Neurocomputing 711 (2027) 135484

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Emotional description-guided vision–language semantic alignment for group emotion recognition☆

Hongxi Yi[a], Dong Zhang[a,*], Yufei Xie[a], Wenxing Liu[b], Ming Li[c,d], Dah-Jye Lee[e]

[a] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China
[b] The School of Computer Science, Wuhan University, Wuhan, 430072, Hubei, China
[c] School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, 518172, Guangdong, China
[d] The Digital Innovation Research Center, Duke Kunshan University, Kunshan, 215316, Jiangsu, China
[e] Department of Electrical and Computer Engineering, Brigham Young University, Provo, USA

HIGHLIGHTS

- Propose a novel idea to refine GER as a vision-language alignment problem.
- Highlight the role of explicit emotional semantics in collective affect modeling.
- Propose a framework using VLM-generated group level emotional descriptions as semantic intermediates between images and emotion labels.
- Design an attention-enhanced, contrastively constrained architecture to integrate visual features with emotional language representations.

ARTICLE INFO

Communicated by Y. Long

Keywords:
Group emotion recognition
Emotional language modeling
Vision-language models
Contrastive learning

ABSTRACT

Group Emotion Recognition (GER) aims to infer the collective emotional state of a group from visual observations, which remains challenging due to visual ambiguity and insufficient high-level semantic modeling. Most existing methods rely solely on empirically defined visual cues and struggle to capture socially grounded and context-dependent emotional semantics. In this work, we propose ED-VLAlign, an emotional description-guided vision–language semantic alignment framework that explicitly incorporates emotional language modeling into group emotion understanding, reinterpreting GER as a vision-language alignment problem. Specifically, we leverage vision-language models to generate group-level emotional descriptions, which serve as semantic intermediates between images and emotion labels. These descriptions are aligned with visual representations using a CLIP-based architecture enhanced by multi-head cross-attention with residual fusion and contrastive learning constraints. Extensive experiments on four benchmark datasets—GAF2, GAF3, GroupEmoW, and the video-based VGAF benchmark—demonstrate that the proposed approach consistently outperforms state-of-the-art visual-only methods, validating the effectiveness and robustness of the proposed vision-language semantic alignment framework.

1. Introduction

Group Emotion Recognition (GER) aims to infer the overall emotional state of a group from visual inputs, such as images or videos depicting collective activities and social interactions [1]. Unlike individual emotion recognition, which identifies discrete emotions of a single person, GER captures collective affect emerging from interactions among multiple individuals within a shared context [2]. Due to its applications in public safety, crowd behavior analysis, social media sentiment understanding, and intelligent surveillance, GER has attracted growing research interest [3]. Group emotion is not a simple aggregation of individual emotions but a high-level construct involving group dynamics, social relations, and contextual cues [4]. Real-world GER is challenged by subtle facial expressions, varying group scales, and large intra-group appearance variations [5], making it a high-level semantic understanding problem rather than a direct extension of individual emotion analysis.

☆ This research is funded in part by the National Natural Science Foundation of China (62173353, 62171207), the Guangdong Science and Technology Plan (2023A1111120012), and the Science and Technology Program of Guangzhou, China (2023B03J1327).
* Corresponding author.
Email address: zhangd@mail.sysu.edu.cn (D. Zhang).

https://doi.org/10.1016/j.neucom.2026.135484
Received 30 June 2026; Received in revised form 17 September 2026; Accepted 24 October 2026
Available online 26 October 2026
0925-2312/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Existing GER methods are predominantly vision-driven. Early approaches focus on detecting and aggregating individual facial emotions [6], while later studies incorporate scene context and global appearance cues to capture environmental information correlated with group affect [7]. More recent methods employ structured modeling, such as graph-based representations and Transformer architectures, to capture interactions among individuals and context [8]. Although these approaches enhance representation by modeling spatial relations and hierarchical dependencies, GER research remains primarily vision-centric, relying on visual signals from images or videos.

Despite remarkable progress in recent years, vision-driven GER is limited by dependence on superficial cues, often restricted to faces, objects, and scenes. Other informative signals, such as body language, posture, and group engagement, are often overlooked. Psychological studies suggest these methods also lack access to culturally grounded and socially contextual emotional semantics, such as symbolic meanings of activities and social backgrounds. For example, dragons symbolize prosperity and celebration in Chinese culture but are often associated with mystery or destruction in Western cultures; white can likewise convey different emotional meanings in funerals and weddings across cultures. As a result, visually similar scenes with distinct emotional connotations across cultures may be misinterpreted [9]. High-level concepts like group relationships, collective behaviors, and atmosphere are difficult to capture through visual features alone, limiting semantic abstraction and interpretability [8]. Visual noise from complex scenes further degrades generalization in real-world settings [10].

Vision-language models (VLMs) can offer a promising solution. Trained on large-scale multimodal data, VLMs encode rich world knowledge and semantic priors, enabling explicit emotional language modeling from visual inputs [11]. They capture affective cues beyond faces and scenes, extract culturally grounded emotional semantics, and model high-level group concepts like collective behavior and atmosphere [12]. However, existing VLMs show limited zero-shot performance on GER due to domain gaps and coarse emotion alignment, highlighting the need for task-specific adaptation.

This paper proposes an emotional description-guided vision–language semantic alignment framework (ED-VLAlign) that explicitly integrates emotional language modeling for GER. This new framework employs VLMs to generate group-level emotional descriptions for each image or video clip, and to extract surface-level affective cues and culturally grounded semantics. These emotional descriptions serve as high-level intermediaries bridging visual content and emotion labels. By reformulating GER as an image/video–text alignment task, this new framework enhances generalization across diverse scenes and group compositions. Built on CLIP-based representations, our approach introduces multi-head cross-attention with residual fusion and contrastive learning to align visual features with emotional language representations. Experiments on three image benchmarks and one video benchmark show competitive performance and outperform the representative vision-only baselines used in our comparisons.

The main contributions of this work include:

- We introduce a novel idea to refine GER as a vision-language alignment problem, and highlight the role of explicit emotional semantics in collective affect modeling.
- We propose a framework using VLM-generated group-level emotional descriptions as semantic intermediates between images and emotion labels.
- We design an attention-enhanced, contrastively constrained architecture to integrate visual features with emotional language representations.
- Experimental results demonstrate consistent performance improvements across multiple benchmarks, validating the effectiveness and generalizability of the proposed method.

2. Related work

2.1. Group-level emotion recognition

Early studies on group-level emotion recognition primarily followed visually driven paradigms by extending individual emotion recognition techniques to groups [3]. A common approach estimates facial expressions of group members and aggregates them with scene-level context to infer overall group affect. To compensate for missing or unreliable facial information in crowded or unconstrained settings, scene context is often incorporated [5], including environmental attributes and background semantics. Hierarchical models further capture relationships between individual-level representations and group-level states, enabling coarse-to-fine reasoning from local faces to global scenes [13]. Region-based attention selectively focuses on emotionally salient individuals or spatial regions, suppressing irrelevant background noise and enhancing discriminative cues [14]. Despite their effectiveness, these approaches remain highly dependent on visual signal quality: faces are frequently occluded, blurred, or of insufficient resolution, and expressions can be subtle or ambiguous, limiting robustness in complex scenes.

