



Unsupervised multimodal emotion-unified representation learning with dual-level language-driven cross-modal emotion alignment

Shaoze Feng ^a, Qiyin Zhou ^a, Yuanyuan Liu ^{a,*}, Ke Wang ^a, Kejun Liu ^a,
Chang Tang ^b

^a China University of Geosciences (Wuhan), Wuhan, China

^b Huazhong University of Science and Technology, Wuhan, China

ARTICLE INFO

Keywords:

Multimodal emotion recognition
Cross-modal alignment
Cross-modal contrastive learning
Unsupervised learning
Text-guided reconstruction

ABSTRACT

Unsupervised Multimodal Emotion Recognition (UMER) aims to infer affective states by integrating unannotated multimodal data, such as text, speech, and images. A key challenge is the substantial semantic gaps between modalities, spanning both global cross-modal emotion cues and local fine-grained emotion changes within each modality. Without annotation, existing methods struggle to effectively align and fuse cross-modal emotional semantics, resulting in suboptimal UMER performance. To address this challenge, we propose **DLCEA**, a *Dual-level Language-Driven Cross-Modal Emotion Alignment* framework for robust unsupervised multimodal emotion representation learning. DLCEA leverages intrinsic emotion semantics in text to guide cross-modal alignment and introduces a dual-level semantic alignment scheme: **Text-guided Cross-modal Global Emotion Alignment (TGEA)** and **Text-guided Cross-modal Local Emotion Alignment (TLEA)**. Specifically, the TGEA module treats text as an alignment anchor and applies text-guided contrastive learning to align the global emotional features of audio and visual modalities with those of the text, achieving global emotion-level consistency across all three modalities. In parallel, TLEA incorporates an emotion-aware text masking strategy and text-guided audio/video reconstruction, enabling the model to capture subtle emotional cues and reinforce local-level cross-modal consistency, thereby further addressing fine-grained emotional alignment. By jointly modeling global and local emotional alignment, DLCEA learns unified and robust multimodal emotion representations in a fully unsupervised manner. Extensive experiments on multimodal datasets such as MAFW, MOSEI, and IEMOCAP demonstrate that DLCEA outperforms existing methods by a significant margin, achieving state-of-the-art performance. These results confirm the critical role of language-driven cross-modal emotional alignment in UMER. Code is available on <https://github.com/Tank9971/DLCEA>.

1. Introduction

Multimodal Emotion Recognition aims to automatically identify human emotional states by integrating information from multiple modalities, such as video, audio, and text, and plays a vital role in natural human-computer interaction, digital humans, and embodied intelligence [1]. Recent advances in deep learning, particularly large-scale pre-trained multimodal models [2], have significantly improved emotion recognition performance. These approaches typically rely on large amounts of high-quality labeled data. However, emotional annotation across modalities is inherently subjective and complex, being influenced by cultural, contextual, and individual variability, which makes data collection time-consuming, costly, and difficult to scale [3]. As a result, **Unsupervised Multimodal Emotion Recognition (UMER)** has emerged

as a promising alternative. UMER leverages unsupervised learning techniques to integrate unlabeled multimodal data, automatically uncover latent emotional patterns, and learn unified emotion representations, thereby enabling emotion recognition without the need for manual annotations.

Recently, UMER has gained increasing attention due to its potential for scalable deployment in real-world applications [4,5]. For instance, Riccardo et al. [6] focus on discrete emotion classification by incorporating text, audio, and visual modalities within a framework based on contrastive learning between modality pairs. Mult [7], a Transformer-based architecture, is introduced to handle unaligned multimodal sequences while capturing long-range dependencies across modalities. Khare et al. [8] further explore large-scale unlabeled multimodal datasets to pre-train cross-modal Transformer models. They remove the decoder after

* Corresponding author.

E-mail addresses: fengshaoze@cug.edu.cn (S. Feng), zhouqiyin@cug.edu.cn (Q. Zhou), liuyy@cug.edu.cn (Y. Liu), WK2023@cug.edu.cn (K. Wang), liukejun@cug.edu.cn (K. Liu), tangchang@cug.edu.cn (C. Tang).

<https://doi.org/10.1016/j.patcog.2026.113160>

Received 30 August 2025; Received in revised form 25 December 2025; Accepted 25 January 2026

Available online 28 January 2026

0031-3203/© 2026 Published by Elsevier Ltd.

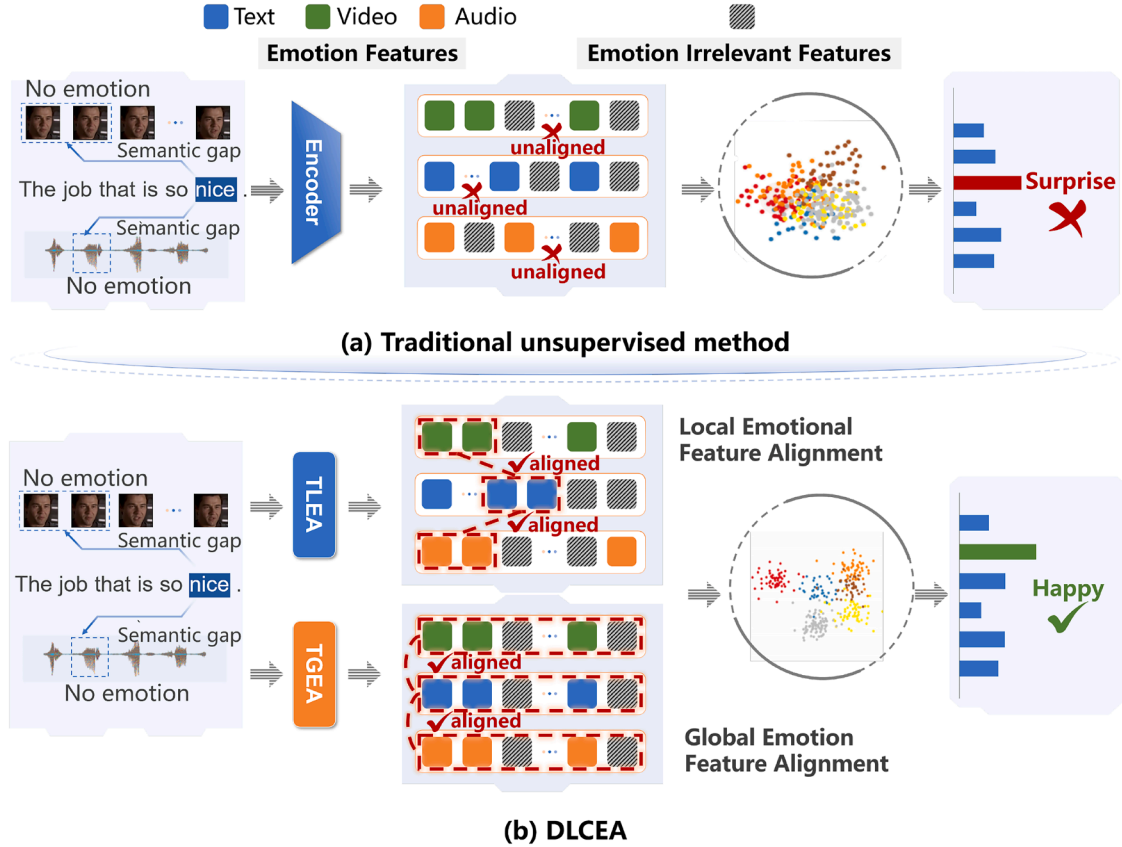


Fig. 1. Motivation of our proposed method for UMER. (a) illustrates the traditional UMER framework, which struggles to effectively align emotional cues across different modalities, often resulting in sub-optimal performance. (b) our proposed approach, which introduces two novel modules, TLEA and TGEA, to facilitate both global and local cross-modal emotional semantic feature alignment. This enables more accurate emotion recognition by bridging semantic gaps and enhancing emotion-unified multimodal representation.

pretraining and employ average pooling followed by a fully connected layer for fine-tuning on downstream emotion recognition tasks.

Despite recent progress in UMER, most existing methods primarily focus on multimodal emotion fusion while overlooking intrinsic semantic discrepancies across modalities, which often leads to suboptimal and incoherent multimodal emotion representations. Two key challenges remain. *First*, achieving globally consistent emotion semantics is difficult due to the substantial heterogeneity among modalities (text, audio, and video) in terms of structure, representation, and expressiveness; *Second*, emotional cues in each modality are often subtle, ambiguous, and highly context-dependent, making it challenging to capture fine-grained emotional signals both within and across modalities. As illustrated in Fig. 1(a), current approaches tend to neglect cross-modal semantic inconsistencies, e.g., pairing the word “nice” with a “neutral” facial expression, which can lead to misleading fused representations and compromise the robustness of UMER.

To address the two major challenges of the global cross-modal affective semantic gap and the inconsistent fine-grained emotional dynamics within modalities, we propose a novel unsupervised multimodal emotion-unified representation learning approach with Dual-level Language-Driven Cross-Modal Emotion Alignment (DLCEA) for robust UMER. DLCEA leverages the fact that the textual modality (language) typically carries richer emotional semantic information than other modalities in real-world affective interactions. As illustrated in Fig. 1(b), our DLCEA method innovatively introduces two key emotional semantic alignment modules: Text-guided Cross-modal Global Emotion Alignment (TGEA) and Text-guided Cross-modal Local Emotion Alignment (TLEA). To ensure global emotional semantic consistency across modalities, the TGEA module introduces a text-guided cross-modal con-

trastive learning strategy that aligns audio and visual features with textual semantics, respectively, achieving unsupervised global emotion alignment. To further achieve local emotional semantic consistency, the TLEA module devises an emotion-aware masking strategy and text-guided audio/video reconstruction for fine-grained emotion cue interaction and local-level cross-modal alignment. By leveraging the synergy between these two modules, our DLCEA effectively exploits the rich emotional information in the textual modality and addresses cross-modal sentiment alignment challenges at both global and local levels, resulting in unified and robust unsupervised multimodal emotion representations. In summary, the main contributions of this paper are as follows:

(1) We propose a novel unsupervised multimodal emotion recognition framework, termed DLCEA, to tackle two challenges: the global cross-modal affective semantic gap and the inconsistent fine-grained emotional dynamics within modalities, enabling accurate emotional representation and recognition without any manual annotations.

