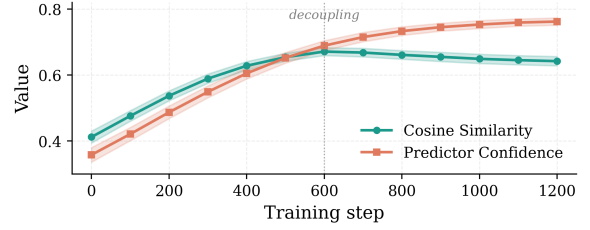


Figure 4: Retrieval-prediction alignment on ED. Predictor confidence and query-demo cosine similarity co-rise initially and decouple after ~ 600 steps.

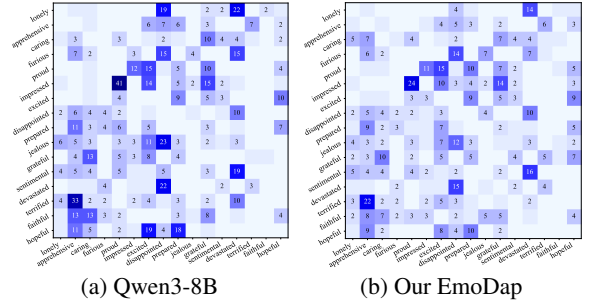


Figure 5: Confusion matrices on the ED dataset. Heatmap intensity shows confusion strength, with the diagonal set to zero.

Shared-Reward Convergence. A defining property of our joint optimization is that $\text{Agent}_{\text{demo}}$ and $\text{Agent}_{\text{emo}}$ are driven by a *single* prototype-aware reward $r^{(g)}$ (Eq. 6): the same scalar credits both retrieval and prediction, binding the two agents into one cohesive policy rather than two competing objectives. Figure 3 shows the smoothed shared reward on ED (3 seeds, window 50): it rises monotonically from ~ 0.45 to ~ 0.72 in two phases (fast warm-up before ~ 600 steps, slow refinement after), with the cross-seed std shrinking from 0.027 to below 0.01 at convergence and the KL to π_{ref} plateauing near 0.05. The absence of oscillation or seed divergence verifies that one shared reward is sufficient to coordinate the dual-agent system without auxiliary per-agent objectives.

Retrieval-Prediction Alignment. To examine *how* the retriever drifts away from similarity heuristics during training, we track two quantities along GRPO updates on ED (Figure 4): predictor confidence on the ground-truth emotion conditioned on the retrieved demonstration, and cosine similarity between the query embedding and the retrieved demonstration. The two curves co-rise during the first ~ 600 steps, when the retriever is still anchored on similarity, and then *decouple*: cosine similarity plateaus and slightly declines while predictor confi-

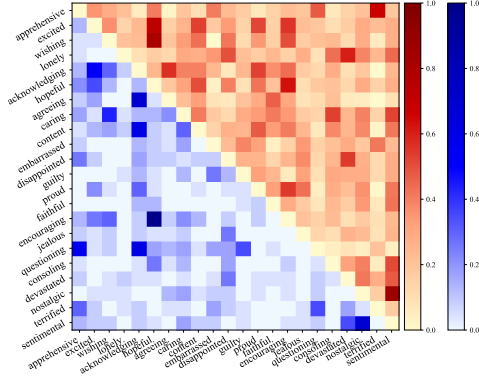


Figure 6: Combined heatmap of the confusion matrix (lower/blue) and cosine similarity (upper/red) for complex emotions in the EDOS dataset.

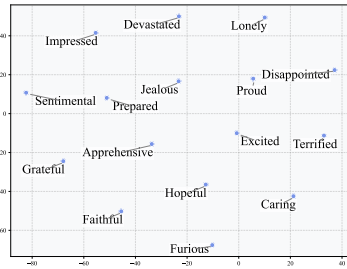


Figure 7: t-SNE visualization of emotion prototype embeddings in the ED dataset.

dence keeps climbing, the signature of preference-driven retrieval that occasionally trades surface similarity for downstream prediction quality.

4.5 Qualitative Analysis & Interpretability

Confusion Mitigation & Prototype Correlation.

Figure 5 shows that EmoDap systematically suppresses inter-class leakage on ED, particularly for challenging pairs (e.g., *impressed* vs. *proud*) that frequently mislead conventional models. We attribute this to prototype-guided filtering and contextually aligned demonstration retrieval. To quantify the tie between prototype geometry and residual errors, Figure 6 overlays the normalized confusion intensity with prototype cosine similarity on EDOS: emotion pairs with higher prototypical similarity (e.g., *encouraging* vs. *hopeful*, *sentimental* vs. *nostalgic*) exhibit stronger residual confusion, and vice versa, confirming that prototype quality dominates the framework’s residual discriminative limits. Extended confusion analyses across all four datasets and a per-emotion gain breakdown are provided in Appendices A.8 and A.9.

Prototype Geometry. We visualize the geometric structure of learned prototypes via t-SNE on

Model	Cost	Time	ACC.	wF1
Deepseek-V3	0.6\$	0.32×5 sec	52.33	53.04
Claude-3.7-sonnet	9.8\$	0.34×5 sec	54.55	54.66
Qwen3-8B	0\$	0.31×5 sec	51.62	52.85
EmoDap (Ours)	0\$	0.74 sec	57.81	58.01

Table 5: Comparison of call costs (per 1000 instances) and inference time (per instance) under one-shot and self-consistency baselines, averaged over four datasets.

Predictor	1-shot	+ EmoDap retr.	Δ
Llama3-8B	48.06	52.43	+4.37
Qwen2.5-7B	51.24	55.34	+4.10
GPT-4o-mini	51.62	55.26	+3.64
DeepSeek-V3	53.04	56.75	+3.71

Table 6: Cross-predictor transfer of the Qwen3-8B trained Agent_{demo} (averaged wF1, %).

the 16 unseen emotions of ED (Figure 7). Emotion clusters are clearly separated, and previously confusable pairs (e.g., *impressed* vs. *proud*, *sentimental* vs. *nostalgic*) occupy distinct regions, providing a geometric explanation for the confusion-matrix mitigation reported below. A detailed case study of demonstration retrieval on a hard instance is provided in Appendix A.10.