To address these issues, recent works explore structured modeling and contextual fusion. Graph-based methods construct relational graphs among detected individuals, with nodes representing people and edges encoding spatial proximity or interactions, enabling direct modeling of inter-person relationships for collective emotion [8]. Transformer-based architectures use self-attention to model long-range dependencies and global context [15]. Multi-scale context modeling captures both fine-grained individual details and holistic group semantics [16], while uncertainty-aware learning improves robustness under noisy or ambiguous conditions [10]. These advances enhance modeling flexibility and representation capacity, improving benchmark performance.

Nevertheless, existing methods remain constrained by a reliance on visual information alone. Structured models such as graphs and Transformers improve contextual reasoning but operate entirely in a visual feature space, lacking explicit semantic supervision at the group emotion level [17]. Psychologically, group emotion involves collective intent, social atmosphere, and shared interpretation, which visual cues alone cannot fully capture [1]. Consequently, visual-only methods struggle with abstract group semantics and generalization across diverse social contexts.

Overall, group-level emotion recognition remains dominated by visual-centric paradigms. While progress has been made in modeling spatial structures, interactions, and context, gaps remain in explicit language-based semantic modeling. Group emotion is inherently a high-level cognitive construct extending beyond observable patterns. In contrast, our approach introduces language as an explicit semantic medium, leveraging affective representations from VLMs to capture high-level group emotional semantics that are inaccessible to purely visual methods.

2.2. Vision–language models for emotion analysis

Large language models (LLMs), such as the GPT-series, demonstrate strong capabilities in capturing high-level semantics and contextual reasoning from large-scale textual data [12]. Building on this, vision–language models (VLMs) extend LLMs to the multimodal domain by jointly learning aligned representations of images and text via large-scale pretraining. Representative VLMs, such as BLIP [18] and GLM [19], achieve cross-modal semantic alignment by linking visual patterns with natural language descriptions.

Compared with visual-only models, VLMs offer clear advantages for emotion analysis. Leveraging rich multimodal priors, they capture affective cues beyond faces and objects, including social roles and culturally grounded symbols. More importantly, VLMs enable explicit high-level semantic modeling of group behaviors, relationships, and collective atmosphere rather than relying solely on low-level

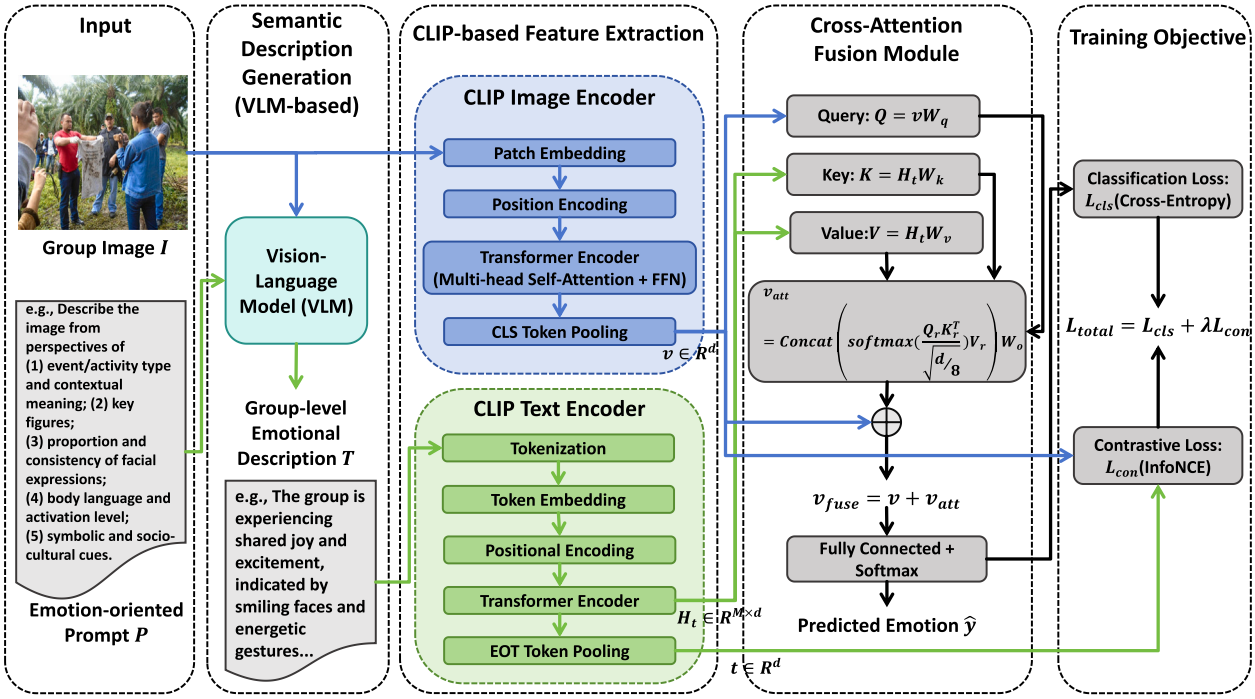


Fig. 1. Overview of the proposed ED-VLAlign framework for group emotion recognition with global-image-to-text-token cross-attention.

visual affect. Recent VLM studies demonstrate transferable cross-modal representations across diverse vision–language tasks [20,21].

Motivated by these properties, an increasing number of studies explores vision–language learning for emotion analysis. Early works use textual emotion labels or attributes as auxiliary supervision to enhance visual recognition [22]. More recent approaches adopt prompt-based strategies, where designed emotion prompts adapt pretrained VLMs to affective tasks with minimal retraining [12], activating emotional concepts through language for classification with limited labels. Beyond prompt engineering, some studies reformulate emotion recognition as cross-modal alignment by matching visual features with emotion-related descriptions or captions [23], shifting analysis from closed-set classification to flexible semantic matching and improving interpretability.

Recent Neurocomputing studies further support this direction from complementary perspectives. Pan et al. [24] systematically reviewed multimodal emotion recognition and emphasized that robust emotion understanding depends not only on feature extraction but also on how heterogeneous affective cues are aligned and fused. Li et al. [25] introduced EmoVerse to enhance multimodal large language models for affective computing through multitask learning, showing the growing importance of language-centered affective reasoning. Cai et al. [26] proposed prompt-based hybrid supervised contrastive learning for emotion recognition in conversation, demonstrating that prompt-derived emotional semantics and contrastive objectives can improve semantic discrimination under class imbalance. These works motivate our use of VLM-generated emotional descriptions as structured semantic supervision rather than treating language as a passive auxiliary modality.

Despite these advances, existing vision–language approaches face limitations for group emotion recognition. Most focus on individual or non-group scenarios, where emotional cues are explicit and localized [27]. In contrast, group emotion arises from collective interactions and shared social context, making it abstract and harder to model. Directly transferring individual-level methods often yields suboptimal results due to a lack of group-specific semantic modeling. Additionally, VLM predictions remain imperfect because models differ from humans in their emotion understanding and lack targeted sentiment training [23]. Finally, existing multimodal methods rarely integrate vision–language learning with structured group modeling, often treating images holistically without considering group composition or relational structure [15].