(2) In DLCEA, we design two complementary emotional semantic alignment modules: TGEA and TLEA. The TGEA module aligns global emotional cues across modalities via text-guided cross-modal contrastive learning, ensuring emotional consistency between audio/video and text while mitigating the global affective semantic gap. Meanwhile, the TLEA module focuses on fine-grained local alignment by reconstructing masked visual/audio features under the guidance of emotion-aware textual cues, capturing subtle emotional variations. The two modules collaborate closely, effectively improving unsupervised multimodal emotion-unified representations for UMER.

(3) Compared to existing UMER methods, our proposed DLCEA achieves state-of-the-art performance on three popular multimodal

emotion datasets: MAFW, CMU-MOSEI, and IEMOCAP. Specifically, DLCEA demonstrates improvements of 1.19% in Weighted Average Recall (WAR) and 1.29% in F1-score on the CMU-MOSEI dataset, a 1.04% WAR enhancement on MAFW, and respective increases of 1.5% WAR and 3.4% F1-score on IEMOCAP. These results clearly demonstrate the model's effectiveness in handling emotional semantic discrepancy and enhancing unsupervised emotion recognition.

With these contributions, our study provides valuable insights and advances in unsupervised multimodal emotion recognition, which contributes to the dual-level emotional semantic alignment between different modalities. The rest of this paper is as follows. [Section 2](#) presents related work. [Section 3](#) presents the details of our proposed DLCEA. [Section 4](#) discusses our experimental results on publicly available datasets. [Section 5](#) summarises the research in this paper.

2. Related work

2.1. Multimodal emotion recognition

Multimodal emotion recognition (MER) has attracted increasing research interest in recent years, with a wide range of methodologies proposed to integrate and model emotional cues from multiple modalities, such as text, visual, and acoustic signals.

Existing MER methodologies can be broadly classified into three main categories based on their core approaches: traditional methods, deep learning-based approaches, and advanced fusion strategies. Traditional methods encompass early systems that primarily relied on classifiers such as Support Vector Machines (SVMs) and Logistic Regression, combined with handcrafted features extracted from each modality. While foundational, these methods often suffered from limited generalization capabilities [9,10]. Deep learning-based approaches emerged with the advent of deep learning. Convolutional Neural Networks (CNNs) became the standard for processing spatial data, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory networks (LSTMs), were widely adopted for modeling temporal sequences. Subsequently, Transformer architectures further enhanced the modeling of complex cross-modal interactions and long-range dependencies. Advanced fusion strategies, representing the focus of recent research, center on developing sophisticated techniques to better integrate information across modalities [11]. Ghaleb et al. [12] proposed a model that combines deep metric learning with LSTM components to effectively capture temporal variations in emotional states. Unlike traditional late fusion strategies, which typically concatenate modality-specific features either directly or in a pairwise fashion, their approach focuses on modeling emotion transitions over time. To improve multimodal fusion strategies, various alternatives have been explored. For instance, the M3ER framework [13] introduces a multiplicative data-driven fusion technique that dynamically accentuates more informative modalities while suppressing less reliable ones. Furthermore, attention-based mechanisms [14], particularly sparse cross-modal attention, have proven highly effective in dynamically focusing on salient emotional cues within and across modalities, thereby enhancing alignment and representation learning.

While most existing approaches rely heavily on supervised learning paradigms requiring large amounts of labeled data, our work explores an alternative direction by addressing the MER task in a fully unsupervised setting [15].

2.2. Unsupervised emotion representation and recognition

Unsupervised emotion representation and recognition seeks to infer emotional states from multimodal data without labeled supervision. Existing methods in this field can be broadly categorized into several representative directions. **Clustering-based methods.** Early efforts predominantly relied on clustering techniques, grouping hand-engineered

features, such as Mel-frequency cepstral coefficients in speech or facial landmarks in images, into clusters representing potential emotional states [16]. While foundational, these clustering-based methods were fundamentally limited by their reliance on shallow, often hand-engineered features, hindering robust semantic understanding of complex emotions. **Cross-modal alignment-based methods.** Subsequent research shifted towards leveraging cross-modal alignment as a source of self-supervision. Recognizing the inherent co-occurrence of emotional cues across modalities [17], these methods aim to learn joint representations by maximizing agreement or minimizing discrepancy between corresponding segments of different modalities, such as aligning prosodic speech features with facial expressions or textual sentiment with vocal tone [7,18]. Although this approach implicitly captures complementary emotional information, it often struggles to achieve precise semantic consistency both at the global level and the local level [19]. **Generative modeling approaches.** Another important line of work employs generative models, including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). These models learn latent emotion representations through reconstruction objectives or data generation, aiming to disentangle underlying emotional factors and yielding interpretable embeddings useful for downstream tasks like clustering or zero-shot classification. However, effectively isolating emotion-specific factors from confounding variables within the latent space remains a persistent challenge. **Masked modeling methods.** More recently, inspired by the success of BERT and GPT, transformer-based masked modeling has been adapted for multimodal emotion data. By predicting masked parts within input sequences, these models learn contextual representations capturing dependencies within and potentially across modalities. While powerful for context modeling, achieving optimal cross-modal semantic alignment, particularly encompassing both global sentiment consistency and local cue correspondence, through masking alone proves difficult.

Despite these varied methodological advances, significant challenges persist in unsupervised MER. A core limitation lies in effectively establishing consistent semantic alignment across modalities, addressing both global emotional coherence and local emotional cue interactions. Furthermore, disentangling emotion-related features from confounding factors and the scarcity of large-scale, high-quality unlabeled affective datasets pose substantial hurdles. To bridge this gap and tackle the critical challenge of achieving comprehensive cross-modal emotional alignment at multiple granularities, our work proposes a novel unsupervised framework centered on deep text guidance.

3. Method

3.1. Overview

To tackle the challenges of cross-modal emotion alignment and generate unified, robust emotion representations, we propose a novel unsupervised framework—**Dual-level Language-Driven Cross-Modal Emotion Alignment (DLCEA)**. The overall training pipeline is illustrated in [Fig. 2](#), which consists of three main components: Multimodal Feature Extraction, Text-guided Cross-modal Global Emotion Alignment (TGEA) and Text-guided Cross-modal Local Emotion Alignment (TLEA). Specifically, we first utilize large-scale pre-trained foundation models to extract different modal features in multimodal feature extraction module. Then, to enhance global emotional consistency across different modality features, TGEA introduces text-guided cross-modal contrastive learning that aligns the global emotional semantics of audio and visual modalities with those of the text, respectively, thus achieving global emotion semantic alignment across all modalities. Furthermore, to further address fine-grained emotion inconsistencies, TLEA introduces emotion-aware masking and text-guided cross-modal reconstruction, enabling local-level emotion alignment for each modality. Together, TGEA and TLEA enhance the unified representation of multimodal emotional semantics at both global and local levels, significantly

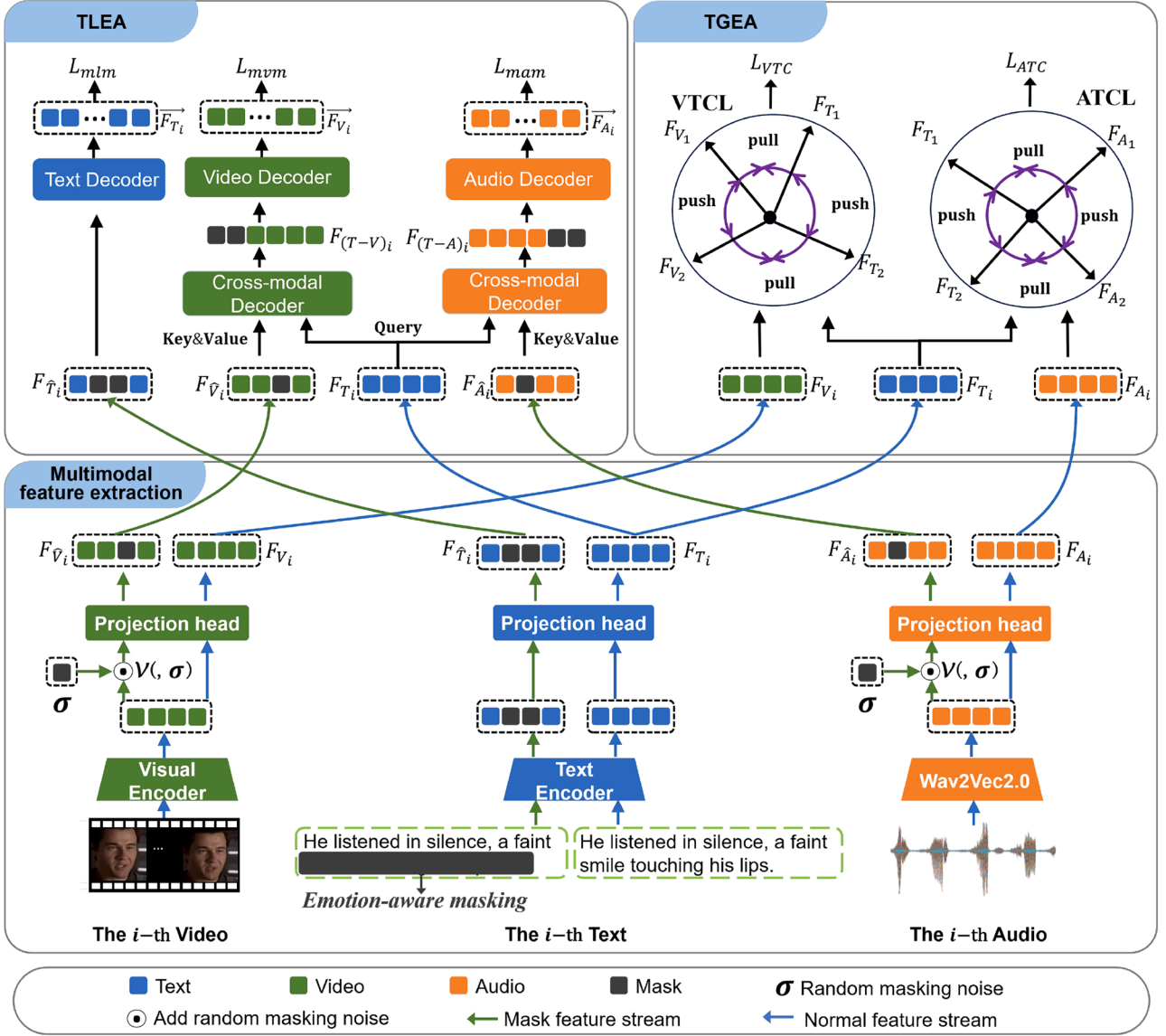


Fig. 2. The proposed DLCEA framework consists of three components: Multimodal Feature Extraction, Text-guided Cross-modal Global Emotion Alignment (TGEA), and Text-guided Cross-modal Local Emotion Alignment (TLEA). TGEA employs text-guided contrastive learning to align global emotional semantics across modalities, while TLEA enhances local emotional alignment through emotion-aware masking and text-guided reconstruction in each modality stream.

improving the robustness of the proposed UMER framework. In the following sections, we will elaborate on these three components in detail.