4.6 Efficiency and Transferability

Inference Efficiency. Table 5 compares API cost and inference latency. Baselines employing self-consistency voting require five LLM calls per instance ($0.32\text{--}0.37 \times 5$ sec), whereas EmoDap applies self-consistency only during precomputation, achieving lower latency (0.74 sec) and zero API cost with open-source backbones.

Cross-Predictor Transferability. To test whether Agent_{demo} learns Qwen3-8B-specific habits or a universal notion of demonstration quality, we freeze the retriever trained with Qwen3-8B and pair it with four *unseen* predictors, comparing against each predictor’s cosine-similarity 1-shot baseline averaged over the four benchmarks. Table 6 reports a consistent $\sim 4\%$ wF1 gain across all four predictors, comparable to the 5.16% wF1 in-distribution gain on Qwen3-8B. This indicates that Agent_{demo} captures predictor-agnostic demonstration quality and can be trained once on a small LLM and reused as a drop-in module for stronger but costlier predictors.

5 Conclusion

In this paper, we present EmoDap, a GRPO-optimized dual-agent framework for zero-shot FER.

By jointly learning a prototype-guided emotion predictor and a preference-driven demonstration retriever, EmoDap effectively mitigates semantic interference and counterproductive retrieval. Comprehensive evaluations across four benchmarks establish consistent gains over strong baselines, validating our approach as a robust, zero-shot paradigm for affective NLP. Future work will extend this framework to multimodal emotion understanding.

6 Acknowledgements

This research is supported by the National Key R&D Program of China (Grant No. 2023YFC3303800) and the National Natural Science Foundation of China (Grant Nos. U25B2029 and 62322202).

7 Limitations

We identify four limitations in our work.

First, our approach is limited to single-modality text analysis, whereas real-world emotional expression often involves multiple modalities, such as memes, which imposes constraints on practical applications.

Second, our emotion prototypes are derived from Oxford English Dictionary definitions, which may reflect Western-centric affective categorizations. We acknowledge that automated emotion recognition should not replace professional psychological assessment, and future work will validate EmoDap across multilingual and culturally diverse corpora to mitigate demographic bias.

Third, the dominant residual error mode is *prototype collapse on near-synonym pairs*. A manual audit of 100 random EmoDap errors on ED attributes 43% of failures to emotion pairs whose definitions overlap heavily (e.g., *sentimental* vs. *nostalgic*, *apprehensive* vs. *hopeful*), so $\mathcal{V}_{\text{cand}}$ retains both and the geometric guidance flattens. This matches the small per-class gains for these pairs in Appendix A.9 and points to richer (e.g., multi-modal or contextually grounded) prototypes as a concrete next step.

Fourth, both the prototype filter and the demonstration pool are built from sincere, literal expressions, leaving EmoDap blind to *sarcasm and figurative language* (29% of audited errors); a sarcasm-aware reranker or non-literal augmentation of the demonstration pool would be required to close this gap. A small share of remaining errors stems from over-aggressive ρ filtering and noisy ground-truth

labels, with the full breakdown in Appendix A.11.

References

- Anthropic. 2024. Claude 3.7 sonnet: The first hybrid reasoning model. Accessed: 2025-07-28.
- Chenran Cai, Qianlong Wang, Bin Liang, Bing Qin, Min Yang, Kam-Fai Wong, and Ruifeng Xu. 2023. [In-context learning for few-shot multimodal named entity recognition](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2969–2979, Singapore. Association for Computational Linguistics.
- Chih Yao Chen, Tun Min Hung, Yi-Li Hsu, and Lun-Wei Ku. 2023. [Label-aware hyperbolic embeddings for fine-grained emotion classification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10947–10958.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hila Gonen, Sridhar Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. [Demystifying prompts in language models via perplexity estimation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148.
- Junqing Gong, Binhan Yang, and Wei Shen. 2025. [A triple-view framework for fine-grained emotion classification with clustering-guided contrastive learning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4970–4984, Vienna, Austria. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- James A Hampton. 2006. Concepts as prototypes. *Psychology of learning and motivation*, 46:79–113.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

