

UniMER: Multilingual Multimodal Emotion Recognition with Prototype-Guided Attention Pooling

Yinghua Qin*, David C. Anastasiu

Department of Computer Science and Engineering
Santa Clara University, CA, USA

Abstract

Conversational emotion recognition suffers from two compounding limitations: models trained on imbalanced data report near-zero F1 on minority emotions, and evaluation has been almost exclusively English, leaving it unknown whether current methods generalize across languages. We propose UniMER, a parameter-efficient framework built on the insight that natural language is the universal interface: visual, acoustic, and facial cues are converted into text by a multimodal LLM, enabling a single LoRA-adapted backbone to process all modalities and languages through one shared representation. To prevent minority-class collapse under extreme imbalance, we introduce class-prototype attention pooling, which anchors each attention head to a learnable class prototype so every emotion retains dedicated representational capacity. With only 2.4% trainable parameters, UniMER achieves state-of-the-art results on three benchmarks across three language families: MELD (English, +14.0pp macro F1), M3ED (Chinese, +3.2pp wF1, text-only), and AkaCE (Akan, +11.5pp macro F1), lifting AkaCE Fear and Disgust F1 from near-zero (<0.1%) to 58% and 40%. All three benchmarks share the 7-class Ekman taxonomy, enabling direct cross-dataset transfer evaluation. Cross-lingual analysis reveals universal arousal-based confusions alongside culturally specific patterns: heightened neutral absorption in Chinese and prosodic confusion in Akan absent in English.¹

1 Introduction

Emotion recognition in conversations has advanced rapidly, yet almost exclusively in English and almost exclusively on common emotions. On benchmarks like MELD (Poria et al., 2019), state-of-the-art methods routinely report 65–70% overall

accuracy while yielding *near-zero* F1 on minority classes like Fear and Disgust. These aggregate metrics obscure a fundamental failure: existing models learn to predict the majority class while neglecting minority categories. This is not merely a measurement artifact. Emotions that are rare in data (fear, disgust, surprise) are precisely those most consequential in applications like mental health monitoring and crisis detection.

Compounding this, the overwhelming focus on English has left open whether emotion recognition generalizes across languages at all. Emotions exhibit complex interactions between universal and culture-specific manifestations (Jack et al., 2012; Zhao et al., 2022): indirect emotional expression in Chinese (Tsai et al., 2007), prosodic emphasis patterns in Akan (Sasu et al., 2025), and differing politeness conventions all shape how the same underlying affect surfaces in conversation. Without multilingual evaluation, it remains unknown whether our models capture emotion or merely English-specific surface patterns.

Meanwhile, the standard approach to multimodal emotion recognition, which learns modality-specific encoders and aligns their embedding spaces through fusion mechanisms (Mao et al., 2021; Zou et al., 2022; Huang et al., 2024), creates engineering complexity that scales poorly to new modalities and new languages. Every new combination requires its own fusion architecture, making it difficult to build systems that are simultaneously multimodal, multilingual, and efficient.

We propose a simple resolution to all three problems. Our key insight is that *natural language is the universal interface*: if a multimodal LLM converts visual, acoustic, and facial cues into descriptive text (e.g., “trembling voice with rising pitch, furrowed brows”), then a single text-based backbone can process all modalities and all languages through one shared representation, eliminating modality-specific encoders, fusion layers, and language-

*Currently at Microsoft, Redmond, USA

¹Code available at <https://github.com/davidanastasiu/unimer>.

specific architectures. This LLM-guided captioning approach leverages the rich world knowledge of instruction-tuned models for multimodal interpretation, produces human-readable intermediate representations, and reduces multimodal fusion to string concatenation.

Building on this insight, we present UniMER (Unified Multilingual Emotion Recognition), a parameter-efficient framework that combines LLM-guided multimodal captioning with LoRA-adapted classification under class-balanced training. A single frozen Gemma-3-12B backbone with per-dataset LoRA adapters (2.4% trainable parameters) handles English, Chinese, and Akan, including AkaCE (Sasu et al., 2025), the first multimodal emotion dataset for an African language. We focus on three benchmarks that share the standard 7-class Ekman taxonomy (Ekman, 1992) (neutral, anger, disgust, fear, joy, sadness, surprise), enabling direct cross-dataset transfer evaluation across languages. Imbalance ratios range from 11:1 to 25:1; we report means over five random seeds alongside best-seed results.

Our contributions are:

- We demonstrate that natural language captioning of multimodal signals, combined with LoRA-adapted classification, achieves state-of-the-art macro-F1 and rare-class performance under extreme imbalance across three benchmarks and three language families: MELD (71.1% accuracy, 58.7% macro F1), M3ED (56.9% accuracy, 56.1% weighted F1, text-only), and AkaCE (59.3% accuracy, 45.0% macro F1), with $3.7\text{--}4.5\times$ improvements on MELD rare emotions (Fear, Disgust) and lifting AkaCE Fear and Disgust F1 from near-zero ($<0.1\%$) to 58% and 40%, respectively.
- We propose class-prototype attention pooling, where each emotion class conditions a dedicated attention head over LLM token representations, preventing head collapse under extreme imbalance. Prototype pooling improves macro F1 on all three benchmarks (AkaCE +4.0pp, MELD +1.4pp, M3ED +0.4pp), with gains scaling with imbalance severity.
- We show that LLM-guided multimodal captioning provides substantial gains over text-only classification (+14.1pp macro F1 on

AkaCE, +2.7pp on MELD), offering an interpretable and language-agnostic alternative that avoids the engineering complexity of modality-specific encoders and fusion architectures.

- Through cross-lingual analysis spanning Indo-European, Sino-Tibetan, and Niger-Congo language families, we identify both universal patterns (arousal-based confusions across all languages) and culturally specific ones (heightened neutral absorption in Chinese, prosodic confusion in Akan), providing empirically grounded guidance for culturally-aware system design.

2 Related Work

Following Wu et al. (2025a), conversational emotion recognition methods can be categorized into graph-based, fusion-based, and LLM-based approaches.

Graph-Based Methods. These methods model conversational structure through GNNs (Ghosal et al., 2019; Ren et al., 2022; Ghosal et al., 2020; Li et al., 2021; Meng et al., 2024). Despite strong aggregate performance, these methods consistently report near-zero F1 on minority emotions.

Fusion-Based Methods. Fusion-based methods integrate textual, acoustic, and visual signals through cross-modal mechanisms (Mao et al., 2021; Li et al., 2022; Zou et al., 2022; Huang et al., 2024; Ma et al., 2024a,b; Xie et al., 2025), but generally do not address class imbalance.

LLM-Based Approaches. Recent instruction-tuned methods frame emotion recognition as generation (InstructERC (Lei et al., 2023), DialogueLLM (Zhang et al., 2025)), while multimodal extensions (BeMERC (Fu et al., 2025), MERITS-L (Dutta and Ganapathy, 2025)) integrate visual cues. Emotion-LLaMA (Cheng et al., 2024) instruction-tunes a multimodal LLM but targets Chinese benchmarks with different emotion taxonomies, precluding direct comparison (see Section 4.3). An emerging trend uses LLMs as multimodal interpreters, converting non-textual signals into text; our work adopts this captioning paradigm with LoRA-based classification under extreme class imbalance.

Prototype-Based Emotion Recognition. Prior prototype-based methods such as ProEmoTrans (Peng et al., 2025) use semantic prototypes as label-description embeddings, matching pooled

representations to prototypes via cosine similarity for zero-shot recognition of unseen emotions where training and test label sets are disjoint. Our class-prototype attention pooling (Section 3.2.3) differs in both task and mechanism: each prototype p_k is a learned attention query that conditions attention over LLM token representations during pooling, giving every class—including rare ones—a dedicated attention head under a single fixed taxonomy shared between training and testing. This directly addresses majority-class collapse under extreme class imbalance.