3.2. Multimodal feature extraction

To capture rich multimodal features, we leverage pre-trained foundation models whose representational power is derived from large-scale, diverse datasets. Specifically, we use the pre-trained CLIP [1] to encode visual and textual modalities, and Wav2Vec 2.0 [2] for audio modality embedding. For detailed explanation, we compare different modal encoders for multimodal feature extraction in Table 6 of Section 4.3.3. In our framework, we use these encoders to generate two different sets of features through parallel extraction paths: (1) Normal Feature Extraction. It is used to obtain standard feature representations; (2) Noise-masked Feature Extraction. It is used to generate emotion-sensitive features for masked emotions. The implementation details of these two paths will be discussed below.

3.2.1. Normal feature extraction

For each modality, the feature extraction pipeline comprises a modality-specific pre-trained encoder and a multi-layer nonlinear projection head. The projection heads share the same architecture across modalities, while their parameters are not shared.

Text Encoder for Textual Feature Extraction. The text encoder employs a pre-trained CLIP text encoder followed by a multi-layer nonlinear projection head. Inspired by SimCLR [20], the projection head is used to enhance the quality of unimodal representations. Formally, given an input text sequence T_i , the encoder extracts the normal text feature $F_{T_i} \in \mathbb{R}^{32 \times 512}$:

$$F_{T_i} = f_P(f_{CLIP-T}(T_i)), \quad (1)$$

where $f_{CLIP-T}(\cdot)$ and $f_P(\cdot)$ denote the pre-trained CLIP text encoder and nonlinear projection head.

Video Encoder for Visual Feature Extraction. The video encoder architecture utilizes the visual encoder of CLIP [1] followed by a multi-layer projection head to extract visual features. Given an input video V_i , the

encoder produces the video features $F_{V_i} \in \mathbb{R}^{32 \times 512}$:

$$F_{V_i} = f_P(f_{CLIP-V}(V_i)), \quad (2)$$

where $f_{CLIP-V}(\cdot)$ and $f_P(\cdot)$ denote the pre-trained CLIP visual encoder and the projection head, respectively.

Audio Encoder for Audio Feature Extraction. The audio encoder consists of a pre-trained Wav2Vec2.0 [2] model followed by a multi-layer projection head for feature extraction. Given an input audio signal A_i , the encoder extracts the audio features $F_{A_i} \in \mathbb{R}^{32 \times 512}$:

$$F_{A_i} = f_P(f_{Wav2Vec2.0}(A_i)), \quad (3)$$

where $f_{Wav2Vec2.0}(\cdot)$ and $f_P(\cdot)$ denote the pre-trained Wav2Vec2.0 encoder and the projection head, respectively.

3.2.2. Noise-masked feature extraction

For the textual modality T_i , we introduce the emotion-aware masking strategy to generate noise-masked textual features $F_{\hat{T}_i}$, which is highly sensitive to emotion information. In practice, we first mask emotion-related words in the textual modality to obtain the emotion-masked text input as T'_i , as shown in Fig. 2. With the T'_i as input, we use the shared process described in Section 3.2.1 to extract the noise-masked textual features $F_{\hat{T}_i} \in \mathbb{R}^{32 \times 512}$. The process is given by

$$F_{\hat{T}_i} = f_P(f_{CLIP-T}(T'_i)). \quad (4)$$

In addition, for the video/audio modalities, we extract video/audio features using pre-trained CLIP [1] and Wav2Vec2.0 [2], respectively. We use the $\mathcal{V}(\sigma)$ to represent the process that explicitly adds the random masked noisy information σ to the video and audio features. Thus, the noise-masked video features $F_{\hat{V}_i} \in \mathbb{R}^{32 \times 512}$ and audio features $F_{\hat{A}_i} \in \mathbb{R}^{32 \times 512}$ are formulated respectively as:

$$F_{\hat{V}_i} = f_P(\mathcal{V}(f_{CLIP-V}(V_i), \sigma)), F_{\hat{A}_i} = f_P(\mathcal{V}(f_{Wav2Vec2.0}(A_i), \sigma)). \quad (5)$$

Notably, our proposed emotion-aware masking strategy is fully implemented in an unsupervised manner, without relying on any human-annotated emotion labels or emotion classifiers. Specifically, we first use a generic face detector to locate the facial region in each video frame (which contains no emotion-related information), and then apply random local masking only within the detected facial area. This design keeps the process entirely unsupervised while making the masked regions more likely to correspond to emotion-relevant cues.

3.3. Text-guided cross-modal global emotion alignment

To mitigate inter-modal emotion gap for cross-modal emotional semantic alignment, we propose **Text-guided Cross-modal Global Emotion Alignment (TGEA)** module, which leverages text-guided cross-modal contrastive learning to align the global emotional semantics of audio and video with those of text, enabling high-quality multimodal emotion representations without relying on supervised annotations. Specifically, TGEA consists of two contrastive learning streams: (1) Video-Text Contrastive Learning (VTCL), which aligns global emotional semantics between video and text with a focus on affective cues; (2) Audio-Text Contrastive Learning (ATCL), which captures global emotional correlations between audio and text. Implementation details of VTCL and ATCL are provided in the following sections.

Video-Text Contrastive Learning (VTCL). VTCL aims to align text and video features in a shared embedding space by introducing video-text contrastive learning loss L_{VTC} , yielding globally consistent video-text emotion representations. Specifically, for the i th multimodal sample, we treat the text and video features (F_{T_i}, F_{V_i}) as a positive sample pair, while features from all other instances j serve as negative samples, where $i \neq j$. The learning objective of VTCL is to maximize the cosine similarity between (F_{T_i}, F_{V_i}) , and minimize it between F_{T_i} and F_{V_j} . Based on this,

the video-text contrastive learning loss L_{VTC} can be expressed as:

$$L_{VTC} = -\frac{1}{N} \sum_{i=1}^N \left(\log \frac{\exp\left(F_{T_i} \cdot \frac{F_{V_i}}{\tau}\right)}{\sum_{j=1}^N \exp\left(F_{T_i} \cdot \frac{F_{V_j}}{\tau}\right)} + \log \frac{\exp\left(F_{V_i} \cdot \frac{F_{T_i}}{\tau}\right)}{\sum_{j=1}^N \exp\left(F_{V_i} \cdot \frac{F_{T_j}}{\tau}\right)} \right), \quad (6)$$

where N and τ are the batch size and temperature parameters, respectively.

Audio-Text Contrastive Learning (ATCL). ATCL employs an audio-text contrastive learning loss L_{ATC} to align audio and text features in a shared embedding space, capturing global audio-text representations with consistent emotional semantics. Specifically, for the i th multimodal sample, the text and audio features (F_{T_i}, F_{A_i}) first produce a positive sample pair, while all other instances j are treated as negatives. The learning objective of ATCL is to maximize the cosine similarity between F_{T_i} and F_{A_i} , and minimize it between F_{T_i} and F_{A_j} for all $i \neq j$. Mathematically, the L_{ATC} is formulated as:

$$L_{ATC} = -\frac{1}{N} \sum_{i=1}^N \left(\log \frac{\exp\left(F_{T_i} \cdot \frac{F_{A_i}}{\tau}\right)}{\sum_{j=1}^N \exp\left(F_{T_i} \cdot \frac{F_{A_j}}{\tau}\right)} + \log \frac{\exp\left(F_{A_i} \cdot \frac{F_{T_i}}{\tau}\right)}{\sum_{j=1}^N \exp\left(F_{A_i} \cdot \frac{F_{T_j}}{\tau}\right)} \right). \quad (7)$$

Overall, in TGEA, VTCL and ATCL enable direct global emotion alignment between text and video, and text and audio modalities, respectively. By using text as a bridge, these mechanisms also induce indirect alignment between audio and video modalities, thus improving global multimodal feature alignment for unsupervised multimodal emotion representation.

3.4. Text-guided cross-modal local emotion alignment

To further achieve fine-grained local emotional semantic alignment and reduce intra-modal emotion gap for each modality, we propose **Text-guided Cross-modal Local Emotion Alignment (TLEA)** module. TLEA introduces an emotion-aware masking strategy combined with dual text-guided cross-modal reconstruction streams, enabling nuanced affective understanding both across and within modalities in an unsupervised manner for finer emotion representation.

TLEA consists of three learning streams: (1) Emotion-aware text reconstruction, which adopts an emotion-aware masking strategy applied to text features, encouraging the model to learn robust, emotion-sensitive textual representations; (2) Text-guided audio reconstruction, using the textual representations to guide the reconstruction of masked audio segments, improving fine-grained and local emotional alignment between textual and audio modalities; (3) Text-guided video reconstruction, where textual context aids in reconstructing masked local regions in video frames, enhancing local visual emotion modeling for fine-grained text-video emotion alignment. Details of each module are elaborated in the following.

Emotion-aware Text Reconstruction. To help the model more effectively capture the fine-grained emotion semantic information in the text modality, we introduce the emotion-aware masking strategy for facilitating the text-masked noise reconstruction. With the emotion-masked text feature $F_{\hat{T}_i}$, we employ a text decoder $Decoder(\cdot)$ with an emotion-masked text reconstruction loss to generate the reconstructed emotion-sensitive text feature $\bar{F}_{T_i} \in \mathbb{R}^{32 \times 512}$:

$$\bar{F}_{T_i} = Decoder(k/q/v = F_{\hat{T}_i}), \quad (8)$$

where $Decoder(\cdot)$ consists of three self-attention blocks. $F_{\hat{T}_i}$ serves as the Query (q), Key (k), Value (v) in this decoder, simultaneously. The

emotion-masked text reconstruction loss L_{mlm} is defined as:

$$L_{mlm} = \frac{1}{N} \sum_{i=1}^N l(\bar{F}_{T_i}, F_{T_i}), \quad (9)$$

where $l(\cdot)$ is the L_1 loss. By minimizing L_{mlm} , the model reconstructs the textual features masked with emotion to learn finer emotion-sensitive text representation, providing higher quality guidance for the audio and video noise-masked reconstruction.