- Zhong Ji, Biying Cui, Yunlong Yu, Yanwei Pang, and Zhongfei Zhang. 2021. [Zero-shot classification with unseen prototype learning](#). *Neural Comput. Appl.*, 35(17):12307–12317.
- Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. [Attribute-based classification for zero-shot visual object categorization](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465.
- Itay Levy, Ben Bogin, and Jonathan Berant. 2023. [Diverse demonstrations improve in-context compositional generalization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.
- Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. [Unified demonstration retriever for in-context learning](#). In *ACL*, pages 4644–4668.
- Xiaonan Li and Xipeng Qiu. 2023. [MoT: Memory-of-thought enables ChatGPT to self-improve](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6354–6374.
- Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. [OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *CoRR*.
- Siya Liu, Yang Li, Jiang Li, Shan Yang, and Yunshi Lan. 2024b. [Unleashing the power of large language models in zero-shot relation extraction via self-prompting](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13147–13161, Miami, Florida, USA. Association for Computational Linguistics.
- 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. *arXiv preprint arXiv:1907.11692*.
- Yuanyuan Liu, Ke Wang, Lin Wei, Jingying Chen, Yibing Zhan, Dapeng Tao, and Zhe Chen. 2024c. Affective computing for healthcare: Recent trends, applications, challenges, and beyond. *arXiv preprint arXiv:2402.13589*.
- Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024d. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.
- Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. [In-context learning with retrieved demonstrations for language models: A survey](#). *Preprint*, arXiv:2401.11624.
- Minxue Niu and Emily Mower Provost. 2025. Contrastive distillation of emotion knowledge from llms for zero-shot emotion recognition. *arXiv preprint arXiv:2505.18040*.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2024-01-18.
- Rosalind W Picard. 2000. *Affective computing*. MIT press.
- Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.
- Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *EMNLP-IJCNLP*, pages 3982–3992.
- 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 *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25*, page 2503–2513, New York, NY, USA. Association for Computing Machinery.
- Peter J. Rousseeuw. 1987. [Silhouettes: A graphical aid to the interpretation and validation of cluster analysis](#). *Journal of Computational and Applied Mathematics*, 20:53–65.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. [Learning to retrieve prompts for in-context learning](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Alexander Shvets. 2025. [Emo pillars: Knowledge distillation to support fine-grained context-aware and context-less emotion classification](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 174–191, Vienna, Austria. Association for Computational Linguistics.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MpNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.
- Ala N Tak and Jonathan Gratch. 2023. Is gpt a computational model of emotion? detailed analysis. *arXiv preprint arXiv:2307.13779*.
- Ala N Tak and Jonathan Gratch. 2024. Gpt-4 emulates average-human emotional cognition from a third-person perspective. *arXiv preprint arXiv:2408.13718*.
- Gemini Team. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Xingchen Wan, Ruoxi Sun, Hootan Nakhost, Hanjun Dai, Julian Martin Eisenschlos, Sercan O Arik, and Tomas Pfister. 2023. Universal self-adaptive prompting. *arXiv preprint arXiv:2305.14926*.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. [Text embeddings by weakly-supervised contrastive pre-training](#). *ArXiv*, abs/2212.03533.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025a. [GPT-NER: Named entity recognition via large language models](#). In *Findings of NAACL*, pages 4257–4275.
- Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.
- Zihan Wang, Ziqi Zhao, Yougang Lyu, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2025b. A co-operative multi-agent framework for zero-shot named entity recognition. In *Proceedings of the ACM on Web Conference 2025*, pages 4183–4195.
- Anuradha Welivita and Pearl Pu. 2020. [A taxonomy of empathetic response intents in human social conversations](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. [A large-scale dataset for empathetic response generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. [Self-improving for zero-shot named entity recognition with large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 583–593, Mexico City, Mexico. Association for Computational Linguistics.
- Xin Zhou Xu, Jun Deng, Zixing Zhang, Zhen Yang, and Björn W. Schuller. 2023. [Zero-shot speech emotion recognition using generative learning with reconstructed prototypes](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.
- Pinyi Zhang, Jingyang Chen, Junchen Shen, Zijie Zhai, Ping Li, Jie Zhang, and Kai Zhang. 2024. Message passing on semantic-anchor-graphs for fine-grained emotion representation learning and classification. In *EMNLP*, pages 2771–2783.
- Weixiang Zhao, Zhuojun Li, Shilong Wang, Yang Wang, Yulin Hu, Yanyan Zhao, Chen Wei, and Bing Qin.

2024. Both matter: Enhancing the emotional intelligence of large language models without compromising the general intelligence. *arXiv preprint arXiv:2402.10073*.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chat-gpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Prototype Learning for Zero-Shot Classification

Cognitive models posit that categories form around a central anchor point (Hampton, 2006), a principle that extends to semantic spaces: texts expressing *Angry*, for instance, cluster around a statistical prototype, reducing classification to prototype comparison. For unseen emotions lacking labeled data, prior work (Lampert et al., 2014) shows that models can infer prototypes from textual descriptions alone, enabling zero-shot classification. Despite wide adoption (Ji et al., 2021; Xu et al., 2023), generalization remains limited; recent methods such as TACO (Gong et al., 2025) improve prototype quality via cluster-guided contrastive learning on supervised emotions, but their transfer to unseen emotions is unverified.

A.2 Datasets and Baseline

Datasets

We evaluate on four benchmarks:

- **Empathetic Dialogue (ED)** (Rashkin et al., 2019): two-speaker conversations with 32 emotions (8 basic + 24 complex).
- **GoEmotion (GE)** (Demszky et al., 2020): Reddit comments with 27 emotions plus neutral. Since its label set covers only 6 of Plutchik’s basic emotions, we map the two closest alternatives (*Approval*→*Trust*, *Optimism*→*Anticipation*) and exclude neutral, yielding 8 basic + 19 complex.
- **Empathetic Intents (EI)** (Welivita and Pu, 2020): an ED extension with 32 emotions (8+24).
- **EDOS** (Welivita et al., 2021): curated from OpenSubtitles (Lison et al., 2018), 40 emotions (8+32) plus neutral.

ED and GE are the largest and most challenging FER benchmarks, preprocessed following (Chen et al., 2023; Zhang et al., 2024); EI and EDOS are from the empathetic-response domain.

Automatic Prompt Engineering Baselines

We provide more details about the automatic prompt engineering and ICL demonstration retrieval approaches here:

- **APE** (Zhou et al., 2023) uses an LLM to generate candidate prompts, evaluates their performance on

Algorithm 1 EmoDap Training and Inference

Require: Training set $\mathcal{D}_s$, unlabeled corpus $\mathcal{D}_u$, emotion sets $\mathcal{Y}_s, \mathcal{Y}_t$, cluster size K , filtering ratio ρ , group size G , KL coefficient β .

Ensure: Optimized joint parameters $\Theta = \{\theta_1, \theta_2\}$ and predictions y' for inputs x .