Multilingual ERC. Most ERC research targets English. M3ED (Zhao et al., 2022) introduced the first large-scale Chinese multimodal emotion dataset, and AkaCE (Sasu et al., 2025) the first for an African language (Akan). We evaluate a common architecture across all three languages.

3 Methodology

We present UniMER, a parameter-efficient framework for multilingual multimodal conversational emotion recognition under extreme class imbalance. We formalize the problem (Section 3.1), detail our architecture (Section 3.2), and describe the loss formulation (Section 3.3).

3.1 Problem Formulation

Given a conversational dataset $\mathcal{D} = \{(x_i, v_i, h_i, y_i, \ell_i)\}_{i=1}^N$ with utterance x_i in language $\ell_i \in \{\text{en, zh, ak}\}$, optional video v_i , context h_i (previous k utterances), and emotion label $y_i \in \{1, \dots, K\}$. UniMER operates in two stages: (1) **Preprocessing**: when video is available, a multimodal LLM synthesizes audio, visual, and facial AU evidence into a natural language summary $s_i = \text{Caption}(v_i)$; when video is unavailable, s_i is omitted and the model operates on text alone; (2) **Classification**: a LoRA-adapted LLM learns $f_\theta : (x, s, h, \ell) \rightarrow y$.

3.2 Architecture Overview

Our architecture comprises four key components: (1) multimodal feature extraction, (2) LoRA-adapted base model, (3) class-prototype attention pooling, and (4) classification head. Figure 1 illustrates the complete end-to-end system.

3.2.1 Data Preprocessing and LLM-Guided Multimodal Captioning

Rather than employing modality-specific encoders with late or early fusion, we use a dedicated multi-

Algorithm 1 LLM-Guided Multimodal Captioning

Require: Video V with frames $\{f_1, \dots, f_n\}$, audio A
Ensure: Multimodal summary S in natural language
1: Extract AUs per frame; detect peak frame f^*
2: $D_{\text{audio}} \leftarrow \text{LLM}(A, p_{\text{audio}})$ {Tone, pitch, pace}
3: $D_{\text{image}} \leftarrow \text{LLM}(f^*, p_{\text{image}})$ {Peak frame}
4: $D_{\text{video}} \leftarrow \text{LLM}(V, p_{\text{video}})$ {Scene dynamics}
5: $S \leftarrow \text{LLM}(D_{\text{audio}}, D_{\text{image}}, D_{\text{video}}, \text{AU_emo}; p_{\text{synth}})$
6: **return** S

Utterance (MELD, Fear): “Look, we have enough, just walk away.”

Peak Frame & LLM-Generated Multimodal Caption:



Visual: Five people with expressions of concern or surprise. The overall mood is one of mild tension or uncertainty.

AUs: Lid tightener (3.27), Cheek raiser (1.75), Nose wrinkler (1.62), Upper lip raiser (1.63), Lip corner puller (1.46).

Prediction: Fear ✓ **Text-only prediction:** Neutral ✗

Table 1: Example of LLM-guided multimodal captioning on a MELD utterance (dia29_utt6). The text alone reads as calm advice; the peak frame and caption reveal critical disambiguating cues: a very high lid tightener (3.27) indicates strong apprehension, while the group’s concerned expressions and tense atmosphere signal underlying anxiety, correctly resolving the rare *Fear* emotion.

modal LLM (Gemma-3n-E4B-it) to independently analyze each modality, including audio (vocal tone, pitch, pace), visual (facial expressions from peak emotional frames identified via OpenFace AU detection (Baltrušaitis et al., 2018)), and video (scene dynamics), and synthesize the results into a unified natural language caption. This approach preserves fine-grained emotional cues in text form (e.g., “trembling voice with rising pitch”), avoids aligning heterogeneous embedding spaces, and produces human-readable intermediate representations. Ablation studies confirm consistent gains from this captioning approach: +14.1pp macro F1 on AkaCE and +2.7pp on MELD (Table 7). The complete captioning procedure is summarized in Algorithm 1; full details including AU-to-emotion mappings and modality-specific prompts are provided in Appendix B.

For each utterance, we construct a unified input sequence:

$$\text{Input} = [\text{CTX}] \text{ Context } [\text{UTT}] \text{ Utterance } [\text{MMS}] \text{ Summary } [\text{SEP}], \quad (1)$$

where Context includes the previous k utterances and Summary is the LLM-generated multimodal description. No explicit natural-language instruction is provided.

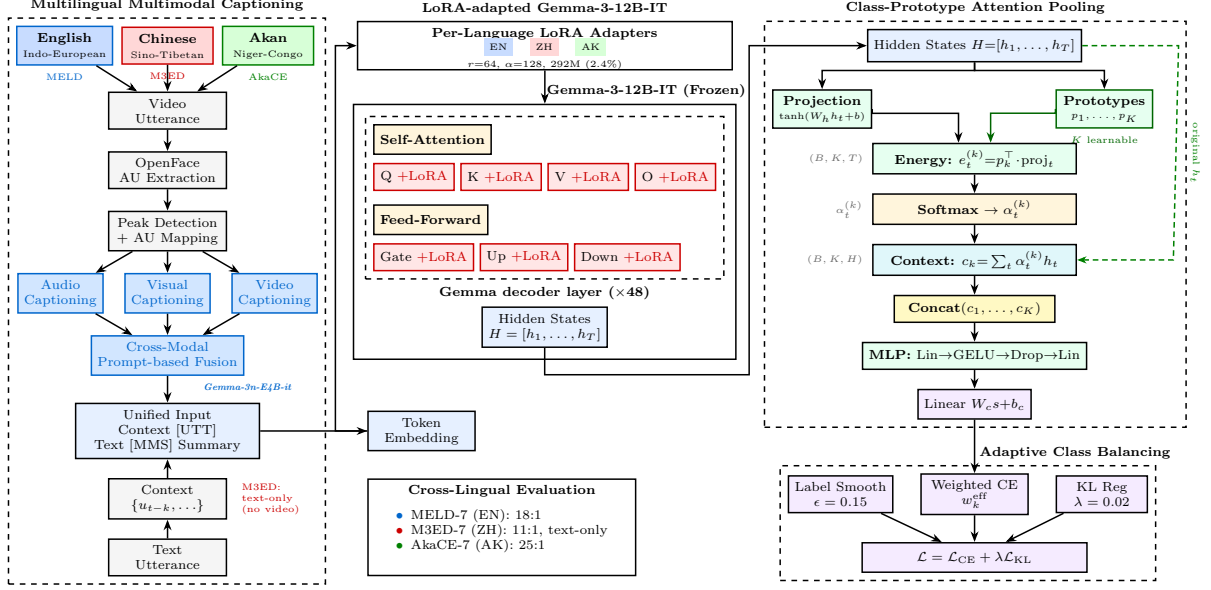


Figure 1: UniMER system architecture. **Left:** LLM-guided multimodal captioning (Algorithm 1). OpenFace extracts facial Action Units and identifies peak emotional frames. Gemma-3n-E4B-it independently analyzes audio, visual, and video modalities, then synthesizes all evidence into a unified natural language summary. **Middle:** Instruction-tuned Gemma-3-12B processes the unified multilingual input (context, transcript, and multimodal summary) with LoRA adaptation (rank=64, 292M trainable params, 2.4% of base model) across all 48 transformer blocks. **Right:** Class-prototype attention pooling (Algorithm 2), where K learnable prototype vectors (one per emotion class) produce class-conditional context representations, followed by a classification head with class-balanced loss across three evaluation settings.