In contrast, this work explicitly targets group-level emotion recognition by using VLMs as semantic generators rather than feature extractors. Instead of predicting labels from images, VLMs generate group-level emotional descriptions encoding both surface affective cues and deeper semantic knowledge grounded in social and cultural context. These descriptions serve as intermediate representations bridging visual observations and emotion categories, enabling more interpretable and semantically grounded reasoning. By reformulating group emotion recognition as a vision–language alignment task and incorporating multi-head cross-attention with residual fusion and contrastive learning, our framework effectively exploits pretrained multimodal priors while addressing limitations in modeling collective affect.

3. Proposed method

In this work, we propose an emotional description-guided vision–language semantic alignment framework (ED-VLAlign) to explicitly integrate emotional language semantics into collective affect modeling (Fig. 1). Motivated by psychological and social theories on the semantic and contextual nature of collective emotions [1], our method extends beyond visual-centric paradigms by introducing language as a semantic bridge between group images and emotion labels.

Specifically, vision–language models (VLMs) generate group-level emotional descriptions from images, reformulating group emotion recognition (GER) as an image–text alignment problem rather than pure visual classification. Visual features and semantic representations derived from emotional language are jointly embedded into a shared space via a CLIP-based architecture. A multi-head cross-attention module uses the global visual representation to selectively aggregate contextual token-level emotional-language features, followed by residual fusion with the original visual embedding. A contrastive learning objective enforces consistency between aligned image–text pairs, improving robustness and generalization.

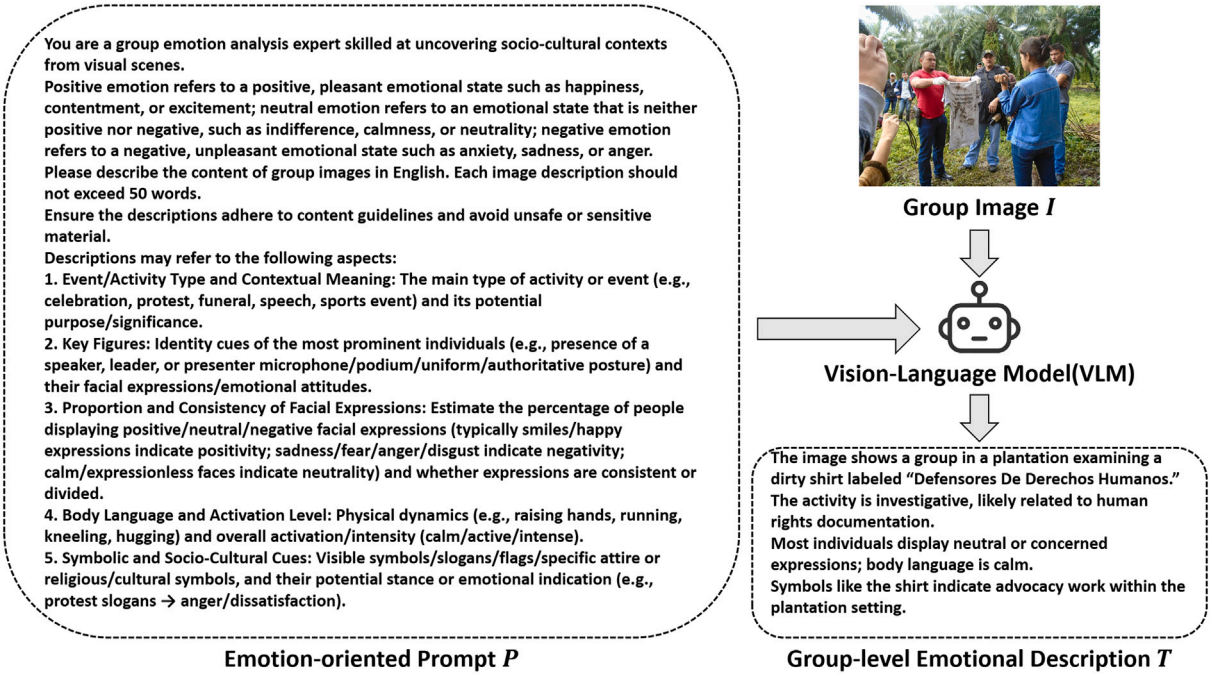


Fig. 2. Example of semantic description generation using the proposed emotion-oriented prompt.

3.1. Semantic description generation

From a psychological perspective, group emotion is not a simple aggregation of individual emotions but an emergent social construct shaped by shared appraisals, collective behavior, social roles, and situational meaning [1]. Social appraisal and collective emotion theories suggest that emotional alignment arises through interaction, shared goals, and common interpretation [4]. Modeling group emotion therefore requires capturing not only individual facial affect but also group-level dynamics such as joint attention, coordinated actions, symbolic activities, and socio-cultural context.

Recent advances in vision–language models (VLMs) demonstrate strong multimodal understanding and long-context reasoning, enabling semantically rich and controllable descriptions from complex scenes. This makes VLMs particularly suitable for group-level emotion understanding, where collective affect is often implicit, context-dependent, and culturally grounded.

Guided by these insights, we design a structured emotion-oriented prompt to guide VLMs to generate group-level semantic descriptions rather than generic captions (Fig. 2). The prompt defines emotion categories (positive, neutral, and negative), constrains the generation length, and induces the pretrained VLM to describe the group image from a psychological perspective. Specifically, it encourages the model to consider event and activity context, key figures and social roles, distribution and consistency of facial expressions, body language and activation level, and symbolic or socio-cultural cues. These aspects correspond to established determinants of collective affect and are discussed below. To make the semantic generation process reproducible, the prompt template used in all experiments is fixed rather than manually adjusted per image.

For reproducibility, all experiments use a fixed prompt the complete content of which is shown in Fig. 2. The VLM is instructed to output only an English group-level emotional description within 50 words, with no sample-specific manual editing performed.

(1) Event type and contextual meaning capture situational appraisal that fundamentally shapes emotional valence and intensity at the group level. Psychological studies show that collective emotions are strongly conditioned on the perceived purpose and significance of an event (e.g., celebration versus protest), often outweighing individual facial cues [4]. By prompting the model to infer activity type and contextual meaning, the generated descriptions encode scene-level emotional priors that remain robust even when faces are partially occluded or visually ambiguous [5].

(2) Key figures and identity cues operationalize social role theory, which posits that leaders, speakers, or authority figures disproportionately influence group affect and emotional alignment [9]. Visual indicators such as microphones, uniforms, or authoritative postures serve as anchors for interpreting collective emotion. Incorporating such cues enables the model to infer group emotion through role-based reasoning rather than relying solely on appearance-level affect [28].

(3) Facial expression distribution and consistency reflect the distinction between emotional homogeneity and divergence within a group. Prior work suggests that emotionally coherent groups tend to convey stronger and more stable collective emotions, whereas divided expressions often correspond to neutral or ambiguous group affect [3]. By explicitly prompting for proportion estimation, the model produces structured semantic signals that implicitly encode group emotional consensus.

(4) Body language and activation level capture dynamic affective information that is often more informative than static facial expressions, especially in crowded or low-resolution scenarios. Collective movements such as cheering, kneeling, running, or hugging have been shown to correlate strongly with emotional arousal and valence at the group level [29]. These motion-related cues provide non-facial affective signals, aligning with our goal of extracting holistic group emotion cues beyond faces [8].