// Stage 1: Precomputation (Frozen)

- 1: $\tilde{\mathcal{D}}_u \leftarrow \text{SelfConsistencyAnnotation}(\mathcal{D}_u, M)$
- 2: $\mathcal{D}_{\text{demo}} \leftarrow \text{ClusteringAndSampling}(\tilde{\mathcal{D}}_u, K)$
- 3: $\mathbf{V}, \text{Enc}(\cdot) \leftarrow \text{ContrastivePrototypeLearning}(\mathcal{D}_s, \mathcal{Y}_s)$

// Stage 2: GRPO-Optimized Dual-Agent Training

- 4: **for** each epoch **do**
- 5: **for** batch $(x, y^*) \in \mathcal{D}_s$ **do**
- 6: $Y_{\text{cand}} \leftarrow \text{Top-}\rho(\{\cos(\text{Enc}(x), \mathbf{v}_y) \mid y \in \mathcal{Y}_s \cup \mathcal{Y}_t\})$
- 7: // Agent_{emo} filtering
- 8: Sample $\mathcal{Z} = \{z^{(1)}, \dots, z^{(G)}\} \sim \pi_{\theta_1^{\text{old}}}(\cdot \mid x)$
- 9: **for** $g = 1$ **to** G **do**
- 10: $S^{(g)} \leftarrow \mathcal{D}_{\text{demo}}[z^{(g)}]$
- 11: $\mathcal{O}^{(g)} \leftarrow \pi_{\theta_2}(x, S^{(g)}, Y_{\text{cand}})$ // Agent_{emo} rollout
- 12: $r^{(g)} \leftarrow \sum_k \cos(\mathbf{v}_{y^*}, \mathbf{v}_{y_k^{(g)}}) \cdot p_k^{(g)}$ // Eq. 6
- 13: $A^{(g)} \leftarrow \text{GroupNormalize}(\{r^{(g)}\}_{g=1}^G)$
- 14: Update $\Theta = \{\theta_1, \theta_2\}$ via $\nabla_{\Theta} \mathcal{L}_{\text{GRPO}}(\Theta)$ (Eq. 8)
- 15: **end for**
- 16: **end for**

// Stage 3: Inference

- 17: $Y_{\text{cand}} \leftarrow \text{Top-}\rho(\{\cos(\text{Enc}(x), \mathbf{v}_y)\})$
- 18: $z^* \leftarrow \arg \max_z \pi_{\theta_1}(z \mid x, Y_{\text{cand}}), S^* \leftarrow \mathcal{D}_{\text{demo}}[z^*]$
- 19: $y' \leftarrow \arg \max \pi_{\theta_2}(x, S^*, Y_{\text{cand}})$
- 20:
- 21: **return** y'

a training set, and applies a Monte Carlo search to optimize for high-quality prompts.

- **OPRO** (Yang et al., 2024) optimizes prompts by leveraging historical iteration data (performance, patterns) and LLM reasoning to guide refinements.
- **MoT** (Li and Qiu, 2023) first pre-processes unlabeled data to build memory, then retrieves the most useful memories at the inference stage.
- **EICL** (Ren et al., 2025) leverages dynamic emotion labels and uses an exclusionary prediction strategy to filter out irrelevant emotional categories.

For a fair comparison, it should be mentioned that these retrieval-based methods share a commonality with EmoDap: both construct demonstrations by retrieving from the same set $\mathcal{D}^{\text{demo}}$.

A.3 GRPO Training Details

Our EmoDap framework’s pseudo algorithm is shown in Algorithm 1.

Prompt Templates. The two prompt templates used in Section 3 are:

$T_{\text{demo}}(\mathbf{x}, \mathcal{Y}_{\text{cand}}, \mathcal{D}_{\text{demo}}) = \text{Given the sentence: } \{\mathbf{x}\} \setminus n. \text{ For each emotion in } \{\mathcal{Y}_{\text{cand}}\}, \text{ please select the most relevant sentence from the demonstration pools: } \{\mathcal{D}_{\text{demo}}\}.$

$T_{\text{emo}}(\mathbf{x}, \mathcal{Y}_{\text{cand}}, S) = \text{Given a sentence, please determine the emotion it conveys. Here are demonstrations of these emotions: } \{S\}. \setminus n \text{ Please understand the meaning of each emotion through these demonstrations. Sentence: } \{\mathbf{x}\} \setminus n \text{ Choose your answer from } \{\mathcal{Y}_{\text{cand}}\}. \setminus n \text{ Output format: emotion: probability}\%.$

Gradient Pathway. Although the final retrieval $z_y = \arg \max_k \cos(\mathbf{h}_{\text{agent}}, \mathbf{h}_{\text{demo}}^{y,k})$ is non-differentiable, gradients to θ_1 flow through the standard policy-gradient pathway: at training time, $\mathbf{h}_{\text{agent}}$ is conditioned on the autoregressive token sequence generated by π_{θ_1} under T_{demo} , and a group of G trajectories $\{z^{(g)}\}$ is obtained by token-level stochastic sampling from π_{θ_1} . The GRPO surrogate then back-propagates the relative advantage $A^{(g)}$ to all sampled tokens of π_{θ_1} , while π_{θ_2} receives gradients from its own emitted probability tokens that produce $\mathcal{O}^{(g)}$. The argmax over cached $\mathbf{H}_{\text{demo}}$ acts only as an inference-time index lookup and is bypassed in gradient computation.

GRPO Objective. GRPO eliminates the need for a separate value network by normalizing rewards within the sampled group. The relative advantage for each trajectory is:

$$A^{(g)} = \frac{r^{(g)} - \text{mean}(\{r^{(i)}\}_{i=1}^G)}{\text{std}(\{r^{(i)}\}_{i=1}^G) + \epsilon_{\text{stab}}}, \quad (7)$$

where ϵ_{stab} ensures numerical stability. Both agents are jointly optimized by maximizing the standard GRPO clipped surrogate objective:

$$\mathcal{L}_{\text{GRPO}}(\Theta) = \mathbb{E}_{x, \{\xi^{(g)}\} \sim \pi_{\Theta_{\text{old}}}} \left[\frac{1}{G} \sum_{g=1}^G \min \left(\omega^{(g)} A^{(g)}, \text{clip}(\rho^{(g)}, 1 - \epsilon, 1 + \epsilon) A^{(g)} \right) - \beta D_{\text{KL}}(\pi_{\Theta} \parallel \pi_{\Theta_{\text{ref}}}) \right]. \quad (8)$$

where $\Theta = \{\theta_1, \theta_2\}$, $\omega^{(g)} = \frac{\pi_{\Theta}(\xi^{(g)}|x)}{\pi_{\Theta_{\text{old}}}(\xi^{(g)}|x)}$ is the importance sampling ratio, and $\pi_{\Theta_{\text{ref}}}$ is a frozen reference policy initialized from $\pi_{\Theta_{\text{old}}}$. The KL penalty regularizes both agents, preventing retrieval degeneracy and maintaining calibrated probability outputs during prediction. Through iterative updates, the dual-agent system learns a co-