3.2.2 Base Model and LoRA Adaptation

We employ Gemma-3-12B-it (Team et al., 2025) as our base LLM with LoRA (Hu et al., 2022a) adaptation:

$$W = W_0 + \Delta W = W_0 + BA, \quad (2)$$

where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and $r \ll \min(d, k)$. We apply LoRA to all attention projections (Q, K, V, O) and FFN layers (gate, up, down) across all 48 transformer layers. For each dataset, we train a separate set of LoRA adapters on the shared frozen base, enabling language-specific adaptation while maintaining a common architecture. With rank $r=64$ and $\alpha=128$, this yields 292M trainable parameters (2.4% of the base model).

3.2.3 Class-Prototype Attention Pooling

Standard multi-head attention pooling uses generic heads that attend to emotionally salient tokens, but under extreme class imbalance, all heads converge to features of majority emotions, leaving minority classes with weak representations. We propose *class-prototype attention pooling*, where each attention head is explicitly conditioned on a learnable class prototype embedding, ensuring that every

emotion, including rare ones, receives a dedicated, specialized representation.

Given token representations $H = [h_1, h_2, \dots, h_T]$ from the LLM, we maintain a set of K learnable prototype vectors $\{p_1, \dots, p_K\}$ (one per emotion class), each $p_k \in \mathbb{R}^d$. Each prototype defines a class-specific attention head:

$$e_t^{(k)} = p_k^\top \tanh(W_h h_t + b_h), \quad (3)$$

$$\alpha_t^{(k)} = \frac{\exp(e_t^{(k)})}{\sum_{j=1}^T \exp(e_j^{(k)})}, \quad (4)$$

$$c_k = \sum_{t=1}^T \alpha_t^{(k)} h_t, \quad (5)$$

where the attention energy $e_t^{(k)}$ scores the relevance of token h_t to emotion class k , $\alpha_t^{(k)}$ are the class-conditional attention weights, and c_k is the class-specific context vector. The K context vectors are concatenated and projected to produce the final pooled representation:

$$s = \text{Proj}(\text{Concat}[c_1, c_2, \dots, c_K]), \quad (6)$$

where the projection is a two-layer MLP with GELU activation and dropout. Each attention head

Algorithm 2 UniMER End-to-End Classification

Require: Utterance x , video V , context $c = \{u_{t-k}, \dots, u_{t-1}\}$
Ensure: Predicted emotion $\hat{y}$
1: $S \leftarrow \text{MULTIMODALCAPTION}(V)$ {Alg. 1}
2: Input $\leftarrow [\text{CTX}]_c [\text{UTT}]_x [\text{MMS}]_S [\text{SEP}]$
3: $\mathbf{t} \leftarrow \text{Tokenize}(\text{Input})$
4: $H = [h_1, \dots, h_T] \leftarrow \text{LLM}_{\text{LoRA}}(\mathbf{t})$ {Frozen + LoRA}
 {Prototype Pooling}
5: **for** $k = 1$ **to** K **do**
6: $\alpha_t^{(k)} = \text{softmax}(p_k^\top \tanh(W_h h_t + b_h))$
7: $c_k = \sum_t \alpha_t^{(k)} h_t$
8: **end for**
9: $s \leftarrow \text{Proj}(\text{Concat}[c_1, \dots, c_K])$
10: $\hat{y} \leftarrow \arg \max_k \text{softmax}(W_c s + b_c)$
11: **return** $\hat{y}$

is *anchored* to a specific emotion class, preventing head collapse under imbalanced training. Ablation studies (Section 4.4) confirm improvements on all three benchmarks, with gains scaling with imbalance severity.

3.2.4 Classification Head

The pooled representation s produces class logits via $p(y \mid x) = \text{softmax}(W_c s + b_c)$, where $W_c \in \mathbb{R}^{K \times d}$. The classification head and prototype vectors are trained from scratch while the base LLM remains frozen except for LoRA adapters. Algorithm 2 summarizes the complete end-to-end pipeline.

3.3 Adaptive Class Balancing and Loss Formulation

We adopt a weighted cross-entropy objective with KL-based regularization:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{CE}}(\theta) + \lambda \mathcal{L}_{\text{KL}}(\theta), \quad (7)$$

where $\mathcal{L}_{\text{CE}} = -\sum_{i=1}^N w_{y_i} \log p_\theta(y_i \mid x_i)$ is the class-weighted cross-entropy with class weights w_k following the effective number of samples scheme (Cui et al., 2019), $w_k^{\text{eff}} = (1 - \beta)/(1 - \beta^{n_k})$ with $\beta = 0.9999$, and $\mathcal{L}_{\text{KL}} = \mathbb{E}_x[D_{\text{KL}}(p_\theta \parallel \text{Uniform})]$ encourages distributional diversity. The KL term ($\lambda=0.02$) prevents majority-class collapse, and label smoothing ($\epsilon=0.15$) reduces overconfidence. Both primarily benefit low-resource datasets.

4 Experiments

We evaluated UniMER on three datasets spanning three languages with varying imbalance (MELD: 18:1, M3ED: 11:1, AkaCE: 25:1), comparing against state-of-the-art methods and conducting

comprehensive ablations on LoRA configuration, class weighting, prototype pooling, and loss formulation.

4.1 Datasets and Evaluation

We evaluated on three benchmark datasets across three languages, all sharing the 7-class Ekman emotion taxonomy (Table 2). This unified label space enables direct cross-dataset transfer evaluation (Section 4.5).

MELD (Poria et al., 2019) contains 13,708 English utterances from *Friends* with 7 emotions.

M3ED (Zhao et al., 2022) is the largest Chinese multimodal emotion dataset with 24,449 utterances from 990 dialogues (mean length 24.7 utterances). We split chronologically: 70%/15%/15%.

AkaCE (Sasu et al., 2025) is the first multimodal emotion dataset for an African language (Akan/Twi), containing 6,162 utterances from 385 dialogues with word-level prosodic annotations. After removing 596 invalid utterances, we use 3,908/821/837 train/dev/test samples.

We report accuracy, weighted F1, and macro F1, using published results for baselines.

4.2 Implementation Details

We used Gemma-3-12B-it as base model with LoRA ($r = 64$, $\alpha = 128$) applied to all attention and FFN layers across all datasets. Multimodal descriptions were generated using Gemma-3n-E4B-it for MELD and AkaCE; M3ED experiments used text-only input because the dataset’s video files are not publicly distributed due to copyright restrictions on the source TV series (Zhao et al., 2022). Context length was 40 utterances for MELD, 20 for M3ED, and 2 for AkaCE.

Training used AdamW optimizer with learning rate 5×10^{-5} , cosine schedule with warmup, label smoothing $\epsilon = 0.15$, and effective-number class weighting. Effective batch sizes were tuned per dataset: 48 for MELD (1 epoch), and 24 for M3ED and AkaCE (3 epochs each). Early stopping with patience=3 on a single A100 80 GB GPU (peak memory ≈ 50 GB, training completes in 1.6–2.8 hours per dataset). We report the mean and standard deviation across five random seeds (42, 123, 456, 789, 2024); per-class breakdowns in SOTA comparison tables are from the best-performing seed, as noted in table captions. We report accuracy, weighted F1, and macro F1 for all datasets to provide a complete picture of overall, distribution-weighted, and per-class performance.