Text-guided Audio Reconstruction. With the finer, emotion-sensitive textual representations, we design the text-guided audio reconstruction stream, aiming to enhance fine-grained, local emotion semantic alignment between audio and text features, as illustrated in Fig. 3. Concretely, using the noise-masked audio feature $F_{\hat{A}_i}$ and normal text feature F_{T_i} mentioned as in Section 3.2.1 are jointly input into a cross-modal decoder $DE_{cross}(\cdot)$. We can obtain a joint text-to-audio emotion representation $F_{(T-A)_i} \in \mathbb{R}^{32 \times 512}$ that captures the cross-modal emotional semantics, which is formulated as:

$$F_{(T-A)_i} = DE_{cross}(q = F_{T_i}, k/v = F_{\hat{A}_i}), \quad (10)$$

where $DE_{cross}(\cdot)$ consists of three blocks of cross-attention. $F_{\hat{A}_i}$ serves as the key and value, while F_{T_i} provides the query in this cross-decoder. This design allows the model to condition the reconstruction process on fine-grained textual cues, thereby enhancing the local emotion alignment between text and audio modalities. The joint representation $F_{(T-A)_i}$ is then passed through an audio decoder $Decoder(\cdot)$ to generate the reconstructed emotion-sensitive audio features $\bar{F}_{A_i} \in \mathbb{R}^{32 \times 512}$:

$$\bar{F}_{A_i} = Decoder(q/k/v = F_{(T-A)_i}). \quad (11)$$

To enforce emotional consistency, we introduce an audio reconstruction loss L_{mam} that aligns the reconstructed audio features with the original ones under text guidance, enhancing fine-grained audio-text emotion alignment. Formally, L_{mam} is defined as:

$$L_{mam} = \frac{1}{N} \sum_{i=1}^N l(\bar{F}_{A_i}, F_{A_i}). \quad (12)$$

Text-guided Video Reconstruction. Similar to the audio reconstruction stream, we adopt a parallel text-guided video reconstruction stream to facilitate fine emotion alignment between video and text, as illustrated in Fig. 3. First, the noise-masked video feature $F_{\hat{V}_i}$ and the corresponding text feature F_{T_i} are jointly fed into the cross-modal decoder $DE_{cross}(\cdot)$ to generate a joint text-to-video emotion representation $F_{(T-V)_i} \in \mathbb{R}^{32 \times 512}$ as:

$$F_{(T-V)_i} = DE_{cross}(q = F_{T_i}, k/v = F_{\hat{V}_i}). \quad (13)$$

Then, the joint representation is fed into the video decoder $Decoder(\cdot)$ to facilitate fine-grained local emotion alignment by reconstructing occluded video features $\bar{F}_{V_i}$ under the guidance of textual contextual information. The joint representation is then fed into the video decoder $Decoder(\cdot)$ to produce the reconstructed emotion-sensitive video features $\bar{F}_{V_i} \in \mathbb{R}^{32 \times 512}$ guided by textual context, enabling fine-grained local emotion alignment. Formally, the process can be written as:

$$\bar{F}_{V_i} = Decoder(q/k/v = F_{(T-V)_i}). \quad (14)$$

To achieve this, we employ the video reconstruction loss L_{mvm} , which is defined as:

$$L_{mvm} = \frac{1}{N} \sum_{i=1}^N l(\bar{F}_{V_i}, F_{V_i}). \quad (15)$$

In summary, TLEA achieves fine-grained cross-modal local emotion alignment by leveraging emotion-aware textual masking scheme and more finer textual cues to guide the reconstruction of masked segments within audio and video modalities, thereby enhancing local emotional consistency for each modality.

3.5. Overall learning objectives

Finally, the overall optimization objective of DLCEA consists of five loss functions, namely, the video-text contrastive learning loss L_{VTC} , the audio-text contrastive learning loss L_{ATC} , the text reconstruction loss L_{mlm} , the audio reconstruction loss L_{mam} , and the video reconstruction loss L_{mvm} . Specifically, L_{VTC} and L_{ATC} perform inter-modal global alignment between text and visual/audio modalities via contrastive learning, aligning the global emotional semantics across modalities. In contrast, L_{mlm} , L_{mam} , and L_{mvm} focus on intra-modal local alignment, enhancing the fine-grained emotional consistency within each modality through masked reconstruction. To balance the contributions of global inter-modal alignment and fine-grained intra-modal consistency, the final training objective is formulated as:

$$L = \lambda_1(L_{VTC} + L_{ATC}) + \lambda_2(L_{mlm} + L_{mam} + L_{mvm}), \quad (16)$$

where λ_1 and λ_2 are hyperparameters controlling the weights of the TGEA and TLEA modules, respectively. In our experiments, we explicitly set $\lambda_1 = \lambda_2 = 1.0$, as our sensitivity analysis (see Table 5) demonstrates that the model is robust to weight variations and achieves optimal performance with equal contribution from both modules.

3.6. Inference for UMER

Following the conventional practice [20], we design the inference pipeline shown in Fig. 4 for the final emotion recognition. Given the j -th prediction sample $\{T_j, V_j, A_j\}$ as input, we first extract the multimodal features (F_{T_j} , F_{V_j} and F_{A_j}) from the trained projection head of each modality. Then, we concatenate them into a 3×512 -dimensional combined vector, forming a unified representation denoted as $F_j \in \mathbb{R}^{3 \times 512}$. The process can be formulated as:

$$F_j = Concat(F_{T_j}, F_{V_j}, F_{A_j}). \quad (17)$$

This vector is then fed into a linear classifier with the structure of $fc + \text{RELU} + fc$ to output the final emotion category. fc is a fully-connected layer.

4. Experiments

In this section, we provide a comprehensive overview of the experimental settings, overall performance, and the analysis of the ablation experiments.

4.1. Experimental settings

4.1.1. Datasets

To thoroughly evaluate our approach, we used three common multimodal sentiment datasets CMU-MOSEI [21], MAFW [22], and IEMOCAP [23].

CMU-MOSEI [21]: is the largest multimodal dataset in the MER domain. It consists of over 23K discourses with more than 1000 speakers, collected from YouTube videos. Each discourse is labelled with six emotions: happiness, sadness, anger, fear, disgust and surprise, and each emotion is represented on a [0,3] Likert scale. Following the dissertation, emotions were treated as either present or absent (i.e., binary categorisation), whereas multiple emotions could be present at the same time, making the task a multi-label problem. The dataset contains approximately 3000 incorrectly aligned sequences. Since our approach does not require strict data alignment, we used all sequences provided in the CMU-MOSEI SDK. In other words, we did not apply any data cleaning. We also used the recommended metric for dataset segmentation and evaluation, weighted accuracy (WAR).

MAFW [22]: is the first large, multimodal, multilabel emotion dataset containing 11 single emotion categories, 32 multiple emotion categories, and descriptive emotion texts. The database contains 10,045 video-audio clips from natural scenes, providing bilingual descriptive

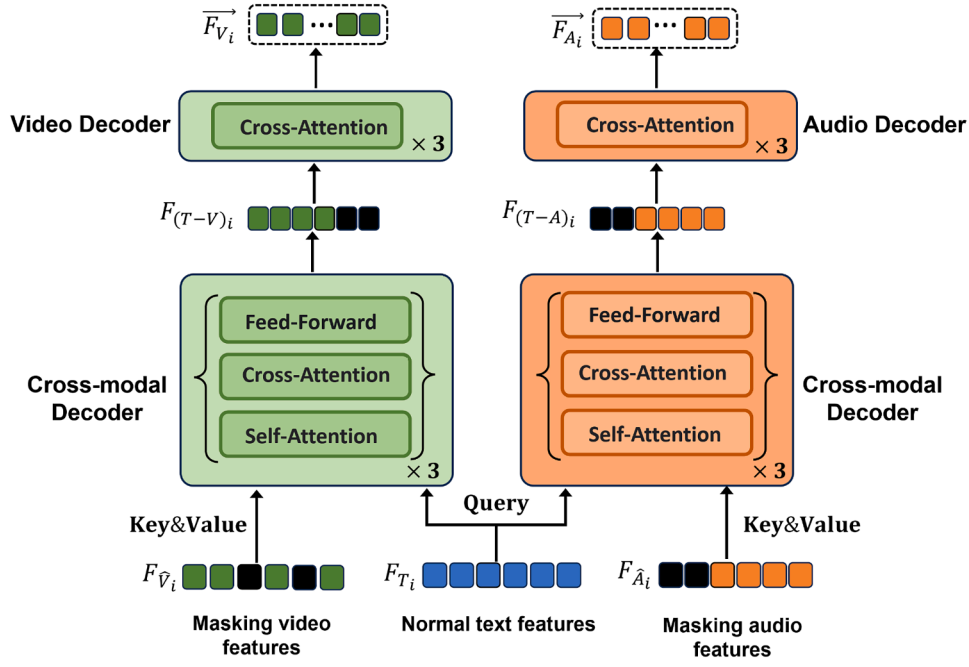


Fig. 3. Text-guided video and audio reconstruction pipeline in TLEA.