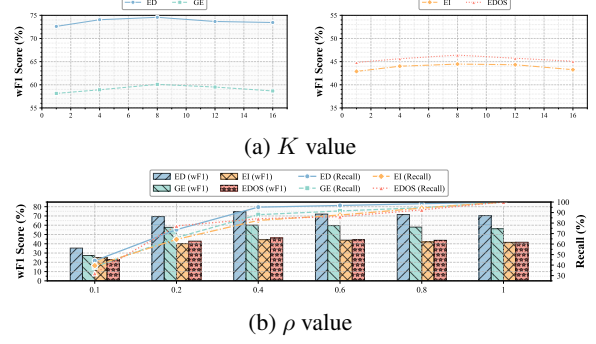


Figure 8: Hyperparameter sensitivity.

Model	ACC.	FI	wFI	APCD
GPT-4o-mini	57.81	55.58	58.01	1.47
w/ GPT-3.5-turbo	57.33	54.61	57.56	1.40
w/ Qwen2.5-7B	57.48	55.22	57.74	1.42

Table 7: The impact of different LLMs as pseudo-annotators, where APCD denotes average pairwise cosine distance.

hesive strategy where $\text{Agent}_{\text{demo}}$ strategically selects demonstrations that maximize $\text{Agent}_{\text{emo}}$ ’s prototype-aligned prediction confidence.

In the dual-agent framework, we set the GRPO group size $G = 8$. The KL regularization coefficient is set to $\beta = 0.02$, with clipping parameter $\epsilon = 0.2$. Both $\text{Agent}_{\text{demo}}$ and $\text{Agent}_{\text{emo}}$ are jointly optimized via the GRPO objective (Eq. (8)) for 3 epochs on a single A100 GPU (80GB) with a batch size of 4. The reference policy $\pi_{\Theta_{\text{ref}}}$ is initialized from the base model and updated at the end of each epoch to stabilize training.

A.4 Hyperparameter Sensitivity

Figure 8 analyzes sensitivity to cluster size K and filtering threshold ρ . Performance peaks at $K = 8$; smaller K restricts demonstration diversity, while larger K introduces redundancy. The optimal threshold is $\rho = 0.4$: lowering ρ severely compromises recall of the ground-truth emotion, whereas higher ρ admits distractor categories that dilute prediction confidence. These results indicate that moderate candidate compression balances recall and precision with computational efficiency.

A.5 Component Sensitivity Studies

We further dissect three input-side design choices that determine the quality of the prototype space and the demonstration pool: (i) the LLM *pseudo-annotator*, (ii) the *sentence encoder* used to embed prototypes and queries, and (iii) the *prototype source*, i.e., the textual definition from which each

emotion’s anchor is derived. All results are averaged over the four datasets.

LLM annotator. The quality of pseudo-labels conditions both the demonstration pool and the GRPO reward signal. We replace the default GPT-4o-mini with two weaker alternatives (GPT-3.5-turbo and Qwen2.5-7B) and report the resulting end-to-end performance and the average pairwise cosine distance (APCD) of the demonstration pool in Table 7. APCD is computed as

$$\text{APCD} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n (1 - \cos(h_i, h_j)), \quad (9)$$

where h_i, h_j are the prototype-encoder representations of two clustered samples; APCD ranges in $[0, 2]$, with larger values indicating greater semantic dispersion. Two observations stand out: (1) end-to-end wF1 varies by only 0.45% across the three annotators, indicating that K-means clustering effectively filters out low-quality pseudo-labels; (2) the diversity of the resulting pool ($\text{APCD} \in [1.40, 1.47]$) is essentially insensitive to the annotator, because the unlabeled corpus is large enough that any reasonable annotator can recover a diverse demonstration set.

Sentence encoder. The encoder $\text{Enc}(\cdot)$ shapes the geometry of the prototype space and is therefore a key knob in EmoDap. We replace the default *all-mpnet-base-v2* with two popular alternatives, SimCSE-RoBERTa_{large} (Gao et al., 2021) and E5-base (Wang et al., 2022), retrain only the contrastive prototype objective $\mathcal{L}_{\text{proto}}$ while keeping the dual-agent stage identical, and re-evaluate end-to-end. Table 8 reports both downstream wF1 and the prototype-space silhouette score (Rousseeuw, 1987), which measures how compactly each emotion clusters around its prototype (higher = better). Two observations follow. (i) All three encoders yield comparable end-to-end performance, confirming that EmoDap’s gain stems from the dual-agent retrieval–prediction loop rather than encoder-specific embeddings. (ii) The default mpnet attains the best balance between silhouette (0.31) and downstream wF1 (58.01%), justifying our choice.

Prototype source. EmoDap derives each emotion prototype from its Oxford English Dictionary (OED) definition. To probe whether this choice is essential, we replace the OED definition with two alternative textual sources of comparable length,

Sentence Encoder	Sil.↑	ACC.	wF1
all-mpnet-base-v2 (ours)	0.31	57.81	58.01
SimCSE-RoBERTa _{large}	0.27	57.12	57.40
E5-base	0.29	57.35	57.61

Table 8: Effect of the sentence encoder on the prototype-space silhouette score and the downstream end-to-end performance, averaged over four datasets.

Prototype source	ACC.	wF1
Oxford Dictionary (ours)	57.81	58.01
WordNet glosses	57.43	57.66
GPT-4o-generated descriptions	57.18	57.34

Table 9: Effect of the textual source from which each emotion prototype is derived, averaged over four datasets.

keeping every other component fixed: (i) WordNet glosses, and (ii) GPT-4o-generated 1–2-sentence descriptions of each emotion (we paraphrase three times per label and average the encoded prototype). Results are summarized in Table 9. OED definitions yield the strongest performance, slightly outperforming both lexical (WordNet) and LLM sources. LLM-generated descriptions, despite being highly fluent, lag by $\sim 0.7\%$ wF1, likely because they paraphrase common usage and therefore underweight the contrastive features that distinguish near-synonym emotions. Crucially, the spread across all three sources is within 1.0% wF1, indicating that EmoDap is robust to the prototype source as long as it carries discriminative semantic content; the dual-agent retrieval–prediction loop compensates for moderate variations in prototype quality.