MELD (7, EN, 18:1)						M3ED (7, ZH, 11:1)						AkaCE (7, AK, 25:1)					
Emotion	Train	Dev	Test	Total	%	Emotion	Train	Dev	Test	Total	%	Emotion	Train	Dev	Test	Total	%
Neutral	4710	470	1256	6436	47.0	Neutral	6821	1462	1463	9746	39.9	Neutral	1945	452	411	2808	50.4
Joy	1743	163	402	2308	16.8	Anger	2576	552	553	3681	15.1	Anger	614	113	177	904	16.2
Surprise	1205	150	281	1636	11.9	Happy	2199	471	472	3142	12.9	Sadness	570	58	128	756	13.6
Anger	1109	153	345	1607	11.7	Sad	2028	435	435	2898	11.9	Joy	390	73	68	531	9.5
<u>Sadness</u>	683	111	208	1002	7.3	Surprise	1542	331	331	2204	9.0	<u>Surprise</u>	210	95	26	331	5.9
<u>Disgust</u>	271	22	68	361	2.6	<u>Disgust</u>	1299	278	279	1856	7.6	<u>Fear</u>	93	15	16	124	2.2
<u>Fear</u>	268	40	50	358	2.6	<u>Fear</u>	649	138	135	922	3.8	<u>Disgust</u>	86	15	11	112	2.0
Total	9989	1109	2610	13708		Total	17114	3667	3668	24449		Total	3908	821	837	5566	

Table 2: Dataset statistics for English (MELD), Chinese (M3ED), and Akan (AkaCE) benchmarks. All three datasets share the 7-class Ekman taxonomy, enabling direct cross-dataset transfer. Headers show (classes, language, imbalance ratio $N_{\max}/N_{\min}$). Minority classes are underlined.

Inference Efficiency. UniMER predicts each label in a single forward pass: prototype pooling converts the LLM hidden states into a fixed-size representation, and a softmax classification head directly outputs the label. No tokens are generated during inference. This contrasts with generation-based ERC methods which autoregressively decode labels or explanations, incurring latency that scales with output length. Multimodal captioning is performed once offline per utterance and is not required during prediction.

4.3 Main Results

On MELD (Table 3), UniMER achieves $70.7 \pm 0.4\%$ accuracy, $70.3 \pm 0.5\%$ weighted F1, and $57.3 \pm 1.3\%$ macro F1 (5-seed means), exceeding the strongest graph-based method on accuracy and macro F1. For rare emotions (Figure 2), Disgust and Fear F1 scores increase from 10.7% to 48.1% and from 10.8% to 39.6%, respectively, mitigating previously near-zero performance. Performance gains are also observed for common emotions, including Neutral (81.8%), Surprise (62.0%), and Joy (67.8%).

On M3ED (Table 4), using text-only input due to video licensing restrictions (Section 4.2), UniMER achieves $56.3 \pm 0.7\%$ accuracy, $45.9 \pm 0.3\%$ macro F1, and $55.5 \pm 0.4\%$ weighted F1. Compared to the recent MERC-PLTAF (Wu et al., 2025b), our 5-seed mean (56.3%) improves over MERC-PLTAF (53.6%) by +2.7pp accuracy. UniMER also surpasses the best multimodal baselines from the original M3ED paper (Zhao et al., 2022): DialogueRNN achieves 52.3% weighted F1 with text+visual fusion, and our 5-seed mean exceeds this by +3.2pp (55.5% vs. 52.3%) using text alone. For rare emotions (Figure 2, center), Disgust F1 reaches 36.4% and Fear 15.7%, notably lower than MELD and AkaCE, reflecting both the text-only limitation and

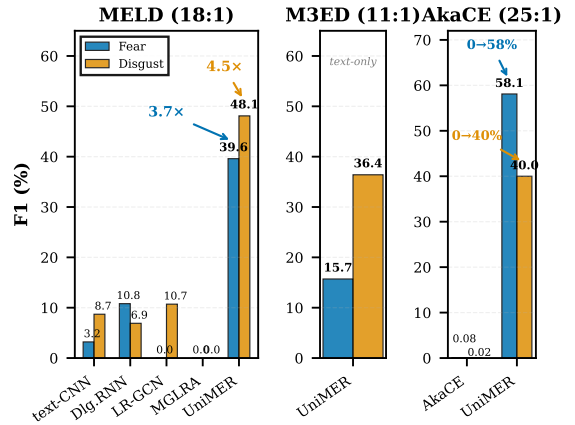


Figure 2: Rare emotion F1 comparison across three datasets and languages. **Left:** On MELD (English), UniMER achieves 3.7 $\times$ (Fear) and 4.5 $\times$ (Disgust) improvements over the best prior method. **Center:** On M3ED (Chinese, text-only), UniMER reaches 15.7% Fear and 36.4% Disgust F1 despite 11:1 imbalance and no multimodal input. **Right:** On AkaCE (Akan), prior methods report near-zero F1 ($<0.1\%$) on Fear and Disgust; UniMER lifts these to 58.1% and 40.0%, respectively.

M3ED’s 11:1 imbalance ratio.

On AkaCE (Table 5), UniMER achieves $57.2 \pm 1.1\%$ accuracy, $55.3 \pm 1.4\%$ weighted F1, and $43.2 \pm 1.5\%$ macro F1. This macro F1 surpasses the multimodal Transformer Fusion baseline (Sasu et al., 2025) (31.7%) by +11.5pp, with the best-seed weighted F1 (57.7%) exceeding the strongest baseline (56.1%). Per-class analysis (Table 5) shows UniMER substantially improves rare-emotion recognition: Fear F1 rises from near-zero (0.08%) to 58.1%, Disgust from 0.02% to 40.0% (Figure 2, right). We note that AkaCE’s rare classes have very small test sets (Disgust: 11 samples, Fear: 16, Surprise: 26), so per-class metrics have high variance across seeds (macro F1 std= ± 1.5 pp);

the reported per-class numbers are from the best-performing seed and should be interpreted cautiously. The consistent improvement pattern across all rare classes, combined with the +11.5pp gain in macro F1 (5-seed mean), supports the effectiveness of our class-balancing and prototype pooling approach.

Beyond graph-based and fusion-based baselines, we also compare UniMER against supervised fine-tuning of smaller language models on MELD (Table 6). UniMER’s weighted F1 (70.7%) exceeds all SFT baselines by ≥ 4.7 pp, surpassing MERITSL (Dutta and Ganapathy, 2025) (66.0%). Although UniMER uses more trainable parameters (292M), these comprise only 2.4% of the base model, and ablations confirm the gains stem from multimodal captioning and class balancing rather than scale alone. On MELD, DialogueLLM (Zhang et al., 2025) reports 71.9% weighted F1 and SpeechCueLLM (Wu et al., 2025c) 67.6%, compared with UniMER’s 70.7%; however, neither reports per-class or macro F1, making rare-emotion performance unclear. UniMER’s class-prototype pooling and adaptive loss formulation explicitly target rare classes, yielding substantial improvements on Disgust and Fear (Figure 2, left), which are not reported in the LLM-based baselines.

4.4 Ablation Studies

We conducted systematic ablations on three core design choices: class-prototype pooling (Proto Pool), LoRA-based supervised fine-tuning (SFT), and LLM-guided multimodal captioning (LMCF). Table 7 reports results across all three datasets. The “All” row represents the full system with class-prototype pooling; the “w/o Proto Pool” row uses standard multi-head pooling, and subsequent component ablations remove one component at a time from this base configuration. Removing prototype pooling causes the largest accuracy drop on AkaCE (−3.7pp), with impact scaling with imbalance severity. LoRA fine-tuning produces the largest drops on MELD and M3ED, while LMCF is most impactful on AkaCE (−14.1pp mF1), where prosodic and visual cues complement limited text, and contributes moderately to MELD (−2.7pp mF1). M3ED experiments use text-only input, so the LMCF ablation is not applicable.