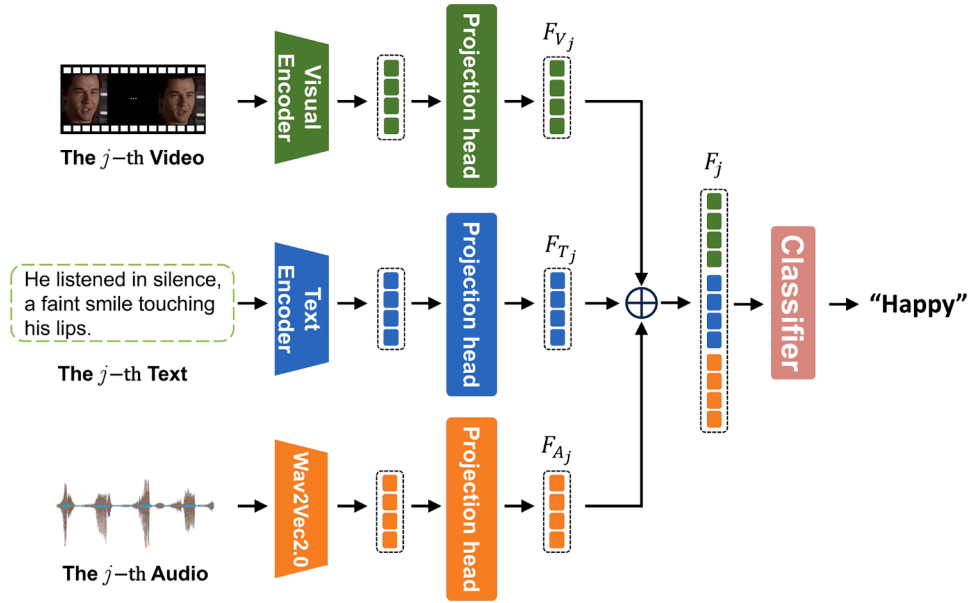


Fig. 4. The inference pipeline for DLCEA framework. The multimodal features of video, text, and audio are fused through concatenation and then input into a linear classifier for emotion recognition.

texts about facial expressions and emotions for videos. We selected its 11 single emotion categories and divided the training and test sets according to 4:1, with 4088 videos in the training set and 1022 videos in the test set.

IEMOCAP [23]: is a multimodal and multispeaker dataset that was collected at the SAIL lab at the University of Southern California (USC). It is designed to facilitate the study and modeling of multimodal and expressive human communication, particularly in the context of emotional expressions.

4.1.2. Experimental details

All models were trained using the SGD optimizer with a momentum of 0.9 and a weight decay of 0.001. The batch size was set to 32 for all experiments. The initial learning rate was 0.001, and a linear learning rate scheduler was employed to adjust the learning rate during training. We do not apply any “spatial” data enhancements (random crops, blurring, or colour distortion), but the data samples can all be trained using random sequences but the same video judgments. For example, video clips from $[t, t+10]$ and video clips from $[t+5, t+15]$ can be used for

Table 1

Compared with other methods on the CMU-MOSEI dataset. The best results for unsupervised methods are in bold.

	Methods	Happy	Sad	Anger	Surprise	Disgust	Fear	WAR	F1
Supervised	FE2E [25]	62.70	54.40	59.60	50.60	66.00	52.90	57.70	76.70
	MESM [14]	64.10	63.00	66.80	65.70	75.60	65.80	66.80	68.80
	CIA [26]	51.90	61.80	67.40	58.20	74.10	63.90	62.88	79.08
	MuT [27]	67.20	64.00	64.90	61.40	71.60	62.90	65.40	70.20
Unsupervised	CAE-LR [28]	64.70	53.20	61.80	57.10	69.00	60.40	61.03	65.52
	CMCR [29]	66.84	55.34	61.94	59.24	71.14	62.54	62.85	66.06
	MM-DFN [30]	-	-	-	-	-	-	63.77	67.13
	MPUC [6]	65.44	57.55	62.53	67.55	69.53	62.11	64.12	68.21
	MKIN-MCL [31]	45.03	70.40	74.62	61.64	63.70	62.14	63.23	66.92
	Qwen2.5-Omni [32]	-	-	-	-	-	-	61.47	65.61
	Qwen3 [33]	71.45	22.65	16.17	25.94	49.50	24.53	51.02	53.49
	Gemma3 [34]	36.98	29.76	25.92	23.96	37.59	33.95	32.73	38.40
	MGAFR [24]	76.25	1.62	16.87	5.15	55.07	2.59	50.62	48.04
	Ours	65.58	58.83	63.92	64.21	70.01	69.31	65.31	69.50

the same training. This is referred to as temporal dimension augmentation. We set the number of epochs to 400, taking the temperature scalar to be 0.07.

For fair reproduction of MGAFR [24], we applied only minimal adjustments: the continuous emotion scores in the three datasets were converted into discrete emotion labels, and a classification head was added to MGAFR's output to enable categorical prediction and metric computation.

4.2. Overall performance

4.2.1. Evaluation on CMU-MOSEI

We evaluated the proposed DLCEA framework on the CMU-MOSEI dataset and compared it with both supervised and unsupervised state-of-the-art methods.

As shown in Table 1, DLCEA achieves the best performance among all unsupervised approaches in the key emotion category of Fear, with improvements of 1.28% and 6.77% in Sad and Fear, respectively, over the second-best methods. While the scores for Surprise and Disgust were slightly lower than the best-performing models, DLCEA still maintained competitive performance. Furthermore, we significantly outperform the recently proposed method, MGAFR [24]. Compared to supervised methods, while DLCEA did not comprehensively outperform them in both WAR and F1, its overall performance was close. Notably, DLCEA surpassed FE2E [25] by 7.61% in WAR and MESM [14] by 0.7% in F1, demonstrating its competitiveness. Overall, DLCEA attained the highest WAR (65.31%) and F1-score (69.5%) among all unsupervised methods, showing better generalization and robustness across emotion categories.

We also included three large pre-trained multimodal models, Qwen2.5-Omni [32], Qwen3 [33] and Gemma3 [34], in the comparison. Despite its scale, it underperformed compared to DLCEA in this task. This result indicated that general-purpose models may not be optimal for fine-grained emotion recognition without task-specific adaptation, whereas DLCEA was better suited for capturing emotional cues through targeted multimodal representation learning.

4.2.2. Evaluation on MAFW

We evaluated the proposed DLCEA framework on the MAFW dataset and compare it with recent state-of-the-art supervised and unsupervised multimodal emotion recognition methods.

As shown in Table 2, our method achieved the best overall performance among unsupervised approaches, with a WAR of 54.19% and an F1 score of 57.43%, surpassing MPUC [6] by 1.04% and 0.12%, respectively. Similarly, we significantly outperform the existing method MAGFR [24]. Compared with the supervised method, although there is still a gap in the overall performance of DLCEA, it is significantly close. For example, its WAR metric is 0.34% higher than that of the supervised method FE2E [25], although its F1 is slightly lower than the latter. This

demonstrates the competitiveness and effectiveness of DLCEA. Notably, DLCEA also outperformed the large model Qwen2.5-Omni [32], Qwen3 [33] and Gemma3 [34], demonstrating the advantage of targeted multimodal alignment over general large-scale pretraining.

Our method achieved the highest accuracy on the “Anger” category, which often involve distinctive multimodal signals. This suggested that the dual-level alignment strategy in DLCEA effectively captures subtle emotional cues, enhancing robustness and generalization in unsupervised settings.

4.2.3. Evaluation on IEMOCAP

We evaluated the proposed DLCEA framework on the IEMOCAP dataset against both supervised and unsupervised baselines as shown in Table 3. Our method achieved a WAR of 71.3% and an F1 score of 76.6%, outperforming the strongest unsupervised competitor MPUC [6], which obtained 69.8% WAR and 73.2% F1, by 1.5% and 3.4% respectively, and still surpassing the recent MGAFR [24] method by a notable margin. DLCEA outperforms the supervised method TGIM [36] by 4.51% in F1, while its WAR performance is very close to that of TGIM [36] (only 0.28% lower), fully demonstrating its superior performance over supervised methods. Furthermore, DLCEA outperforms large foundation models such as Qwen2.5-Omni [32], Qwen3 [33] and Gemma3 [34], highlighting the advantage of task-specific multimodal alignment over general-purpose pretraining.

Notably, DLCEA attained the highest recognition rates on the “Excited” category with 80.5% accuracy and the “Frustrated” category with 81.7% accuracy, matching or even exceeding some supervised methods. This was attributed to its effective disentanglement of distinctive audiovisual cues such as prosodic shifts and facial muscle movements through a dual-level alignment strategy. These results indicated that DLCEA significantly narrowed the performance gap with supervised methods while maintaining robust generalization in unsupervised multimodal emotion recognition.

4.3. Ablation experiments and analysis

4.3.1. Impact of different modules in DLCEA

Table 4 presents the impact of progressively incorporating the TGEA module and the TLEA module into the baseline model (using only the pre-trained foundation models and Concatenation fusion). Specifically, TGEA comprised L_{VTC} and L_{ATC} , while TLEA included L_{mlm} , L_{mam} , and L_{mvm} .

The experimental results demonstrated a consistent improvement in both WAR and F1 scores as additional modules were introduced. For instance, the baseline model alone achieved WAR and F1 scores of 48.47% and 52.17%, respectively. With the incorporation of the L_{VTC} loss from TGEA, these metrics improved to 50.13% and 52.94%. Further enhancements were observed when adding the L_{ATC} loss and the L_{mlm} loss, resulting in WAR and F1 scores of 50.79% and 53.07%, respectively.

Table 2

Compared with other methods on the MAFW dataset. The best results for unsupervised methods are in bold.

	Methods	Happy	Sad	Anger	Surprise	Disgust	Fear	WAR	F1
Supervised	MuT [27]	57.25	54.30	54.84	51.31	61.55	52.76	55.33	63.20
	FE2E [25]	53.40	52.80	58.30	49.50	60.40	48.70	53.85	60.21
	MESM [14]	55.13	57.82	61.33	58.51	70.21	61.29	60.71	63.11
	EmoEmbs [35]	52.11	63.38	48.71	56.12	55.38	51.36	54.51	59.80
Unsupervised	CAE-LR [28]	54.70	43.20	51.80	47.10	59.00	50.40	51.03	54.72
	CMCR [29]	54.30	49.41	52.53	54.39	56.39	50.91	52.86	57.11
	MM-DFN [30]	-	-	-	-	-	-	53.15	57.31
	MPUC [6]	54.78	50.11	53.63	52.91	57.32	53.12	53.15	57.31
	MKIN-MCL [31]	44.09	52.50	57.26	49.45	62.95	54.29	53.42	57.04
	Qwen2.5-Omni [32]	-	-	-	-	-	-	45.32	49.72
	Qwen3 [33]	87.85	90.33	57.59	68.57	6.11	59.36	53.25	48.80
	Gemma3 [34]	52.85	85.17	46.11	59.25	5.95	37.76	52.06	48.23
	MGAFFR [24]	28.91	1.32	24.27	12.31	58.94	3.16	30.83	27.47
	Ours	54.67	51.48	60.12	51.79	54.53	55.33	54.19	57.43

Table 3

Compared with other methods on the IEMOCAP dataset. The best results for unsupervised methods are in bold.