(5) Symbolic and socio-cultural cues enable the extraction of deep, culturally mediated emotional meanings. Symbols, slogans, flags, and attire often function as emotional signifiers that convey shared attitudes, ideological stances, or collective intentions [30]. Unlike facial expressions, these cues support high-level abstraction and cross-domain generalization, allowing the model to infer emotion even under severe visual noise [7].

Formally, given a group image I, the VLM generates a semantic description T conditioned on the prompt P:

$$T = V(I, P), \tag{1}$$

where $V(\cdot)$ denotes the VLM inference function. The generated text T serves as an explicit semantic representation of group emotion, embedding both surface-level affective cues and deep contextual interpretations.

This vision-to-language mapping has several advantages. Encoding body language captures non-facial cues, mitigating limitations of face-centric GER approaches. Modeling social context and symbolic meaning supports culturally grounded emotion understanding. Language further enables high-level abstraction for representing group relations, collective behavior, and overall atmosphere. Finally, semantic descriptions are less sensitive to low-level visual precision, enhancing generalization across diverse scenarios.

3.2. CLIP-based feature extraction

Contrastive Language–Image Pretraining (CLIP) has shown a strong capability for learning aligned visual and linguistic representations from large-scale image–text pairs [20]. Compared with task-specific visual backbones, CLIP provides a semantically structured embedding space where visual concepts and language descriptions are directly comparable. In this work, CLIP serves as the backbone feature extractor for both group images and VLM-generated semantic descriptions.

3.2.1. Text encoder

The CLIP text encoder maps natural language descriptions into the same semantic embedding space as images, enabling direct cross-modal comparison. Given a VLM-generated group emotion description T, it is tokenized into a CLIP text sequence with $M \leq 77$ token positions, including the required special tokens; any padded positions introduced during batching are identified by the attention mask. Each token ID is mapped by CLIP's learnable text-token embedding lookup table to a token embedding $\mathbf{x}_i \in \mathbb{R}^d$, where $d = 512$ is the embedding dimension:

$$\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_M] + \mathbf{E}_{\text{pos}}^{\text{text}}, \tag{2}$$

where $\mathbf{E}_{\text{pos}}^{\text{text}} \in \mathbb{R}^{M \times d}$ denotes the learnable positional embeddings of the text encoder, and "+" represents element-wise addition applied to each token embedding.

The resulting token sequence is processed by a Transformer-based language encoder that models contextual dependencies across the entire description. Through self-attention mechanisms, the encoder captures both local linguistic patterns (e.g., emotion-related adjectives) and global semantic structure (e.g., event descriptions, social roles, and symbolic cues). For cross-attention, the contextualized token outputs are mapped through CLIP's standard text projection into the shared d-dimensional image–text space. The resulting token matrix is denoted by $\mathbf{H}_t$, while the projected EOT token is retained as the global textual feature used by the contrastive objective:

$$\begin{aligned} \mathbf{H}_t &= \text{Proj}_{\text{text}}\big(\text{Transformer}_{\text{text}}(\mathbf{x})\big) \in \mathbb{R}^{M \times d}, \\ \mathbf{t} &= \mathbf{H}_{t,\text{eot}} \in \mathbb{R}^d. \end{aligned} \tag{3}$$

Thus, $\mathbf{H}_t$ preserves token-level contextual semantics for cross-attention, whereas $\mathbf{t}$ is the pooled global text representation used by InfoNCE.

For compactness, Fig. 1 absorbs CLIP's standard visual/text projection layers into the displayed shared-space representations $\mathbf{v}$, $\mathbf{H}_t$, and $\mathbf{t}$.

3.2.2. Image encoder

The CLIP image encoder is designed to extract high-level semantic representations from raw visual inputs while preserving discriminative information relevant to downstream tasks. Specifically, the input image I is divided into N non-overlapping patches and mapped into the internal token space of the CLIP Vision Transformer. A learnable class token $\mathbf{z}_{\text{cls}}$ is prepended to aggregate global image information:

$$\mathbf{z} = [\mathbf{z}_{\text{cls}}, \mathbf{z}_1, \ldots, \mathbf{z}_N] + \mathbf{E}_{\text{pos}}^{\text{img}}, \tag{4}$$

where $\mathbf{E}_{\text{pos}}^{\text{img}}$ denotes the learnable positional embeddings of the image encoder.

The resulting token sequence is processed by a stack of Transformer layers composed of multi-head self-attention and feed-forward networks. The final class-token representation is then mapped by CLIP's visual projection into the shared image–text space:

$$\mathbf{v} = \text{Proj}_{\text{img}}\big(\text{Transformer}_{\text{img}}(\mathbf{z})_{\text{cls}}\big) \in \mathbb{R}^d, \tag{5}$$

where $d = 512$ for the CLIP-ViT-Base-Patch32 projection space and $\mathbf{v}$ has the same dimensionality as the global textual feature $\mathbf{t}$.

Unlike conventional face-centric GER methods that explicitly detect and aggregate individual faces, the CLIP image encoder implicitly captures holistic scene semantics. This property aligns with our goal of modeling collective group emotion rather than relying on isolated facial expressions. Moreover, trained on diverse web-scale data, CLIP representations are less sensitive to background clutter, providing a robust foundation for real-world group emotion analysis.

3.3. Cross-attention vision–language fusion classification

After extracting the global visual representation $\mathbf{v} \in \mathbb{R}^d$ and the contextual textual token matrix $\mathbf{H}_t \in \mathbb{R}^{M \times d}$ from CLIP, we use the visual representation as the query and the token-level textual representations as keys and values. Here, $d = 512$ and $M \leq 77$ are defined in Section 3.2. All tensor shapes below are written per sample, with the leading batch dimension omitted for clarity. This global-to-token design preserves the holistic visual representation used for GER while allowing the model to assign different weights to the emotional-language tokens.

The projected query, key, and value are

$$\mathbf{Q} = \mathbf{v}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{H}_t\mathbf{W}_k, \quad \mathbf{V} = \mathbf{H}_t\mathbf{W}_v, \tag{6}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$ are learnable projection matrices. Thus, $\mathbf{Q} \in \mathbb{R}^{1 \times d}$ and $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{M \times d}$. Multi-head attention uses eight heads, so each head operates on $d/8 = 64$ feature dimensions. For the r-th head,

$$\mathbf{H}_r = \text{softmax}\left(\frac{\mathbf{Q}_r\mathbf{K}_r^\top}{\sqrt{d/8}}\right)\mathbf{V}_r, \tag{7}$$

where $\mathbf{Q}_r \in \mathbb{R}^{1 \times d/8}$ and $\mathbf{K}_r, \mathbf{V}_r \in \mathbb{R}^{M \times d/8}$. Consequently, $\mathbf{Q}_r\mathbf{K}_r^\top \in \mathbb{R}^{1 \times M}$ and the softmax is computed over the M token positions, with padded positions masked so that only valid tokens receive probability mass. For descriptions containing multiple valid tokens, this yields a nontrivial visual-query-conditioned distribution over the text sequence.