Key Findings Prototype pooling is the most impactful addition on AkaCE (−3.7pp Acc, −4.0pp mF1 when removed), with gains scaling with im-

balance severity (AkaCE 25:1 > MELD 18:1 > M3ED 11:1). LoRA fine-tuning produces the largest component drops (−3.9pp accuracy and −7.6pp mF1 on MELD when removed). LLM-guided captioning is most impactful on AkaCE (−14.1pp mF1), where visual and prosodic cues are essential for the low-resource setting, and benefits MELD moderately (−2.7pp mF1).

4.5 Cross-Dataset Generalization

We evaluate zero-shot cross-dataset transfer among the three datasets sharing the 7-class Ekman taxonomy (Table 8). Transfer quality varies with language distance: M3ED→MELD achieves 52.3% accuracy with 37.4% mF1, suggesting partial cross-lingual transfer, while MELD→AkaCE achieves 52.2% accuracy but only 23.3% mF1, indicating majority-class prediction. Akan-trained models reach 46.7–49.7% accuracy on English and Chinese benchmarks despite only 3.9k training samples. These results confirm that Gemma-3’s multilingual representations enable some cross-lingual emotion transfer, but quality depends on language typological distance.

4.6 Cross-Lingual Emotion Patterns

Our multilingual experiments reveal both universal and culturally-specific patterns (see Section 4.7 for detailed confusion matrices). Universal patterns include consistent arousal-based confusions (anger/disgust, joy/surprise) across all languages.

Language-specific differences are substantial. Chinese M3ED exhibits significantly higher neutral absorption than English MELD (mean rate 40% vs. 22%), with the largest gaps on Happy→Neutral (56% vs. 18%) and Fear→Neutral (53% vs. 32%), consistent with cultural emotional restraint (Tsai et al., 2007). AkaCE (Akan) shows a qualitatively different pattern: rather than suppressing emotions toward neutral, it exhibits arousal-based confusion—Fear↔Disgust (25%/27%) absent in English and Chinese—suggesting Akan emotion boundaries are organized by arousal rather than valence.

Context sensitivity also varies across languages: M3ED benefits most from longer context windows, MELD shows limited sensitivity, and AkaCE is optimal at short context (context=2), suggesting that optimal context windows should be adapted based on conversational structure rather than using a fixed value. Appendix A isolates culturally-specific neutral absorption patterns via per-emotion deviations

Methods	Neutral		Surprise		Fear		Sadness		Joy		Disgust		Anger		Overall		
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	wF1	mF1
text-CNN (Kim, 2014)	74.8	73.2	45.4	44.8	3.8	3.2	21.4	22.3	49.3	48.9	8.9	8.7	35.3	35.1	55.3	54.9	33.7
bc-LSTM (Poria et al., 2017)	77.0	76.2	25.1	24.3	9.2	8.7	24.4	23.5	54.6	54.0	4.4	4.2	43.7	42.8	59.3	58.3	33.4
MFN (Zadeh et al., 2018)	76.4	76.2	40.9	39.5	0.0	0.0	14.1	13.7	46.7	45.8	0.0	0.0	40.9	39.6	55.1	54.2	30.7
DialogueRNN (Majumder et al., 2019)	77.4	78.2	52.8	51.9	11.9	10.8	34.5	33.8	54.4	52.9	7.8	6.9	43.7	42.5	60.1	59.4	39.6
LR-GCN (Ren et al., 2022)	79.8	80.3	57.5	57.0	0.0	0.0	<u>37.2</u>	<u>36.3</u>	66.1	65.2	10.9	<u>10.7</u>	<u>54.7</u>	<u>53.9</u>	65.9	<u>65.3</u>	<u>43.3</u>
MGLRA (Meng et al., 2024)	78.2	80.8	59.8	59.5	0.0	0.0	30.8	27.8	<u>68.5</u>	<u>66.5</u>	0.0	0.0	43.6	48.4	<u>66.4</u>	64.9	40.4
UniMER	83.2	81.8	67.3	62.0	36.0	39.6	40.4	47.7	68.9	67.8	45.6	48.1	61.4	63.8	71.1	70.7	58.7

Table 3: Comparison with graph- and fusion-based methods on MELD. UniMER achieves 71.1% accuracy, 70.7% weighted F1, and 58.7% macro F1 (best of 5 seeds; mean±std: 70.7±0.4% / 70.3±0.5% / 57.3±1.3%). Best in bold, second underlined.

Method	Acc	mF1	wF1
<i>Dialogue Models</i>			
DialogueGCN (Ghosal et al., 2019)	—	—	49.9
DialogueRNN (Majumder et al., 2019)	—	—	<u>52.3</u>
MMGCN (Hu et al., 2021)	—	—	51.2
UttBaseline (Zhao et al., 2022)	—	—	48.9
MDI (Zhao et al., 2022)	—	—	52.1
<i>Speech / LLM Methods</i>			
Emotion2vec (Ma et al., 2024c)	49.2	—	44.4
MERC-PLTAF (Wu et al., 2025b)	<u>53.6</u>	—	51.4
UniMER	56.9	46.2	56.1

Table 4: State-of-the-art comparison on M3ED (Chinese). Baselines report only aggregate metrics; per-class F1 is not available in the original publications. UniMER achieves 56.9% accuracy (+3.3pp over MERC-PLTAF) and 56.1% wF1 (+3.8pp over DialogueRNN), using text-only input. UniMER per-class F1 (best seed): Neu 63.0, Ang 65.5, Sur 57.8, Sad 46.4, Hap 38.6, Dis 36.4, Fea 15.7. 5-seed mean±std: 56.3±0.7% / 45.9±0.3% / 55.5±0.4%.

Method	Overall			Per-class F1						
	Acc	mF1	wF1	Neu	Sur	Fea	Sad	Joy	Dis	Ang
AkaCE (Sasu et al., 2025)	—	31.7	56.1	68.6	15.2	0.1	56.5	25.3	0.0	51.5
UniMER	59.3	45.0	57.7	70.0	14.6	58.1	49.1	29.4	40.0	53.6

Table 5: State-of-the-art comparison on AkaCE (Akan). The AkaCE baseline is the best-performing multi-modal Transformer Fusion (Text+Audio+Video) model from Sasu et al. (2025). UniMER achieves 45.0% macro F1 (+13.3pp over baseline). Fear F1 rises from 0.08% to 58.1%, Disgust from 0.02% to 40.0%. Small test sets for rare classes (Disgust: 11, Fear: 16, Surprise: 26) yield high per-class variance; the 5-seed mean (57.2±1.1% acc, 43.2±1.5% mF1, 55.3±1.4% wF1) provides a more robust estimate.

from each dataset’s mean (Figure 6).

4.7 Error Analysis and Confusion Patterns

Confusion matrices (Figure 3, Figure 4, Figure 5), reveal dataset-specific patterns. On MELD, rare emotions frequently misclassify as Neutral (Fear→Neutral 32%, Disgust→Neutral 22%). On M3ED, neutral absorption is even more pronounced (Happy→Neutral 56%, Fear→Neutral

Method	Params	wF1
COSMIC (Ghosal et al., 2020)	355M	65.2
EmoBERTa (Kim and Vossen, 2021)	355M	65.6
UniMSE (Hu et al., 2022b)	355M	65.5
MERITS-L (Dutta and Ganapathy, 2025)	0.6M	66.0
SpeechCueLLM (Wu et al., 2025c) [‡]	—	67.6
DialogueLLM (Zhang et al., 2025) [†]	—	71.9
UniMER	292M	<u>70.7</u>

Table 6: Comparison with SFT-based and LLM-based methods on MELD. UniMER achieves 70.7% weighted F1, compared with 71.9% for DialogueLLM and 67.6% for SpeechCueLLM. Best result is in bold; second best is underlined.