	Methods	Happy	Sad	Anger	Excited	Disgust	Frustrated	WAR	F1
Supervised	MESM [14]	67.92	72.81	68.47	63.29	77.11	67.90	69.58	79.08
	SLCT [8]	68.10	64.30	67.30	65.10	73.60	63.00	66.90	78.68
	TGIM [36]	61.00	65.00	68.30	80.50	73.00	81.70	71.58	72.09
	Joyful [37]	60.94	84.42	69.95	73.54	68.24	67.55	70.77	71.03
Unsupervised	CAE-LR [28]	63.44	57.55	62.53	57.55	67.53	59.11	61.31	67.77
	CMCR [29]	63.50	62.40	66.20	65.10	75.00	65.87	66.35	70.10
	MM-DFN [30]	-	-	-	-	-	-	68.21	68.18
	MPUC [6]	67.10	66.00	69.80	68.70	78.60	68.80	69.80	73.20
	MKIN-MCL [31]	50.00	78.17	82.85	68.44	70.73	69.00	70.21	72.59
	Qwen2.5-Omni [32]	-	-	-	-	-	-	64.31	69.29
	Qwen3 [33]	53.11	15.96	2.08	16.52	56.32	52.57	35.47	31.44
	Gemma3 [34]	14.02	32.23	9.44	37.08	42.58	64.30	38.47	36.10
	MGAFFR [24]	29.77	70.30	59.51	61.73	55.60	56.84	57.30	56.95
	Ours	61.00	65.00	68.30	80.50	71.20	81.70	71.30	76.60

Table 4

Effects of different modules in our method.

Baseline	TGEA		TLEA			WAR	F1
	L_{VTC}	L_{ATC}	L_{mlm}	L_{mam}	L_{mcm}		
✓						48.47	52.17
✓	✓					50.13	52.94
✓	✓	✓				50.79	53.07
✓	✓	✓	✓			52.17	54.69
✓	✓	✓	✓	✓		53.34	55.19
✓	✓	✓	✓	✓	✓	54.19	57.43

Table 5

Sensitivity analysis of loss weights on MAFW.

λ_1	λ_2	WAR	F1
0.1	1.0	52.85	55.60
0.5	1.0	53.57	56.80
1.0	1.0	54.19	57.43
1.0	0.5	52.95	56.39
1.0	0.1	52.10	56.11

In addition, as shown in Table 5, we conducted a sensitivity analysis of loss weights on MAFW. The results indicate that the balanced configuration achieves optimal performance. Reducing any weight value (e.g., to 0.5 or 0.1) degrades overall performance, whereas weights approaching 1.0 better demonstrate the model's robustness.

4.3.2. Effects of varied masking strategies in noise-masked feature extraction

We evaluated the impact of different masking strategies in noise-masked feature extraction, as shown in Fig. 5. The orange denotes ran-

dom masking with varying proportions across modalities, while the blue indicates the emotion-aware masking strategy (e.g., masking emotional words or facial expressions). We also examined the effect of different masking ratios in this figure. Results show that emotion-relevant masking consistently yields the best performance, highlighting the importance of reconstructing emotional components to enhance model capability.

4.3.3. Comparative analysis of various foundation models for multimodal feature extraction

To evaluate the impact of different foundation models on our unsupervised multimodal emotion representation learning method with text-driven modality-hierarchical alignment, we compare the performance of four CLIP improved models (including Vanilla CLIP [1], ViFi-CLIP [38], Open-VCLIP [39], and ActionCLIP [40]) on the emotion recognition task. The results are shown in Table 6. Among them, ActionCLIP [40] performs best, leading in weighted accuracy (WAR) and F1 scores on the three datasets, especially on IEMOCAP, where it achieves 71.83% WAR and 76.89% F1, highlighting its excellent cross-modal association ability. Open-VCLIP [39] also achieved 65.77% WAR and 69.83% F1 on CMU-MOSEI, slightly outperforming other models. Experiments show that our method can effectively improve the performance of various foundation models and show good robustness and generalization.

4.3.4. Impact of feature guidance from different modalities in TGEA and TLEA

To fully validate our text-guided hypothesis, we analyzed the impact of different guidance modalities on the global alignment of the TGEA module. As shown in Table 7, the text-guided approach consistently outperforms the video-guided and audio-guided counterparts across the three benchmark datasets. Specifically, it improves the WAR metric by 1.99% and 0.84% on CMU-MOSEI, 2.67% and 0.38% on MAFW, and

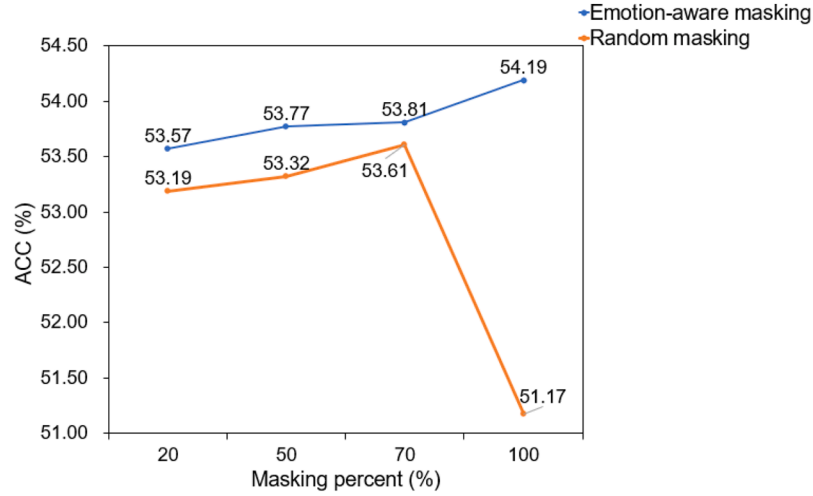


Fig. 5. Different masking strategies in TLEA. The *orange* denotes random masking with varying proportions across modalities, while the *blue* indicates emotion-aware masking. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6

Comparison of various foundation model for feature extraction.

Various CLIPs	CMU-MOSEI		MAFW		IEMOCAP	
	WAR	F1	WAR	F1	WAR	F1
CLIP [1]	65.31	69.5	54.19	57.43	71.30	76.60
ViFi-CLIP [38]	65.42	68.4	54.92	57.66	71.54	76.83
Open-VCLIP [39]	65.77	69.83	54.23	57.52	71.32	76.31
ActionCLIP [40]	65.93	69.58	55.43	57.73	71.83	76.89

Table 7

Compare the influence of using different modalities for guidance in TGEA.

Different modalities	CMU-MOSEI		MAFW		IEMOCAP	
	WAR	F1	WAR	F1	WAR	F1
Video-guided	63.32	67.82	51.52	54.94	70.37	74.83
Audio-guided	64.47	67.51	53.81	55.73	68.93	73.86
Text-guided	65.31	69.50	54.19	57.43	71.30	76.60

Table 8

Compare the influence of using different modalities for guidance in TLEA.

Different modalities	CMU-MOSEI		MAFW		IEMOCAP	
	WAR	F1	WAR	F1	WAR	F1
NO modality	61.54	63.79	52.97	54.02	69.31	73.20
Video-guided	64.61	65.75	52.91	53.91	71.32	75.93
Audio-guided	62.97	65.26	52.33	54.18	70.31	75.72
Text-guided	65.31	69.50	54.19	57.43	71.30	76.60

0.93% and 2.37% on IEMOCAP, respectively. This confirms that text is the most effective anchor for establishing global cross-modal emotional consistency.

We further evaluate the influence of different guidance modalities within the TLEA module for fine-grained local alignment, with the results summarized in Table 8. Consistent with the global alignment results, text-guided reconstruction achieves superior performance in terms of both WAR and F1 scores across all datasets. Notably, on CMU-MOSEI, text-guided TLEA achieves a significant improvement (e.g., 69.50% F1 vs. 65.75% for video-guided and 65.26% for audio-guided). This trend holds on MAFW and IEMOCAP as well, underscoring the advantage of leveraging the richer and more explicit emotional cues embedded in the textual modality. These results validate the effectiveness of using text as a guiding signal for fine-grained intra-modal emotional alignment.

These findings align with prior work, confirming that text typically encodes richer emotional cues. The strong performance of the text modality highlights its superior capability in complex emotion recognition. In contrast, video and audio modalities show weaker performance on certain datasets, likely due to challenges in capturing subtle emotional variations.

4.4. Qualitative analysis and visualization of modal alignment

4.4.1. Visualization of cross-modal global emotion alignment

Fig. 6 presents the t-SNE visualizations of the learned feature distributions from video, text, and audio modalities, using samples labeled as Anger and Happy from the MAFW dataset.

In Fig. 6(a), the baseline model shows dispersed and weakly aligned features across the three modalities under both emotion categories. This suggests that relying solely on the pre-trained foundation models is insufficient for effective cross-modal alignment. Fig. 6(b) incorporates the Text-guided Cross-modal Global Emotion Alignment (TGEA) module, resulting in more compact and cohesive feature distributions. This indicates that TGEA contributes positively to enhancing global emotional alignment across modalities. In Fig. 6(c), the further integration of the Text-guided Cross-modal Local Emotion Alignment (TLEA) module leads to even tighter clustering of features. This demonstrates that TLEA further strengthens intra-modal consistency and helps the model distinguish emotional states more effectively.

Overall, the DLCEA framework, which integrates both TGEA and TLEA, significantly improves cross-modal alignment, resulting in clearer and more concentrated emotion-specific feature distributions. These results highlight the effectiveness of the proposed method in enhancing the performance of unsupervised multimodal emotion recognition.

4.4.2. Visualization of cross-modal local emotion alignment

To intuitively validate the effectiveness of cross-modal local emotion alignment, we analyzed the cross-modal attention map on the MAFW dataset, as illustrated in Fig. 7. The model accurately aligns emotion-related words (e.g., “frown”) with the corresponding video frames. Specifically, lower attention is assigned to the first two frames with limited emotional relevance, while significantly higher attention is focused on the final frames depicting the frowning expression. These results indicate that our method enables fine-grained alignment between text and video modalities, facilitating unified cross-modal emotional representation and contributing to improved UMER performance.