A.6 Comparison with More LLMs

To assess scalability, we additionally evaluate Qwen3-14B against leading commercial LLMs: **GPT-4o-mini**, **DeepSeek-V3** (Liu et al., 2024a), **Claude-3.7-sonnet** (Anthropic, 2024), **Gemini-2.5-flash** (Team, 2025), and **GPT-4o** (Hurst et al., 2024). Results in Table 10 confirm our main findings: EmoDap on Qwen3-14B yields consistent and statistically significant gains over even the strongest baseline (GPT-4o) across all datasets, demonstrating its generalization. Detailed ablation results are reported in Table 11.

A.7 Demonstration Substitution Robustness

To verify that $\text{Agent}_{\text{demo}}$ ’s choice is not arbitrary among same-class candidates, we conduct a substitution study on the ED test set: holding the predic-

Model	ED			GE			EI			EDOS		
	ACC.	F1	wF1	ACC.	F1	wF1	ACC.	F1	wF1	ACC.	F1	wF1
GPT-4o-mini	68.66	64.44	68.48	53.80	41.34	54.60	37.58	39.05	39.73	41.71	39.25	40.75
w/ one-shot	69.39	65.18	69.16	54.62	42.05	55.38	38.23	39.77	40.49	42.54	40.06	41.43
Deepseek-V3	70.03	66.53	70.33	55.17	44.69	55.92	39.09	40.48	41.22	41.86	40.89	41.52
w/ one-shot	70.90	67.52	71.21	55.95	45.57	56.63	39.81	41.25	42.03	42.64	41.62	42.30
Claude-3.7-sonnet	73.26	73.48	72.53	57.87	46.27	58.22	39.93	43.63	42.11	44.52	43.87	42.93
w/ one-shot	74.33	74.21	73.98	58.15	47.79	58.47	40.91	44.85	42.57	44.82	45.27	43.63
Gemini-2.5-flash	74.52	74.97	74.01	58.67	48.39	58.97	41.24	45.31	43.33	45.26	45.82	44.44
w/ one-shot	75.28	75.50	75.45	59.94	48.64	60.50	42.13	45.99	44.36	46.79	46.25	44.98
GPT-4o	75.26	74.66	74.23	59.50	47.69	59.88	40.73	45.29	43.69	45.70	45.68	44.12
w/ one-shot	76.02	75.85	75.90	<u>60.55</u>	48.45	60.78	42.48	45.82	44.51	47.34	46.08	45.22
Llama3-8B	62.96	60.36	61.77	47.01	41.65	47.19	38.95	38.17	38.99	42.86	40.54	40.61
w/ one-shot	64.04	61.18	62.73	48.31	42.43	48.26	40.01	38.98	39.82	43.85	41.40	41.44
w/ EmoDap	72.14	72.41	72.53	53.85	46.88	54.31	42.79	39.52	43.02	47.27	46.67	46.29
Qwen2.5-7B	67.62	68.15	68.43	50.27	44.31	53.70	38.35	37.73	38.87	40.82	40.55	40.77
w/ one-shot	68.56	68.87	69.23	51.10	45.04	54.57	39.13	38.47	39.66	41.55	41.27	41.49
w/ EmoDap	74.42	74.37	74.57	58.18	49.55	60.09	44.38	44.16	44.47	47.90	46.82	46.41
Qwen3-8B	69.65	70.08	70.37	52.10	46.14	55.63	40.18	39.56	40.70	42.65	42.38	42.60
w/ one-shot	70.12	70.39	70.94	52.65	46.50	56.05	40.58	39.96	41.33	43.12	42.89	43.08
w/ EmoDap	<u>76.45</u>	<u>76.37</u>	<u>76.46</u>	59.63	<u>51.10</u>	<u>61.62</u>	<u>45.79</u>	<u>46.57</u>	<u>45.99</u>	<u>49.37</u>	<u>48.29</u>	<u>47.98</u>
Qwen3-14B	72.20	70.61	71.59	55.54	44.49	55.87	38.33	42.02	39.72	42.26	40.88	40.70
w/ one-shot	73.43	71.53	72.96	56.25	45.57	56.52	38.77	42.85	40.29	43.45	41.66	41.13
w/ EmoDap	78.04	77.41	77.71	62.33	50.54	62.58	44.84	47.24	45.34	48.28	47.29	46.81

Table 10: More results on 4 datasets. The best results are in **bold**, while the second best are underlined.

Model	ED			GE			EI			EDOS		
	ACC.	F1	wF1	ACC.	F1	wF1	ACC.	F1	wF1	ACC.	F1	wF1
EmoDap _{Qwen3-8B}	76.45	76.37	76.46	59.63	51.10	61.62	45.79	46.57	45.99	49.37	48.29	47.98
w/ single-agent	73.02	72.21	73.86	57.99	47.31	58.96	42.76	42.07	42.65	46.99	43.98	42.83
w/o GRPO fine-tuning	73.89	71.99	73.99	57.48	50.08	59.65	43.45	42.42	43.32	46.94	46.19	41.86
w/o self-consistency	73.66	71.23	73.84	57.68	48.52	59.08	42.35	42.20	43.02	46.34	44.23	41.51
w/o K-means	73.30	73.00	72.93	58.12	49.13	60.09	42.84	42.72	42.13	47.62	44.88	42.62
w/o prototype	70.45	68.47	71.41	54.59	45.71	56.49	39.27	39.33	39.52	43.21	41.70	39.62
<i>Retrieval policy variants (Agent_{emo} frozen):</i>												
Agent _{demo} $\leftarrow$ adversarial [†]	68.86	68.53	67.92	52.83	44.86	53.11	37.29	38.07	38.74	41.28	40.95	41.05
Agent _{demo} $\leftarrow$ random	72.63	72.56	74.15	58.10	48.98	59.44	43.57	41.79	42.12	46.91	44.79	42.77
Agent _{demo} $\leftarrow$ centroid	74.02	72.00	73.11	57.06	50.27	59.31	42.76	42.43	43.45	47.40	45.43	42.63
Agent _{demo} $\leftarrow$ nearest-NN	74.30	71.93	74.26	58.41	49.24	60.30	42.80	43.53	42.48	46.72	46.20	41.59
Agent _{demo} $\leftarrow$ oracle*	78.12	77.61	78.21	61.27	52.63	63.06	47.25	48.08	47.30	50.69	49.54	49.45