53%), consistent with Chinese emotional restraint (Tsai et al., 2007); cross-confusion between negative emotions is also notable—Fear→Sad 19%, Disgust→Anger 13%—reflecting subtle negative-emotion boundaries in Chinese, while Surprise achieves the highest accuracy (57%), likely due to distinctive linguistic markers. On AkaCE, the 25:1 imbalance produces severe neutral dominance (Joy→Neutral 69%), though UniMER achieves non-trivial recognition on all seven classes including the rarest; small test sets for Disgust (11) and Fear (16) yield high per-class variance, and Fear and Disgust exhibit unique bidirectional confusion (25% and 27%), suggesting shared prosodic cues for these high-arousal negative emotions.

5 Conclusion

The central finding of this work is that natural language is a highly effective universal interface for multimodal emotion recognition. Converting visual, acoustic, and facial cues into text descriptions allows a single LoRA-adapted LLM to handle all modalities and languages, achieving state-of-the-art macro-F1 and rare-class performance, and competitive weighted F1, across three benchmarks in three language families under severe class imbalance, while using only 2.4% trainable parame-

Configuration	MELD			M3ED			AkaCE		
	Acc	wF1	mF1	Acc	wF1	mF1	Acc	wF1	mF1
All (w/ Proto Pool)	71.1	70.7	58.7	56.7	55.6	45.7	59.3	57.7	45.0
w/o Proto Pool	70.6	70.4	57.3	55.8	55.3	45.3	55.6	53.5	41.0
w/o LLM-Guided Capt.	69.9	69.3	54.6	—	—	—	51.5	50.2	26.9
w/o LoRA Fine-Tuning	66.7	65.6	49.7	53.2	52.5	41.8	54.8	53.4	38.0
Δ w/o Proto Pool	-0.5	-0.3	-1.4	-0.9	-0.3	-0.4	-3.7	-4.2	-4.0
Δ w/o LLM-Guided Capt.	-0.7	-1.1	-2.7	—	—	—	-4.1	-3.3	-14.1
Δ w/o LoRA Fine-Tuning	-3.9	-4.8	-7.6	-2.6	-2.8	-3.5	-0.8	-0.1	-3.0

Table 7: Component ablation (seed 42). “All” is the full UniMER system with class-prototype pooling. The “w/o Proto Pool” row uses standard multi-head pooling; subsequent component ablations (indented) remove one component at a time from this base configuration. Proto pooling yields the largest gains on AkaCE (−3.7pp Acc, −4.0pp mF1 when removed), scaling with imbalance severity. LoRA fine-tuning produces the largest component drops (−3.9pp Acc, −7.6pp mF1 on MELD). Entries marked “—” for M3ED indicate that LLM-guided captioning is not applicable: M3ED’s video files are not publicly distributed due to copyright restrictions, so all M3ED experiments use text-only input.

Train	→ MELD			→ M3ED			→ AkaCE		
	Acc	wF1	mF1	Acc	wF1	mF1	Acc	wF1	mF1
MELD	71.1	70.7	58.7	48.7	45.9	33.2	52.2	46.5	23.3
M3ED	52.3	51.9	37.4	56.7	55.6	45.7	40.7	40.4	25.9
AkaCE	46.7	47.1	33.0	49.7	47.5	33.0	59.3	57.7	45.0

Table 8: Zero-shot cross-dataset transfer among the three datasets sharing a 7-class Ekman taxonomy. Bold entries denote in-domain performance with class-prototype pooling. M3ED→MELD achieves 52.3% accuracy with 37.4% mF1, suggesting partial cross-lingual transfer, while MELD→AkaCE achieves 52.2% accuracy but only 23.3% mF1, indicating majority-class prediction. AkaCE-trained models reach 46.7–49.7% accuracy on English and Chinese benchmarks despite only 3.9k training samples.

ters. Class-prototype attention pooling ensures this shared representation serves all emotions equally, with gains scaling with imbalance severity (AkaCE +4.0pp, MELD +1.4pp, M3ED +0.4pp mF1).

Cross-lingual analysis reveals that emotion is both more universal and culturally specific than current systems assume: arousal-based confusions occur across all three language families, while Chinese speakers more often suppress emotion toward neutral and Akan speakers show prosodic confusions absent in English. The primary bottleneck in multilingual emotion recognition is data representation, not model capacity. We posit that this “text as universal interface” paradigm extends beyond emotion recognition and release our code and models to enable further exploration.

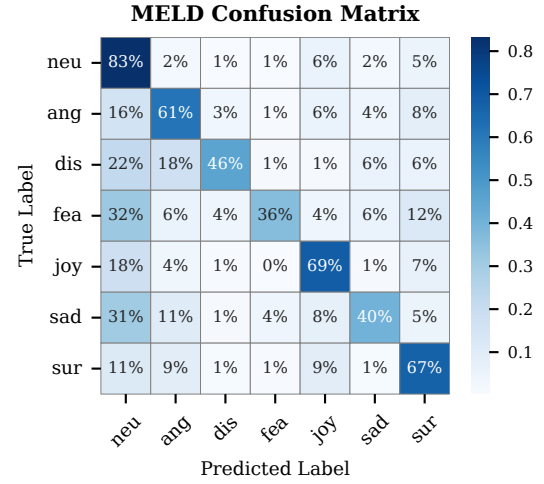


Figure 3: Confusion matrix for MELD (7 classes, 18:1).

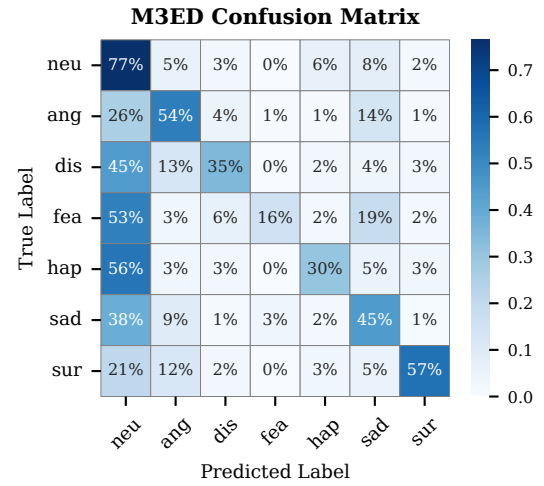


Figure 4: Confusion matrix for M3ED (7 classes, 11:1).

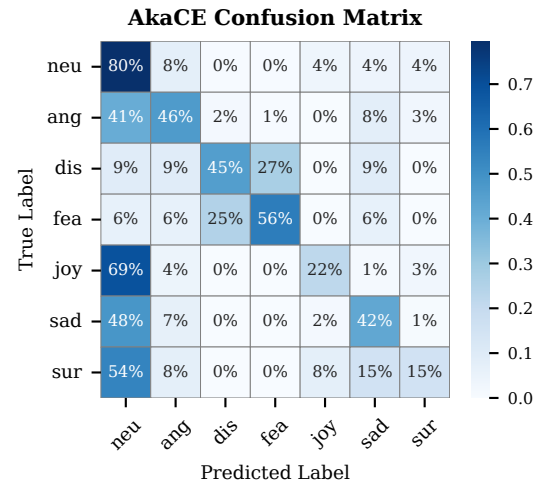


Figure 5: Confusion matrix for AkaCE (7 classes, 25:1).

6 Limitations

M3ED experiments use text-only input because video files are not publicly distributed due to copyright restrictions (Zhao et al., 2022), limiting full multimodal evaluation on that benchmark. In addition, none of the three datasets provide speaker profile information (e.g., age, gender, personality traits), so our model cannot leverage speaker-level priors that may benefit emotion recognition in longer conversations.