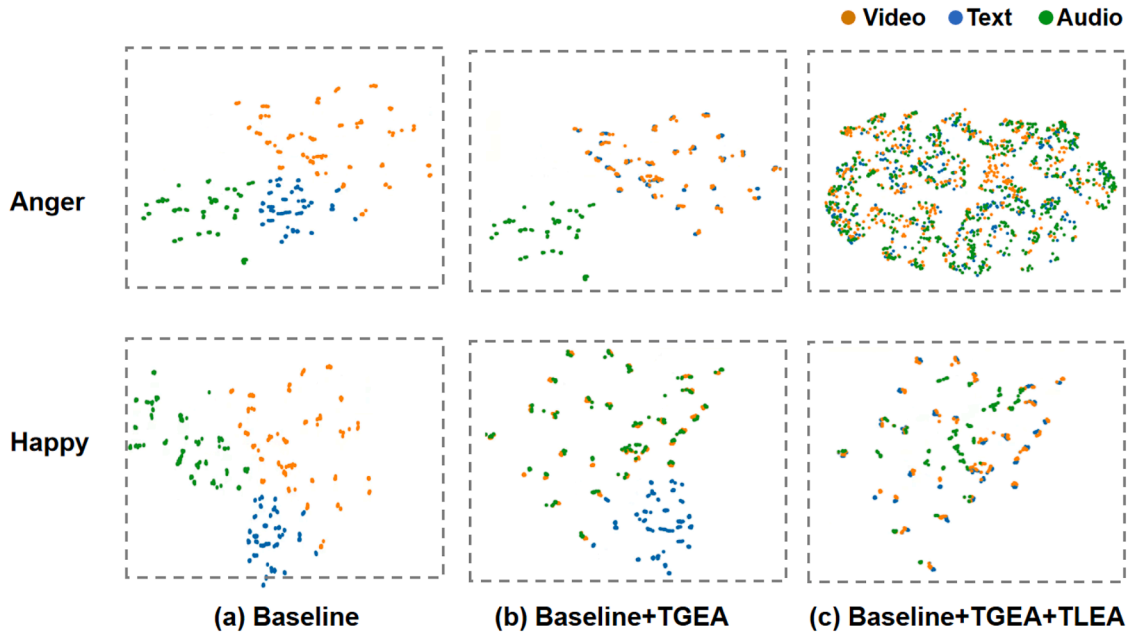


Fig. 6. t-SNE visualizations of video, text, and audio features for “Anger” and “Happy” emotions on the MAFW dataset. Better cross-modal alignment is shown by tighter, overlapping clusters of different modalities. (a) The baseline shows poor alignment with separated modality clusters. (b) The addition of the TGEA module improves alignment, resulting in more compact cross-modal clusters. (c) Our full model, DLCEA (Baseline + TGEA + TLEA), demonstrates the best performance, with features from all three modalities being highly coherent and tightly clustered, signifying superior feature alignment.

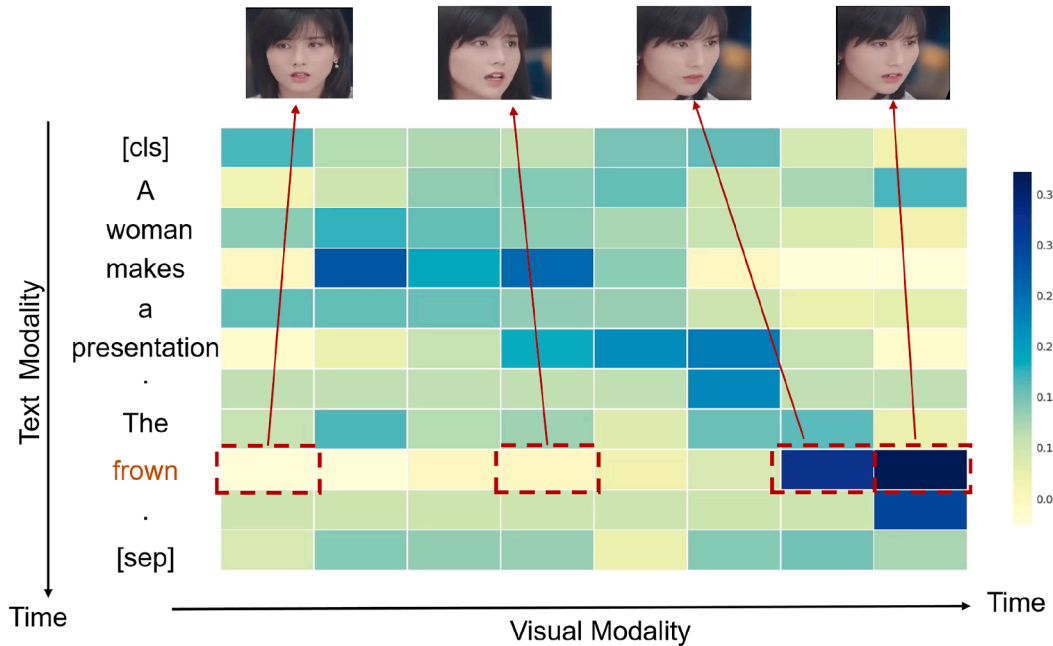


Fig. 7. Cross-model attention map visualization for local emotion alignment on MAFW. Bluer regions indicate higher model focus, while greener regions indicate lower attention.

4.4.3. Visualization of emotion-unified multimodal representation

In Fig. 8, we employed t-SNE visualization to illustrate the fused multimodal feature distributions of CAE-LR [28], CMCR [29], MPUC [6], and our proposed method, on the MAFW dataset. The results revealed that the proposed method achieved more distinct clustering among different emotion categories compared to the other approaches, demonstrating its capability to extract more robust emotional features for UMER.

4.4.4. Visualization of emotion recognition results

Fig. 9 shows a visualization of the results of different representation learning methods in an emotion recognition task. Each sample contained a series of consecutive image frames and the emotion recognition results of the different methods for these image frames. The figure listed three different emotion recognition methods: the method of this paper, MPUC [6], and CMCR [29], and compared the recognition results of each method.

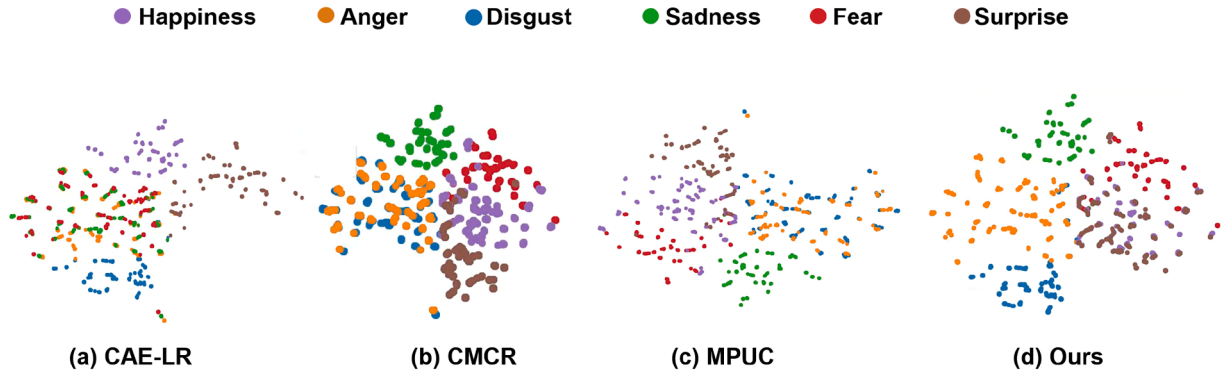


Fig. 8. Comparative visualization of emotional feature distributions across different methods: (a) CAE-LR, (b) CMCR, (c) MPUC, and (d) our proposed method. Each color represents a distinct emotion category, including Happiness, Anger, Disgust, Sadness, Fear, and Surprise. Greater separation between emotion classes indicates better discriminative performance.

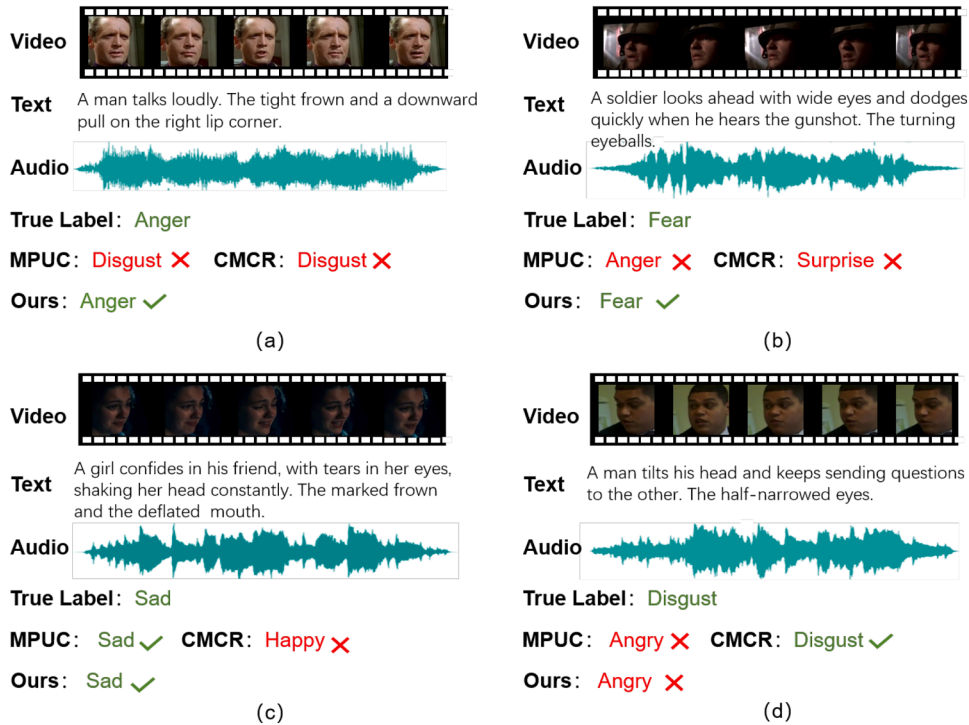


Fig. 9. Comparison of emotion recognition results using different methods. Each example includes video, text, and audio inputs, with ground-truth emotion labels. Predictions from MPUC, CMCR, and the proposed method are shown. The proposed method demonstrates higher prediction accuracy across multiple cases.

In Fig. 9(a), the methods in this paper correctly identified the emotion as “anger”, while both MPUC [6] and CMCR [29] incorrectly identified it as “disgust”. In Fig. 9(b), the methods in this paper correctly identified the emotion as “fear”, but MPUC [6] incorrectly identified it as “anger” and CMCR [29] incorrectly identified it as “surprise”. In Fig. 9(c), both the methods in this paper and MPUC [6] correctly identified the emotion as “sadness”, while CMCR [29] incorrectly identified it as “happiness”. In Fig. 9(d), the method in this paper incorrectly identified the emotion as “anger”, but CMCR [29] correctly identified the emotion as “disgust”, and MPUC [6] also incorrectly identified the emotion as “anger”.