The eight head outputs are concatenated and mapped back to the shared d-dimensional CLIP space:

$$\mathbf{v}_{\text{att}} = \text{Concat}(\mathbf{H}_1, \ldots, \mathbf{H}_8)\mathbf{W}_o, \tag{8}$$

where $\mathbf{W}_o \in \mathbb{R}^{d \times d}$ is the standard multi-head attention output projection and $\mathbf{v}_{\text{att}} \in \mathbb{R}^d$. Fig. 1 depicts the head-wise attention computation schematically; Eq. (8) explicitly includes this output projection. The resulting language-guided feature is combined with the original global visual embedding:

$$\mathbf{v}_{\text{fuse}} = \mathbf{v} + \mathbf{v}_{\text{att}}. \tag{9}$$

The residual path preserves the original CLIP visual evidence, while cross-attention introduces visual-query-dependent aggregation of contextual emotional-language tokens. The w/o attention variant (direct concatenation) and the variant without residual visual fusion are evaluated in Table 7.

The fused representation is passed to a lightweight linear classification head to obtain class logits, from which the final group-emotion label is selected. The standard softmax definition is omitted to keep the presentation focused on the task-specific components.

3.4. Training objective

The training objective is designed to serve two complementary roles: supervised emotion discrimination and semantic alignment between group images and emotional descriptions. Although a standard classifier trained with cross-entropy can learn decision boundaries for the three emotion categories, it does not explicitly require the visual representation of a group image to be close to its language-level emotional interpretation. This is important for GER because visually similar scenes may correspond to different group emotions once activity type, social role, cultural symbols, and collective atmosphere are considered. Inspired by recent studies showing that multimodal emotion recognition benefits from carefully designed fusion and contrastive learning objectives [24,26], we formulate training as a decoupled alignment–classification optimization problem.

Given a group image and its corresponding VLM-generated emotional description, the cross-attention-based residual fusion module produces a fused representation $\mathbf{v}_{\text{fuse}}$ (Eq. 9), which is mapped by the classification head to the predicted probability distribution $\mathbf{p}$. Let C denote the number of emotion categories, and let $\mathbf{y}^* \in \{0,1\}^C$ denote the one-hot encoded ground-truth label. The supervised classification term is defined as:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_c^* \log(p_c), \tag{10}$$

where y_c^* and p_c denote the ground-truth label and predicted probability for class c, respectively. This term directly optimizes task-specific discriminability and stabilizes training by anchoring the fused representation to the final emotion categories.

However, using $\mathcal{L}_{\text{cls}}$ alone means that the language branch is optimized only through the classification pathway. In that case, the network may still overfit dataset-specific visual patterns and ignore fine-grained emotional semantics provided by the generated description. Therefore, we further introduce a symmetric image–text contrastive loss. Instead of applying this loss to the post-fusion prediction space, we apply it to the pre-fusion CLIP image and text embeddings. This design preserves a clean semantic alignment constraint: $\mathcal{L}_{\text{con}}$ shapes the shared image–text representation space, while $\mathcal{L}_{\text{cls}}$ optimizes the label decision after cross-attention-based residual fusion. The two losses therefore play different but compatible roles.

Given a mini-batch of N image–text pairs $\{(I_i, T_i)\}_{i=1}^N$, the visual and textual embeddings are first ℓ_2-normalized:

$$\hat{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|}, \quad \hat{\mathbf{t}}_i = \frac{\mathbf{t}_i}{\|\mathbf{t}_i\|}. \tag{11}$$

The normalization makes the contrastive term depend on angular semantic similarity rather than embedding magnitude, which is consistent with CLIP-style image–text alignment.

Following the InfoNCE formulation, the image-to-text contrastive loss is defined as:

$$\mathcal{L}_{\text{v2t}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\hat{\mathbf{v}}_i^\top \hat{\mathbf{t}}_i/\tau\right)}{\sum_{j=1}^{N}\exp\left(\hat{\mathbf{v}}_i^\top \hat{\mathbf{t}}_j/\tau\right)}, \tag{12}$$

where τ is a temperature parameter controlling the sharpness of the similarity distribution. A smaller τ strengthens the penalty for hard negatives but may amplify noisy or semantically similar negative pairs; a larger τ gives smoother gradients but weakens alignment. We therefore use a moderate value to balance discriminative alignment and optimization stability.

Similarly, the text-to-image contrastive loss is computed as:

$$\mathcal{L}_{\text{t2v}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\hat{\mathbf{t}}_i^\top \hat{\mathbf{v}}_i/\tau\right)}{\sum_{j=1}^{N}\exp\left(\hat{\mathbf{t}}_i^\top \hat{\mathbf{v}}_j/\tau\right)}. \tag{13}$$

The two directional terms are both necessary in our setting. The image-to-text direction encourages each group image to retrieve its own semantic description, while the text-to-image direction prevents descriptions with similar emotional wording from becoming indistinguishable across different images. This bidirectional constraint is particularly useful when several images share the same coarse label but differ in social context or group composition.

Algorithm 1 Semantic alignment training of ED-VLAlign.

Require: Mini-batch $\{(I_i, y_i^*)\}_{i=1}^N$; descriptions $\{T_i\}_{i=1}^N$; temperature τ; loss weight λ

Ensure: Updated parameters of the encoders, fusion module, and classifier

1: Encode group images with the CLIP image encoder to obtain $\{\mathbf{v}_i\}_{i=1}^N$
2: Encode emotional descriptions to obtain token matrices $\{\mathbf{H}_{t,i}\}_{i=1}^N$ and global EOT features $\{\mathbf{t}_i\}_{i=1}^N$
3: Use $\mathbf{v}_i$ as the query and $\mathbf{H}_{t,i}$ as keys/values for multi-head cross-attention, then apply residual fusion to obtain $\mathbf{v}_{\text{fuse},i}$
4: Predict class probabilities $\mathbf{p}_i = \text{softmax}(\mathbf{W}_c \mathbf{v}_{\text{fuse},i} + \mathbf{b}_c)$
5: Compute $\mathcal{L}_{\text{cls}}$ using ground-truth labels and predicted probabilities
6: Normalize $\mathbf{v}_i$ and $\mathbf{t}_i$, then compute $\mathcal{L}_{\text{v2t}}$ and $\mathcal{L}_{\text{t2v}}$
7: Compute $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{con}}$
8: Update trainable parameters by minimizing $\mathcal{L}_{\text{total}}$ with AdamW

The overall contrastive loss is obtained by averaging the two directions:

$$\mathcal{L}_{\text{con}} = \frac{1}{2}\left(\mathcal{L}_{\text{v2t}} + \mathcal{L}_{\text{t2v}}\right). \tag{14}$$

The final training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{con}}, \tag{15}$$

where λ balances supervised classification learning and cross-modal semantic alignment. In practice, λ should not be too large, because an overly dominant contrastive term may over-emphasize instance-level image–text matching and weaken class-level separability. Conversely, a very small value weakens explicit image–text alignment and leaves the language branch to be optimized mainly through the classification objective. Thus, the loss weight is chosen to make the textual branch act as a semantic regularizer rather than a replacement for supervised recognition.

Algorithm 1 summarizes the optimization procedure. During training, the VLM-generated descriptions are treated as fixed semantic supervision, and the trainable CLIP encoders, fusion module, and classifier are updated jointly.