Table 11: Ablation study on four datasets.

tor fixed, we replace the EmoDap-selected demonstration with one of four alternatives and measure the wF1 drop:

- **Same-class 2nd best / Same-class worst:** target-class demonstrations ranked 2nd / last by predictor confidence among the K candidates;
- **Random other-class:** uniformly drawn from a non-target emotion;
- **Adversarial other-class:** non-target demonstration with the highest cosine similarity to the query.

Table 12 shows that even substituting a same-class *but lower-quality* demonstration causes a non-trivial 2.5 wF1 drop, confirming Agent_{emo}’s intra-class selection is meaningful and not interchangeable with any same-label example.

Substitution	wF1	Δ vs. EmoDap
EmoDap (selected)	76.46	—
$\hookrightarrow$ same-class 2nd-best	75.71	−0.75
$\hookrightarrow$ same-class worst	73.92	−2.54
$\hookrightarrow$ random other-class	71.05	−5.41
$\hookrightarrow$ adversarial other-class	67.92	−8.54

Table 12: Demonstration substitution robustness on ED. Replacing EmoDap’s choice with a worse same-class centroid already costs 2.5+ wF1, confirming intra-class selectivity.

A.8 Confusion Mitigation Visualization

To provide a more comprehensive evaluation of our model’s performance, we present detailed confusion analysis visualizations in Figure 9. The results demonstrate a consistent and significant reduction in classification confusion across all emotion categories in the four benchmark datasets. Notably,

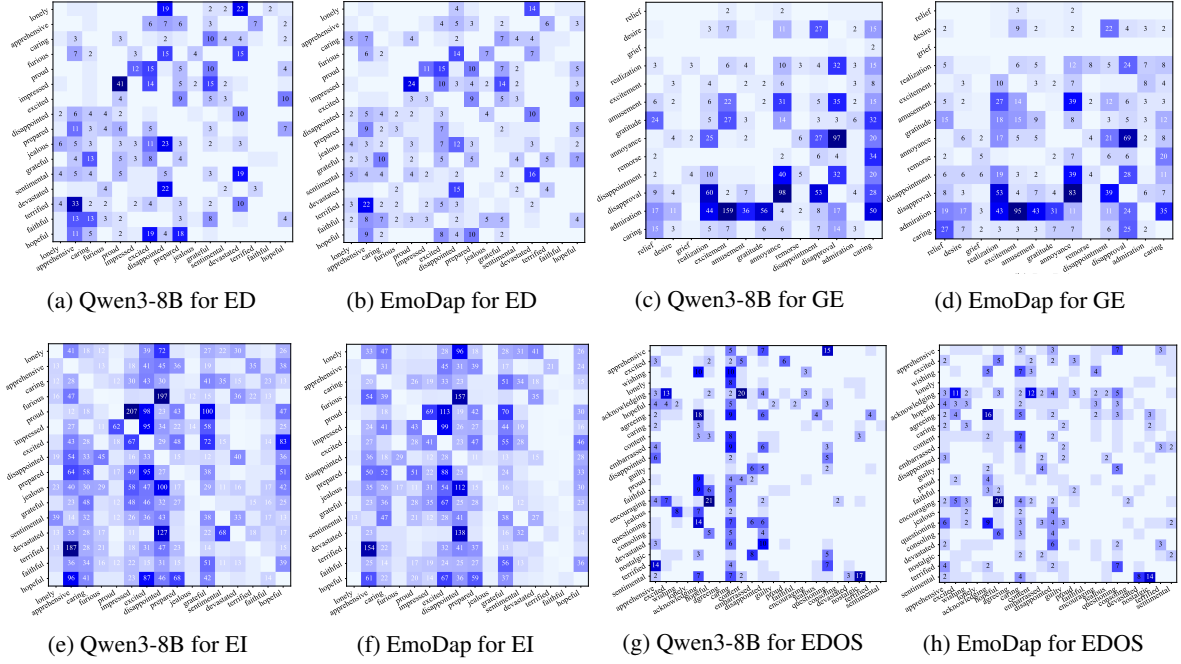


Figure 9: Confusion matrices on four datasets. Heatmap intensity shows confusion strength.

this improvement is not only universal (affecting all emotion classes) but also particularly pronounced for traditionally challenging emotion pairs that were previously prone to misclassification (e.g., *admiration* vs. *excitement* or *proud* vs. *impressed*). Further analysis reveals two key observations: (1) Dataset consistency: The improvements remain robust across all four datasets, demonstrating the generalizability of our method. (2) Hard case improvement: Previously ambiguous emotion pairs exhibit the most substantial reduction in confusion, suggesting that our approach effectively captures subtle discriminative features. This systematic reduction in confusion underscores the effectiveness of our proposed technique in enhancing fine-grained emotion recognition performance.

A.9 Per-Emotion Improvement Breakdown

We further investigate which emotions benefit most from EmoDap. For each unseen emotion $y \in \mathcal{Y}_t$ on ED, we compute (i) the per-class accuracy gain $\Delta\text{ACC}(y)$ of EmoDap over Qwen3-8B, and (ii) the average cosine distance from $\mathbf{v}_y$ to its nearest neighbor in the prototype space, $d_{\min}(y)$. The latter measures how isolated emotion y is in the prototype manifold, where larger values indicate cleaner separability.