These results indicated that the method proposed in this paper could accurately recognize emotions in most cases, especially in recognizing the emotions “anger” and “fear”. However, in some cases, such as Fig. 9(d), the method in this paper misrecognized the emotion, whereas the CMCR [29] method was able to recognize it correctly. This suggested that different representation learning methods might have their own strengths and limitations when dealing with specific emotions. Through

this visual analysis, we gained a deeper understanding of the performance of different methods in emotion recognition tasks and provided references for further optimization and improvement.

5. Conclusion

This paper presents an unsupervised multimodal emotion recognition framework that tackles the challenge of unaligned emotional representations across modalities. The proposed framework introduces a dual-level emotion semantic alignment scheme: Text-guided Cross-modal Global Emotion Alignment (TGEA) and Text-guided Cross-modal Local Emotion Alignment (TLEA), which collaboratively achieve emotional alignment at complementary global and local levels. Specifically, TGEA leverages the rich semantics of the textual modality as a cross-modal anchor and employs contrastive learning to align global affective features of audio and visual modalities with those of text, ensuring semantic consistency across modalities. In parallel, TLEA enhances fine-grained, local emotional understanding by reconstructing masked seg-

ments within text, audio, and video, using emotion-aware masking strategy and textual cues to guide local emotion semantic refinement. This two-tiered alignment strategy enables the model to capture an effective emotion-unified representation with both coarse and subtle emotional cues in a fully unsupervised manner.

Extensive experiments on benchmark datasets demonstrate that our method consistently outperforms existing unsupervised approaches and recent large-scale pre-trained multimodal models, highlighting its effectiveness and robustness for cross-modal emotion alignment and representation. However, the method may still be limited by its reliance on text as the primary semantic anchor, which can introduce biases or reduced effectiveness in scenarios where textual cues are sparse or noisy. In future work, we will explore adaptive or modality-agnostic anchoring strategies that dynamically adjust the reliance on text and better handle cases with sparse or unreliable textual cues.

CRedit authorship contribution statement

Shaoze Feng: Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization; **Qiyin Zhou:** Writing – review & editing, Visualization, Investigation; **Yuanyuan Liu:** Writing – review & editing, Supervision, Conceptualization; **Ke Wang:** Writing – review & editing, Visualization; **Kejun Liu:** Writing – review & editing, Investigation; **Chang Tang:** Writing – review & editing, Supervision.

Data availability

I have shared the link to my code at the Attach File step.

Declaration of competing interest

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

Acknowledgements

This work was supported by the National Natural Science Foundation of China grant (62076227), Natural Science Foundation of Hubei Province grant (2023AFB572) and Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2022-B10).

References

- [1] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
- [2] H.-H. Wu, P. Seetharaman, K. Kumar, J.P. Bello, Wav2CLIP: learning robust audio representations from clip, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 4563–4567.
- [3] Y. Liu, S. Feng, S. Liu, Y. Zhan, D. Tao, Z. Chen, Z. Chen, Sample-cohesive pose-aware contrastive facial representation learning, *Int. J. Comput. Vis.* (2025) 1–19.
- [4] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T.-S. Chua, B.W. Schuller, K. Keutzer, Affective image content analysis: two decades review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (10) (2021) 6729–6751.
- [5] K. Liu, Y. Liu, L. Wei, C. Tang, Y. Zhan, Z. Chen, Z. Chen, Smile on the face, sadness in the eyes: bridging the emotion gap with a multimodal dataset of eye and facial behaviors, *IEEE Trans. Multimed.* (2025).
- [6] R. Franceschini, E. Fini, C. Beyan, A. Conti, F. Arrigoni, E. Ricci, Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 2589–2596.
- [7] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, 2019, NIH Public Access, 2019, p. 6558.
- [8] A. Khare, S. Parthasarathy, S. Sundaram, Self-supervised learning with cross-modal transformers for emotion recognition, in: 2021 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2021, pp. 381–388.
- [9] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases, *Pattern Recognit.* 44 (3) (2011) 572–587.
- [10] S. Zhao, X. Hong, J. Yang, Y. Zhao, G. Ding, Toward label-efficient emotion and sentiment analysis, *Proc. IEEE* 111 (10) (2023) 1159–1197.
- [11] G. Xiang, S. Yao, X. Wu, H. Deng, G. Wang, Y. Liu, F. Li, Y. Peng, Driver multi-task emotion recognition network based on multi-modal facial video analysis, *Pattern Recognit.* 161 (2025) 111241.
- [12] E. Ghaleb, M. Popa, S. Asteriadis, Multimodal and temporal perception of audio-visual cues for emotion recognition, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2019, pp. 552–558.
- [13] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 1359–1367.
- [14] W. Dai, S. Cahyawijaya, Z. Liu, P. Fung, Multimodal end-to-end sparse model for emotion recognition, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 5305–5316. <https://doi.org/10.18653/v1/2021.naacl-main.417>
- [15] C. Gan, J. Zheng, Q. Zhu, Y. Cao, Y. Zhu, A survey of dialogic emotion analysis: developments, approaches and perspectives, *Pattern Recognit.* 156 (2024) 110794.
- [16] X. Liu, T. Chen, J. Zhou, H. Wang, G. Liu, X. Fu, Human emotion and StO2: dataset, pattern, and recognition of basic emotions, *Pattern Recognit.* 161 (2025) 111249.
- [17] S. Zhao, H. Yao, Y. Gao, G. Ding, T.-S. Chua, Predicting personalized image emotion perceptions in social networks, *IEEE Trans. Affect. Comput.* 9 (4) (2016) 526–540.
- [18] Y. Liu, L. Wei, K. Liu, Z. Chen, Z. Chen, C. Tang, J. Chen, S. Shan, Leveraging eye movement for instructing robust video-based facial expression recognition, *IEEE Trans. Affect. Comput.* 16 (4) (2025) 3404–3420. <https://doi.org/10.1109/TAFFC.2025.3599859>
- [19] C. Wang, J. Wang, Z. Gu, J.-M. Wei, J. Liu, Unsupervised feature selection by learning exponential weights, *Pattern Recognit.* 148 (2024) 110183.
- [20] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
- [21] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.
- [22] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, S. Shan, MAFW: a large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 24–32.
- [23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, *Lang. Resour. Eval.* 42 (4) (2008) 335–359.
- [24] Y. Deng, J. Bian, S. Wu, J. Lai, X. Xie, Multiplex graph aggregation and feature refinement for unsupervised incomplete multimodal emotion recognition, *Inf. Fusion* 114 (2025) 102711.
- [25] J.-B. Delbrouck, N. Tits, M. Brousmiche, S. Dupont, A transformer-based joint-encoding for emotion recognition and sentiment analysis, in: A. Zadeh, L.-P. Morency, P.P. Liang, S. Poria (Eds.), Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML), Association for Computational Linguistics, Seattle, USA, 2020, pp. 1–7. <https://doi.org/10.18653/v1/2020.challengehml-1.1>
- [26] D.S. Chauhan, M.S. Akhtar, A. Ekbal, P. Bhattacharyya, Context-aware interactive attention for multi-modal sentiment and emotion analysis, 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), 2019, pp. 5647–5657.
- [27] W. Dai, Z. Liu, T. Yu, P. Fung, Modality-transferable emotion embeddings for low-resource multimodal emotion recognition, in: K.-F. Wong, K. Knight, H. Wu (Eds.), Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 269–280. <https://doi.org/10.18653/v1/2020.aacl-main.30>
- [28] P. Koromilas, T. Giannakopoulos, Unsupervised multimodal language representations using convolutional autoencoders, (2021). [arXiv:2110.03007](https://arxiv.org/abs/2110.03007)
- [29] Z. Wang, Y. Zhao, H. Huang, J. Liu, A. Yin, L. Tang, L. Li, Y. Wang, Z. Zhang, Z. Zhao, Connecting multi-modal contrastive representations, *Adv. Neural Inf. Process. Syst.* 36 (2023) 22099–22114.
- [30] D. Hu, X. Hou, L. Wei, L. Jiang, Y. Mo, MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 7037–7041.
- [31] X. Shen, X. Huang, S. Zou, X. Gan, Multimodal knowledge-enhanced interactive network with mixed contrastive learning for emotion recognition in conversation, *Neurocomputing* 582 (2024) 127550. <https://doi.org/10.1016/j.neucom.2024.127550>
- [32] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, et al., Qwen2.5-omni technical report, (2025). [arXiv:2503.20215](https://arxiv.org/abs/2503.20215)
- [33] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, (2025). [arXiv:2505.09388](https://arxiv.org/abs/2505.09388)
- [34] A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, et al., Gemma 3 technical report, *CoRR* abs/2503.19786 (2025).

- [35] M.S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, P. Bhattacharyya, Multi-task learning for multi-modal emotion recognition and sentiment analysis, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 370–379.
- [36] Y.-H.H. Tsai, M.Q. Ma, M. Yang, R. Salakhutdinov, L.-P. Morency, Multimodal routing: improving local and global interpretability of multimodal language analysis, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020, NIH Public Access, 2020, p. 1823.
- [37] D. Li, Y. Wang, K. Funakoshi, M. Okumura, Joyful: joint modality fusion and graph contrastive learning for multimodal emotion recognition, in: H. Bouamor, J. Pino, K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 2023, pp. 16051–16069. <https://doi.org/10.18653/v1/2023.emnlp-main.996>
- [38] H. Rasheed, M.U. khattak, M. Maaz, S. Khan, F.S. Khan, Finetuned CLIP models are efficient video learners, in: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [39] Z. Weng, X. Yang, A. Li, Z. Wu, Y.-G. Jiang, Open-VCLIP: transforming CLIP to an open-vocabulary video model via interpolated weight optimization, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, 2023 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 36978–36989. <https://proceedings.mlr.press/v202/weng23b.html>.
- [40] M. Wang, J. Xing, J. Mei, Y. Liu, Y. Jiang, ActionCLIP: adapting language-image pretrained models for video action recognition, *IEEE Trans. Neural Netw. Learn. Syst.* (2023) 1–13. <https://doi.org/10.1109/TNNLS.2023.3331841>