This objective transforms emotional descriptions from auxiliary text into an explicit training signal. The classification loss ensures category-level correctness, the contrastive loss imposes instance-level semantic consistency, and cross-attention-based residual fusion connects the aligned representations to the final decision. As a result, the model is encouraged to learn not only which class an image belongs to, but also why its visual content is semantically consistent with a particular group-level emotional interpretation.

4. Experiments

4.1. Datasets

To evaluate the effectiveness and generalization ability of the proposed ED-VLAlign framework, experiments are conducted on four widely used benchmark datasets for group-level emotion recognition: GAF2, GAF3, GroupEmoW, and VGAF. These datasets cover diverse real-world social scenarios with varying group sizes, visual complexity, emotional ambiguity, and temporal dynamics, providing a comprehensive testbed for collective affect modeling.

The Group Affect Dataset 2.0 (GAF2) was released for the EmotiW 2017 Group-level Emotion Recognition Challenge [3]. It contains 3630

training images and 2065 validation images collected from Google and Flickr using event-related keywords, including festivals, parties, calm gatherings, protests, and violent activities. Each image depicts a group of people and is annotated by multiple annotators as one of three emotion categories: positive, neutral, or negative. As the official test set is not publicly available, only the training and validation sets are used for evaluation.

The Group Affect Dataset 3.0 (GAF3) was introduced in the EmotiW 2018 Challenge as an extension of GAF2 [31]. It includes 9815 training images, 4346 validation images, and 3011 test images, collected and annotated following a similar protocol to GAF2. Since test annotations are not released, experiments are conducted using only the training and validation splits.

GroupEmoW is a large-scale public dataset specifically designed for group emotion recognition [32]. It contains 15,894 images with at least two people per image, split into 11,127 training, 3178 validation, and 1589 test images. Images were collected from multiple search engines using social-event-related keywords, and each image was labeled by five independent annotators as positive, neutral, or negative categories. Unlike GAF2 and GAF3, all splits of GroupEmoW are publicly accessible.

VGAF (Video-level Group Affect) is an unconstrained video benchmark for recognizing collective affect in short social-event clips [33]. The 4183 clips were gathered from online videos through event-oriented queries spanning celebratory, routine, and conflict-related situations. Each clip is assigned one of three group-level valence labels: Positive, Neutral, or Negative. The official partition contains 2661 training clips, 766 validation clips, and 756 test clips. Because the labels of the official test split are withheld, we optimize the model on the training split and report all VGAF results on the validation split.

4.2. Settings and configuration

This section details the settings used to train and evaluate the proposed ED-VLAlign framework, including backbone selection, preprocessing, optimization, loss configuration, and evaluation protocols. All settings are kept consistent across datasets for fair comparison and reproducibility.

We use GLM-4.1 V-Thinking-Flash, a VLM with strong visual reasoning and instruction-following ability, to generate semantic descriptions [19], and adopt the pretrained CLIP-ViT-Base-Patch32 as the vision–language encoder [20]. Image and text embeddings are both 512-dimensional, aligned with the CLIP projection layer. All CLIP parameters are fine-tuned jointly with the attention fusion and classification modules.

For images, training involves resizing the shorter side to 256 pixels, random resized cropping to 224×224, and random horizontal flipping; validation and testing use 256-pixel resizing and center-crop to 224×224. Images are normalized with CLIP's default mean $(0.4814, 0.4578, 0.4082)$ and standard deviation $(0.2686, 0.2613, 0.2758)$. For VGAF, 20 frames are sampled at approximately equal temporal intervals over the complete clip and independently encoded by the CLIP visual encoder. The 20 frame embeddings form a temporal feature sequence of size 20×512; they are not concatenated spatially or stacked into a single image. Mean pooling is then applied across the 20-frame dimension to obtain one 512-dimensional clip-level visual feature. A clip-level emotional description is encoded by the CLIP text branch, after which the pooled clip-level visual feature serves as the query and the contextual text-token sequence serves as keys and values for multi-head cross-attention; residual fusion, contrastive alignment, and classification then follow the same pipeline used for the image datasets.

Validation is performed after each training epoch. The model with the highest validation accuracy is saved for testing. The maximum number of epochs is 50, with early stopping after 5 non-improving epochs. The InfoNCE temperature is 0.07, and the contrastive loss weight is $\lambda = 0.3$.

Optimization uses AdamW with a learning rate of 1×10^{-6} for all trainable parameters and a batch size of 32. Experiments are conducted on a single NVIDIA TITAN GPU. To reduce random-seed bias and provide a statistically transparent evaluation, all main experiments on GAF2, GAF3, GroupEmoW, and VGAF are repeated across 10 independent runs. We report the mean and sample standard deviation across these runs, while the class-wise precision, recall, and F1 statistics are computed from the per-run classification reports. The best run is also retained for direct comparison with prior single-run results. For VGAF, the official training split is used for optimization and the validation split is used for evaluation; all remaining hyperparameters are identical to those used for GAF and GroupEmoW.

4.3. Overall performance

To evaluate the effectiveness of the proposed method for group emotion recognition, we compare it with representative state-of-the-art methods on four public benchmarks: GAF2, GAF3, GroupEmoW, and VGAF. The compared approaches cover hierarchical modeling, graph-based relational reasoning, context-aware attention mechanisms, and recent deep learning frameworks for group affect analysis. Performance is evaluated using standard metrics, including overall accuracy, class-wise precision, recall, and F1-score.

4.3.1. Results on the GAF2 dataset

Table 1 reports the performance comparison on the GAF2 dataset, which features complex crowd compositions, frequent occlusions, and large intra-class variations. To complement the original single best-run result, the table now reports statistics from 10 independent classification reports. Across these runs, ED-VLAlign achieves a mean accuracy of $81.29 \pm 0.70\%$, with the best run reaching 82.86%. The best-run accuracy remains higher than all the compared methods, while the 10-run mean provides a more conservative estimate of stability under different random initializations.

From the class-wise reports, the 10-run mean precision values for Positive, Neutral, and Negative emotions are $92.95 \pm 0.99\%$, $76.11 \pm 2.10\%$, and $75.01 \pm 1.68\%$, respectively, and the corresponding recall values are $82.41 \pm 2.34\%$, $74.14 \pm 2.60\%$, and $88.99 \pm 1.14\%$. The mean macro-F1 score is $81.26 \pm 0.66\%$. Compared with the best run, the repeated-run statistics show that ED-VLAlign is consistently strong rather than relying on a single favorable initialization. In particular, the Negative class maintains high recall across runs, reflecting the advantage of emotional language descriptions in capturing implicit group affect under visually ambiguous scenes.

4.3.2. Statistical significance analysis on GAF2

To complement the repeated-run reporting in Table 1, we evaluate whether the accuracy improvement of ED-VLAlign over the reproduced Wang et al. baseline on GAF2 is statistically significant. Both methods are evaluated over 10 independent runs under the same GAF2 setting. ED-VLAlign obtains $81.288 \pm 0.702\%$ accuracy, whereas the Wang et al. baseline obtains $80.553 \pm 0.484\%$, corresponding to a mean difference of 0.735 percentage points. Table 2 summarizes the statistical tests, effect size, and confidence interval.