7 Ethics Statement

We acknowledge and thank Poria et al. (Poria et al., 2019), Zhao et al. (Zhao et al., 2022), and Sasu et al. (Sasu et al., 2025) for their valuable work in constructing and sharing the MELD, M3ED, and AkaCE datasets. MELD is licensed under the GNU General Public License v3.0², M3ED is publicly available under its original release terms³, and AkaCE is publicly available⁴. Our multimodal pre-processing pipeline uses publicly available tools, including OpenFace (Baltrušaitis et al., 2018) for facial Action Unit extraction and Gemma-3n-E4B-it for visual description generation, operating solely on existing video data without collecting new personal information.

While our work relies on publicly available research datasets with established ethical protocols, we acknowledge broader concerns surrounding emotion recognition technologies, as discussed by Stark and Hutson (2022) and Stark and Hoey (2021). If deployed in real-world settings, such systems may raise ethical considerations related to user privacy, informed consent, and algorithmic bias. We emphasize that any practical application should: (1) obtain explicit informed consent for emotion analysis and facial data processing; (2) implement appropriate privacy protections, including data anonymization and secure storage; (3) evaluate and mitigate potential demographic biases; and (4) comply with relevant regulations and ethical guidelines.

We are committed to transparency in our research methods, data sources, and limitations. All code and materials are released on GitHub⁵ and HuggingFace Hub under an open-source MIT li-

cense to facilitate independent verification and responsible reuse. Future work should prioritize cross-cultural validation on more diverse datasets and develop robust bias mitigation strategies to ensure equitable performance across demographic groups.

References

- Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. [Openface 2.0: Facial behavior analysis toolkit](#). In *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 59–66.
- Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G. Hauptmann. 2024. [Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 110805–110853. Curran Associates, Inc.
- Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. [Class-balanced loss based on effective number of samples](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277.
- Soumya Dutta and Sriram Ganapathy. 2025. [LLM supervised pre-training for multimodal emotion recognition in conversations](#). In *ICASSP 2025 - IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.
- Paul Ekman. 1992. [An argument for basic emotions](#). *Cognition & Emotion*, 6(3-4):169–200.
- Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Yulin Wu, and Bingquan Liu. 2025. [Be-MERC: Behavior-aware MLLM-based framework for multimodal emotion recognition in conversation](#). *arXiv preprint arXiv:2503.23990*.
- Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. [COSMIC: COmmonSense knowledge for eMotion identification in conversations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.
- Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. [DialogueGCN: A graph convolutional neural network for emotion recognition in conversation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

²<https://github.com/declare-lab/MELD/blob/master/LICENSE>

³<https://github.com/aim3-ruc/rucm3ed>

⁴<https://github.com/sasudavid/AkaCE>

⁵<https://github.com/davidanastasiu/unimer>

- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. [Lora: Low-rank adaptation of large language models](#). In *Proceedings of the International Conference on Learning Representations*.
- Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022b. [UniMSE: Towards unified multimodal sentiment analysis and emotion recognition](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. [MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675. Association for Computational Linguistics.
- Zilong Huang, Man-Wai Mak, and Kong Aik Lee. 2024. [Mm-nodeformer: Node transformer multimodal fusion for emotion recognition in conversation](#). In *Proceedings of Interspeech 2024*, pages 4069–4073. ISCA.
- Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. 2012. [Facial expressions of emotion are not culturally universal](#). *Proceedings of the National Academy of Sciences*, 109(19):7241–7244.
- Taewoon Kim and Piek Vossen. 2021. [Emoberta: Speaker-aware emotion recognition in conversation with roberta](#). *Preprint*, arXiv:2108.12009.
- Yoon Kim. 2014. [Convolutional neural networks for sentence classification](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751. Association for Computational Linguistics.
- Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2023. [Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models](#). *arXiv preprint arXiv:2309.11911*.
- Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. [Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1204–1214, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. [EmoCaps: Emotion capsule based model for conversational emotion recognition](#). In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.
- Yuxiang Lin and Shunchao Zheng. 2025. [MER-Factory: Automated factory for multimodal emotion recognition and reasoning data](#). Software, version 0.1.0.
- Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2024a. [A transformer-based model with self-distillation for multimodal emotion recognition in conversations](#). *IEEE Transactions on Multimedia*, 26:776–788.
- Wenxuan Ma, Shuang Li, Lincan Cai, and Jingxuan Kang. 2024b. [Learning modality knowledge alignment for cross-modality transfer](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33777–33793. PMLR.
- Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024c. [emotion2vec: Self-supervised pre-training for speech emotion representation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand. Association for Computational Linguistics.
- Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. [Dialoguernn: An attentive rnn for emotion detection in conversations](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.
- Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. 2021. [DialogueTRM: Exploring multimodal emotional dynamics in a conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tao Meng, Fuchen Zhang, Yuntao Shou, Hongen Shao, Wei Ai, and Keqin Li. 2024. [Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4298–4312.
- Kun Peng, Cong Cao, Hao Peng, Guanlin Wu, Zhifeng Hao, Lei Jiang, Yanbing Liu, and Philip S. Yu. 2025. [Emotion transfer with enhanced prototype for unseen emotion recognition in conversation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 596–608, Suzhou, China. Association for Computational Linguistics.
- Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. [Context-dependent sentiment analysis in user-generated videos](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 873–883.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. [Meld: A multimodal multi-party dataset for emotion recognition in conversations](#). In

- Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536. Association for Computational Linguistics.
- Minjie Ren, Xiangdong Huang, Wenhui Li, Dan Song, and Weizhi Nie. 2022. [Lr-gcn: Latent relation-aware graph convolutional network for conversational emotion recognition](#). *IEEE Transactions on Multimedia*, 24:4422–4432.
- David Sasu, Zehui Wu, Ziwei Gong, Run Chen, Pengyuan Shi, Lin Ai, Julia Hirschberg, and Natalie Schluter. 2025. [Akan cinematic emotions \(ACE\): A multimodal multi-party dataset for emotion recognition in movie dialogues](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9820–9831. Association for Computational Linguistics.
- Luke Stark and Jesse Hoey. 2021. [The ethics of emotion in artificial intelligence systems](#). In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 782–793.
- Luke Stark and Jevan Hutson. 2022. [Physiognomic artificial intelligence](#). *Fordham Intell. Prop. Media & Ent. LJ*, 32:922.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, and 1 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Jeanne L. Tsai, Felicity F. Miao, Emma Seppala, Helene H. Fung, and Dannii Y. Yeung. 2007. [Influence and adjustment goals: Sources of cultural differences in ideal affect](#). *Journal of Personality and Social Psychology*, 92(6):1102–1117.
- ChengYan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025a. [Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6257–6274. Association for Computational Linguistics.
- Yuezhou Wu, Siling Zhang, and Pengfei Li. 2025b. [Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features](#). *Scientific Reports*, 15:8855.
- Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2025c. [Beyond silent letters: Amplifying LLMs in emotion recognition with vocal nuances](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2202–2218, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yunhe Xie, Chengjie Sun, Ziyi Cao, Bingquan Liu, Zhenzhou Ji, Yuanchao Liu, and Lili Shan. 2025. [A dual contrastive learning framework for enhanced multimodal conversational emotion recognition](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4055–4065, Abu Dhabi, UAE. Association for Computational Linguistics.
- Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. [Memory fusion network for multi-view sequential learning](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2025. [Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations](#). *Neural Networks*, 192:107901.
- Jinming Zhao, Tengan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. [M3ED: Multi-modal multi-scene multi-label emotional dialogue database](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710. Association for Computational Linguistics.
- Shihao Zou, Xianying Huang, Xudong Shen, and Hankai Liu. 2022. [Improving multimodal fusion with main modal transformer for emotion recognition in conversation](#). *Knowledge-Based Systems*, 258:109978.