Figure 10 reveals a clear positive correlation (Pearson $r = 0.67$, $p < 0.01$): emotions whose prototypes are well-separated from neighboring

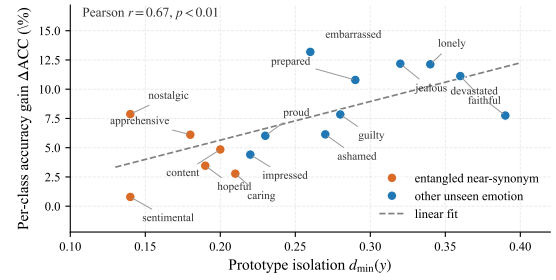


Figure 10: Per-emotion accuracy gain of EmoDap over Qwen3-8B on ED, plotted against the prototype-space isolation $d_{\min}$. Pearson $r = 0.67$, $p < 0.01$: emotions with cleaner prototype separation benefit more from EmoDap.

categories enjoy substantially larger gains, while geometrically entangled pairs (e.g., *sentimental*–*nostalgic*, $d_{\min}=0.14$) see smaller improvements. This finding has two implications: (i) prototype quality is the primary bottleneck of our framework’s residual error, providing an actionable direction for future work, e.g., richer prototype representations from multi-modal cues; and (ii) our gains are not uniform "free lunches" but principled improvements concentrated in the regions where prototype guidance is most informative.

A.10 Case Study

Figure 11 illustrates retrieval behavior on a nuanced instance (“Dad, what should I do? I love him so much it’s frightening.”). DeepSeek-V3’s

Emotions	Ashamed	Guilty
1.	I can't believe I lied to my mother.	I stole money from my friend.
2.	I felt really bad when I cheated on my math test.	I accidentally threw out my co-worker's lunch today at work. I felt so bad.
3.	I was ashamed when I decided to drop out of school.	Last week, I was unfaithful to my wife. I feel bad about it.
4.	I mistakenly stepped on someone without saying sorry, and she got furious; people attacked me, I felt so bad.	When I mistakenly hit someone into the gutter with a car that caused him to miss his interview.
	Annoyed	Furious
1.	I was irked when I saw someone put a dent in my door.	Someone cut me off in traffic today.
2.	There is a fly in my room and it is annoying the heck out of me!	When someone treated me badly and talked down to me.
3.	I am very irritated with people who want to make everyone else miserable when they are.	Someone rear-ended me and wrecked my car.
4.	My upstairs neighbor is making too much noise with her heels; sometimes she wakes me up with the noise.	I was watching my dad's friend's dog and when I left him in his dog pen, he jumped out and made a mess of my living room!

Table 13: Collected demonstrations for similar emotion pairs from the ED dataset.

Sentence: Dad, what should I do? I love him so much it's frightening.	
Top-4 emotions:	Deepseek-V3
1. **anxious** (at a loss because of loving too deeply)	
2. **terrified** (the intensity of love is described as "frightening")	
3. **hopeful** (implied in seeking guidance and expressing deep love)	
4. **sentimental** (the emotional weight of "I love him so much")	
Final prediction: anxious ❌	
Prototype recalled top-4 emotions: [terrified-0.51, embarrassed-0.41, apprehensive-0.37, anxious-0.15]	
↓	
Nearest retrieval:	w/ nearest demos.
• Joe? I don't think we should have a child. (<i>apprehensive</i>)	
• If it's not too late ... because I love you ... and I want to marry you. (<i>anxious</i>)	
Qwen3-8B prediction: anxious ❌	
Agent _{demo} retrieval:	Our EmoDap
• why don't you give me your number? I'm not sure that my fiancée would approve. (<i>apprehensive</i>)	
• It's so dark, I can't see anything. (<i>anxious</i>)	
Agent _{emo} prediction: apprehensive ✅	

Figure 11: Case study.

top-4 predictions omit the ground-truth (*apprehensive*), revealing limitations in capturing delicate emotional gradations. EmoDap successfully recalls the correct label by leveraging prototype similarity to filter candidates and retrieve contextually aligned demonstrations. In contrast, nearest-neighbor retrieval surfaces semantically proximate but emotionally mismatched examples, leading to incorrect predictions. This case confirms that agent-driven, preference-aligned retrieval substantially outperforms static similarity matching in zero-shot FER.

A.11 Failure Case Analysis

We complement the success case in Appendix A.10 with a failure analysis based on 100 randomly sampled errors of EmoDap on ED. Errors cluster into three patterns:

(i) Prototype collapse on near-synonym pairs

(43%). When two emotions share highly overlapping definitions (e.g., *sentimental* vs. *nostalgic*), their prototypes are insufficiently separated, and the candidate set $\mathcal{Y}_{\text{cand}}$ retains both, leaving the predictor to disambiguate without further geometric guidance. This is the dominant residual error and aligns with the per-emotion analysis in Appendix A.9.

(ii) Sarcasm and figurative language (29%).

Sentences expressing one emotion through the surface form of another (e.g., ironic praise) defeat both the prototype filter and the retriever, because the demonstration pool, built from sincere expressions, contains no aligned supervision. Multi-modal cues or a sarcasm-aware reranker would be needed.

(iii) Over-aggressive candidate filtering (17%).

When ρ is set too low, the GT emotion is occasionally pruned out of $\mathcal{Y}_{\text{cand}}$ (a hard ceiling that no downstream stage can recover from). The remaining 11% are noisy ground-truth labels confirmed by manual inspection. These observations directly motivate our Limitations section and inform several future directions, including richer prototype representations and adaptive ρ scheduling.

A.12 Demonstration Pool Quality

To investigate the quality of the demonstration pool, we present in Table 13 four examples for similar emotion pairs from the ED dataset. These examples were obtained by clustering the unlabeled corpora. As the most representative samples at the cluster centers, they show rich diversity in content and variation in length, thus effectively covering queries of different contexts and sizes.