Welch's independent two-sample t-test, the Mann–Whitney U test, and a 10,000-resample permutation test all yield $p < 0.05$. Cohen's $d = 1.2186$ indicates a large standardized effect, and the bootstrap 95% confidence interval for the mean difference, of [0.264, 1.264] percentage points, excludes zero. These complementary results consistently support a statistically significant repeated-run accuracy advantage for ED-VLAlign in this GAF2 comparison. The published Wang et al. result in Table 1 is retained for literature comparison; the reproduced 10-run values are used only for the significance analysis.

4.3.3. Results on the GAF3 dataset

The GAF3 dataset presents greater challenges due to more diverse activities, complex backgrounds, and increased visual ambiguity. As shown in Table 3, ED-VLAlign achieves a 10-run mean accuracy of $83.76 \pm 0.33\%$, and the best run reaches 84.31%, outperforming the

Table 1
Comparison of classification performance of different methods on the GAF2 dataset. ED-VLAlign mean results are computed from 10 per-run classification reports and are shown as mean±SD. The best-run row reports the best ED-VLAlign run. Bold values indicate the best results among the reported baselines and ED-VLAlign (best); ED-VLAlign (mean) is excluded from the best-value comparison. The symbol "–" indicates "not reported". Unit: %.

Methods	Acc.	Precision				Recall				F1			
		Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.
Wang et al. [8]	81.97	89.52	80.96	72.51	81.00	–	–	–	–	–	–	–	–
Zhu et al. [10]	79.52	87.57	73.93	76.08	79.19	84.16	75.18	78.62	79.32	85.83	74.55	77.33	79.23
Guo et al. [16]	79.36	81.87	75.60	**80.71**	79.39	–	–	–	–	–	–	–	–
Zhu et al. [28]	80.95	–	–	–	–	87.84	**77.55**	74.10	80.56	–	–	–	–
Gong et al. [34]	79.60	86.88	77.36	71.58	78.99	–	–	–	78.60	–	–	–	78.72
Fujii et al. [35]	80.41	–	77.51	74.10	80.19	**88.41**	72.51	79.64	80.19	88.12	74.95	76.76	79.95
Li et al. [36]	–	88.29	**83.91**	79.06	**83.75**	87.94	76.21	80.42	81.07	86.02	74.68	78.17	80.12
ED-VLAlign (best)	**82.86**	**91.15**	81.16	75.23	82.51	87.97	72.80	**88.83**	**83.20**	**89.53**	**76.76**	**81.46**	**82.58**
ED-VLAlign (mean)	81.29 ± 0.70	92.95 ± 0.99	76.11 ± 2.10	75.01 ± 1.68	81.36 ± 0.65	82.41 ± 2.34	74.14 ± 2.60	88.99 ± 1.14	81.84 ± 0.62	87.34 ± 1.06	75.06 ± 1.09	81.38 ± 0.61	81.26 ± 0.66

Table 2
Statistical significance analysis of ED-VLAlign versus the reproduced Wang et al. baseline on GAF2. Both methods use 10 independent runs. The observed mean accuracy difference is +0.735 percentage points.

Analysis	Statistic/estimate	*p*-value	Interpretation
Welch *t*-test	$t = 2.7248$	0.0150	Significant
Mann–Whitney *U*	$U = 84.0000$	0.0112	Significant
Permutation test (10,000)	$\Delta = +0.7350\%$	0.0095	Significant
Cohen's *d*	1.2186	–	Large effect
Bootstrap 95% CI	[0.2640, 1.2640]%	–	Excludes zero

strongest reported visual-only baseline, Xu et al. [37]. The added F1 columns further summarize the per-run classification reports and provide a more complete view of class-wise behavior.

Across 10 runs, the model obtains mean precision values of 89.11 ± 0.87%, 78.83 ± 1.90%, and 81.53 ± 2.07% for Positive, Neutral, and Negative classes, with mean recall values of 90.73 ± 1.07%, 74.07 ± 2.71%, and 84.46 ± 2.04%, respectively. The mean macro-F1 score is 83.04 ± 0.40%. These results indicate that the proposed framework remains robust on the larger and more visually diverse GAF3 validation split. The relatively small standard deviation in accuracy also suggests that the benefits of emotion-oriented semantic descriptions and vision–language alignment are stable across repeated trials.

4.3.4. Results on the GroupEmoW dataset

Table 4 presents a comparison on the GroupEmoW dataset, which contains richer scene context and more consistent group-level annotations. In addition to the original best-run result of 92.07%, the supplementary 10-run evaluation gives a mean accuracy of 91.28 ± 0.40%. Although the best-run gain over the strongest baseline is moderate because GroupEmoW already contains strong prior results, the repeated-run statistics confirm that ED-VLAlign remains consistently competitive across random seeds.

The 10-run classification reports show balanced class-wise performance: precision values for Positive, Neutral, and Negative classes are 93.70 ± 0.95%, 88.48 ± 1.25%, and 90.65 ± 0.47%, while recall values are 95.21 ± 0.59%, 85.44 ± 1.72%, and 91.72 ± 1.10%, respectively. The mean macro-F1 score is 90.90 ± 0.42%. Compared with Li et al. [36], the best run improves Neutral precision, indicating better handling of emotionally ambiguous group scenes. The repeated-run results further indicate stable class-wise performance across random initializations, while the remaining differences among classes are reflected in the reported precision and recall values.

4.3.5. Results on the VGAF dataset

Although the primary scope of ED-VLAlign is image-based GER, we conduct this preliminary VGAF experiment to assess the potential of the proposed semantic-alignment strategy for video-based GER when a clip is represented by uniformly sampled frame features. Table 5 reports the results on the official VGAF validation split using the 10-run evaluation. The proposed method obtains a mean accuracy of 75.90 ± 0.49%, and the best run reaches 76.50%. Among the published baselines listed in Table 5, the strongest reported accuracy is 70.23%, achieved by several visual or audiovisual methods. ED-VLAlign (best) therefore improves this value by 6.27 percentage points, indicating that clip-level emotional descriptions remain useful when the visual representation is formed by pooling uniformly sampled video frames.

The repeated-run class-wise precision values are 81.53 ± 2.11%, 69.61 ± 1.58%, and 78.01 ± 2.98% for Positive, Neutral, and Negative clips, respectively. The corresponding recalls are 77.81 ± 2.48%, 74.89 ± 4.60%, and 74.29 ± 4.39%. In the best run, Positive precision reaches 84.27%, whereas Neutral precision (69.87%) remains below the 73.00% reported by Evtodienko and Negative precision (76.47%) remains below the 80.00% reported by Petrova et al. The mean macro-F1 is 75.86 ± 0.54%, showing stable but non-uniform class behavior. These

Table 3
Comparison of classification performance of different methods on the GAF3 dataset. ED-VLAlign mean results are computed from 10 per-run classification reports and are shown as mean±SD. The best-run row reports the best ED-VLAlign run. Bold values indicate the best results among the reported baselines and ED-VLAlign (best); ED-VLAlign (mean) is excluded from the best-value comparison. The symbol "–" indicates "not reported". Unit: %.