A Error Analysis and Confusion Patterns

The MELD, M3ED, and AkaCE confusion matrices appear in Section 4.7 (Figure 3, Figure 4, and Figure 5, respectively). Key MELD-specific patterns beyond those discussed in Section 4.7—beyond neutral absorption, Sadness→Anger (8%) and Disgust→Anger (16%)—indicate confusion among negative-valence emotions sharing similar lexical cues in English sitcom dialogue. Figure 6 isolates these culturally-specific effects from raw confusion rates by plotting, for each emotion, how far its neutral-absorption rate deviates from the dataset’s overall mean; this removes general dataset difficulty as a confounding factor and highlights which neutral-absorption and arousal-based confusions are distinctively cultural rather than universal.

B LLM-Guided Multimodal Captioning Details

This section provides full implementation details for the captioning pipeline summarized in Algorithm 1 and Section 3.2.

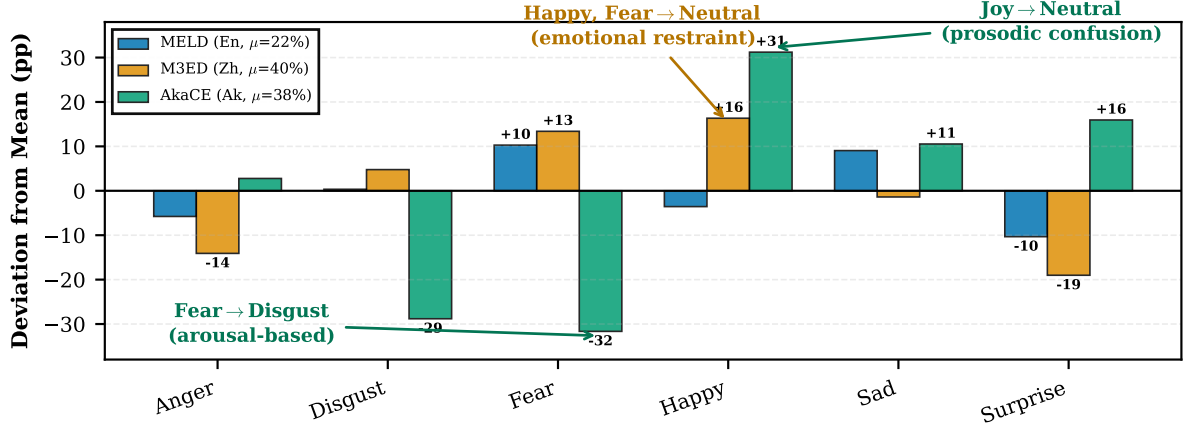


Figure 6: Deviation of per-emotion neutral absorption rates from each dataset’s mean (MELD $\mu=22\%$, M3ED $\mu=40\%$, AkaCE $\mu=38\%$), isolating culturally-specific patterns from general dataset difficulty. Chinese M3ED shows a +16.3pp excess for Happy→Neutral and +13.4pp for Fear→Neutral, consistent with cultural emotional restraint (Tsai et al., 2007). AkaCE shows disproportionate Joy→Neutral (+31.2pp) and Surprise→Neutral (+15.9pp) absorption, while Fear is instead confused with Disgust (−31.7pp), suggesting arousal-based rather than suppressive confusion patterns in Akan.

AU Peak Detection. We extract AUs frame-by-frame using OpenFace (Baltrušaitis et al., 2018), then apply peak detection (`scipy.signal.find_peaks`, threshold $\tau=0.8$) following MER-Factory (Lin and Zheng, 2025; Cheng et al., 2024). AU activation patterns at peaks are mapped to emotion categories (e.g., AU06+AU12 → happy).

Modality-Specific Prompts. Gemma-3n-E4B-it analyzes audio, peak frame, and video independently. All prompts request only observable features without emotion labeling. The four prompts are:

p_{audio}: “Describe all audible characteristics in this clip. If speech is present, describe the speaker’s pitch contour (rising, falling, flat, tonal variations), speech rate (rushed, hesitant, steady), voice quality (tense, breathy, trembling, loud, soft), rhythm and emphasis patterns, and any non-verbal vocalizations (pauses, sighs, laughter, exclamations). If speech is absent or minimal, describe background music (tempo, key mode, instrumentation, dynamics) and ambient sounds (silence, environmental noise, sound effects).”

p_{image}: “Describe the person’s observable facial configuration (brow position, eye openness, mouth shape, muscle tension), body posture, hand gestures, and gaze direction. ”

p_{video}: “Describe what happens in this video clip. Note changes over time in the speaker’s facial con-

figuration, body movement, gestures, and spatial behavior toward others. ”

p_{synth}: “You are an expert in multimodal behavioral analysis across cultures. Given the following observational field notes, systematically organize the evidence by modality (audio, visual, video, facial action units). For each piece of evidence, describe what behavioral pattern it reflects, noting that the same cue can carry different meanings across languages and cultures. Produce a structured summary of all observed behavioral signals. ”

All prompts are culturally neutral and label-free; emotion classification is deferred to the supervised LoRA classifier (Section 3.2). This decouples perception from judgment, allowing the classifier to learn culture-specific cue–emotion mappings. The synthesis prompt receives the concatenation of all modality analyses and AU-derived emotion peaks as its “Observational Field Notes” context.

C Detailed Configuration for Reproducibility

Tables 9 and 10 list all hyperparameters for our methods.

Parameter	MELD	M3ED	AkaCE
<i>Training</i>			
Context length	40	20	2
Effective batch size	48	24	24
Grad. accum. steps	24	12	12
Epochs	1	3	3
KL loss weight (λ)	0.02	0.02	0.02
Modality	Text+Caption	Text-only	Text+Caption
<i>Evaluation</i>			
5-seed set	42, 123, 456, 789, 2024		
SOTA table seed	42	789	42
Ablation seed	42	42	42

Table 9: Per-dataset configuration. Parameters not listed are shared across all datasets (Table 10). SOTA table seeds are the best-performing seed from 5-seed evaluation.

Parameter	Value
<i>Base Model</i>	
Model	Gemma-3-12B-it
Hidden size	3840
Precision	bfloat16
Base model	Frozen
<i>LoRA</i>	
Strategy	Comprehensive (Q/K/V/O + FFN)
Rank (r)	64
Alpha (α)	128
LoRA dropout	0.05
Bias	None
Trainable params	292M (2.4%)
<i>Prototype Pooling (Proto-Lite)</i>	
Num prototypes (K)	7 (= num classes)
Projection dim	256
Prototype dropout	0.3
Output MLP	$K \times 3840 \rightarrow 256$ (GELU) $\rightarrow 3840$
Init	Xavier uniform, bias = 0
Proto vectors init	$\mathcal{N}(0, 0.02)$
<i>Optimizer & Scheduler</i>	
Optimizer	AdamW ($\beta_1=0.9, \beta_2=0.999$)
Learning rate	5×10^{-5}
LR scheduler	Cosine annealing
Warmup ratio	0.1
Weight decay	0.01
Max gradient norm	1.0
<i>Loss</i>	
Class weight method	Effective samples
Class weight β	0.9999
Label smoothing (ϵ)	0.15
KL loss type	Label smoothing
<i>Data & Training</i>	
Max sequence length	1024 tokens
Per-device batch size	2
Early stopping patience	3 (on eval macro F1)
Hardware	1 $\times$ A100 80 GB
Deterministic	cuda.deterministic=True

Table 10: Shared configuration across all datasets. LoRA is applied to all attention projections (Q, K, V, O) and FFN layers (gate, up, down) across all 48 transformer blocks.