Methods	Acc.	Precision				Recall				F1			
		Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.
Wang et al. [8]	79.71	87.52	**79.02**	69.37	78.63	–	–	–	–	–	–	–	–
Zhu et al. [10]	77.98	84.46	72.62	74.21	77.10	87.12	69.81	74.09	77.01	–	–	–	–
Khan et al. [14]	79.13	–	–	–	–	86.89	75.00	72.71	78.20	–	–	–	–
Guo et al. [16]	79.47	86.70	78.51	70.27	78.49	–	–	–	–	–	–	–	–
Zhu et al. [28]	81.29	–	–	–	–	82.29	**79.03**	**82.19**	81.17	–	–	–	–
Guo et al. [32]	79.08	–	–	–	–	–	–	–	–	–	–	–	–
Fujii et al. [35]	77.54	82.88	72.64	74.32	76.61	89.83	70.92	67.41	76.05	–	–	–	–
Xu et al. [37]	83.97	–	–	–	–	–	–	–	–	–	–	–	–
ED-VLAlign (best)	**84.31**	**89.11**	77.32	**85.30**	**83.91**	**91.30**	78.73	80.58	**83.54**	**90.19**	**78.02**	**82.87**	**83.69**
ED-VLAlign (mean)	83.76 ± 0.33	89.11 ± 0.87	78.83 ± 1.90	81.53 ± 2.07	83.16 ± 0.45	90.73 ± 1.07	74.07 ± 2.71	84.46 ± 2.04	83.09 ± 0.50	89.92 ± 0.33	76.30 ± 0.83	82.89 ± 0.57	83.04 ± 0.40

Table 4
Comparison of classification performance of different methods on the GroupEmoW dataset. ED-VLAlign mean results are computed from 10 per-run classification reports and are shown as mean±SD. The best-run row reports the best ED-VLAlign run. Bold values indicate the best results among the reported baselines and ED-VLAlign (best); ED-VLAlign (mean) is excluded from the best-value comparison. The symbol "–" indicates "not reported". Unit: %.

Methods	Acc.	Precision				Recall				F1			
		Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.
Wang et al. [8]	90.19	94.58	85.63	88.63	–	–	–	–	–	–	–	–	–
Khan et al. [14]	90.18	94.95	85.05	88.50	–	–	–	–	–	–	–	–	–
Zhu et al. [28]	91.88	–	–	–	–	–	–	–	–	–	–	–	–
Guo et al. [32]	89.93	–	–	–	–	–	–	–	–	–	–	–	–
Li et al. [36]	91.77	**96.21**	86.38	90.16	–	–	–	–	–	–	–	–	–
ED-VLAlign (best)	**92.07**	94.07	**90.25**	**90.95**	**91.76**	**95.63**	**86.23**	**93.27**	**91.71**	**94.85**	**88.20**	**92.10**	**91.71**
ED-VLAlign (mean)	91.28 ± 0.40	93.70 ± 0.95	88.48 ± 1.25	90.65 ± 0.47	90.94 ± 0.27	95.21 ± 0.59	85.44 ± 1.72	91.72 ± 1.10	90.79 ± 0.55	94.45 ± 0.49	87.06 ± 0.65	91.18 ± 0.46	90.90 ± 0.42

Table 5
Comparison of classification performance on the VGAF validation set. Bold values identify the best result for each metric when ED-VLAlign (best) is compared with the published baselines; ED-VLAlign (mean) is reported for stability and is excluded from this comparison. The symbol "–" denotes an unavailable or unreported value. Mean results are shown as mean±SD over 10 runs. For published methods, Ave. precision is the macro-average calculated from the three reported class-wise precision values when available. (Unit: %).

Methods	Acc.	Precision				Recall				F1			
		Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.
Pinto et al. [38]	62.40	69.20	52.50	66.30	62.67	–	–	–	–	–	–	–	–
Sun et al. [39]	70.23	–	–	–	–	–	–	–	–	–	–	–	–
Petrova et al. [40]	52.36	60.00	40.00	**80.00**	60.00	–	–	–	–	–	–	–	–
Evtodienko [41]	60.37	51.00	**73.00**	56.00	60.00	–	–	–	–	–	–	–	–
Savchenko et al. [42]	70.23	73.20	64.30	70.70	69.40	–	–	–	–	–	–	–	–
Wang et al. [43]	70.23	–	–	–	–	–	–	–	–	–	–	–	–
Huang and Xu [44]	66.42	60.41	70.26	71.32	67.33	–	–	–	–	–	–	–	–
ED-VLAlign (best)	**76.50**	**84.27**	69.87	76.47	**76.87**	**74.50**	**77.86**	**77.72**	**76.69**	**79.09**	**73.65**	**77.09**	**76.61**
ED-VLAlign (mean)	75.90 ± 0.49	81.53 ± 2.11	69.61 ± 1.58	78.01 ± 2.98	76.38 ± 0.84	77.81 ± 2.48	74.89 ± 4.60	74.29 ± 4.39	75.66 ± 0.66	79.57 ± 0.69	72.05 ± 1.58	75.97 ± 1.74	75.86 ± 0.54

results support the extension to video clips while also confirming that average pooling can smooth short-lived affective evidence, motivating more explicit temporal modeling.

4.4. Ablation study

To analyze the contribution of different components in the proposed ED-VLAlign framework, we conduct ablation experiments on the GAF2 dataset. The study includes two parts: (1) model-level ablations evaluating the vision–language architecture and learning objectives, and (2) prompt-level ablations examining the role of each semantic aspect in the emotion description prompt. All variants are trained and evaluated under identical settings for fair comparison.

Fig. 3 provides a direct cross-dataset view of the benefit of adding emotional language representations. Relative to the visual-only configurations, visual–language integration raises accuracy from 77.81% to 82.86% on GAF2, from 81.41% to 84.31% on GAF3, and from 90.94% to 92.07% on GroupEmoW. Macro-averaged precision improves from 77.92% to 82.51%, from 81.00% to 83.91%, and from 91.00% to 91.76%, respectively. The gains are largest on GAF2 and GAF3, where scene ambiguity and class overlap are more pronounced, and remain positive on the stronger GroupEmoW baseline.

4.4.1. Prompt-level ablation

Table 6 presents the prompt-level ablation results, where one semantic aspect is removed at a time while keeping the other prompt components unchanged. The five aspects include: (1) event type and contextual meaning, (2) key figures and identity cues, (3) facial expression distribution, (4) body language and activation level, and (5) symbolic and socio-cultural cues. Removing any single aspect consistently degrades performance compared with the full prompt, which achieves the highest accuracy of 82.86%.

Among the variants, removing key figures and identity cues (w/o (2)) causes the largest accuracy decrease, highlighting the importance of social-role information in group emotion understanding. Removing facial expression distribution (w/o (3)) and symbolic and socio-cultural cues (w/o (5)) also produce comparatively large declines, while removing event/context information (w/o (1)) or body language (w/o (4)) results in smaller but still consistent decreases. Together, these results indicate that the five prompt components provide complementary semantic cues rather than contributing equally.

4.4.2. Model-level ablation

Table 7 reports five model-level ablations and unimodal baselines compared with the full model. When removing the textual branch (w/o text), the model becomes a purely visual baseline and achieves 77.81% accuracy and 77.83% macro-F1, confirming that visual evidence alone leaves a substantial performance gap.

Conversely, the w/o image baseline (i.e., the requested Text-only setting) performs classification using only the generated emotional-description feature and obtains 80.41% accuracy, 80.15% macro-precision, 80.78% macro-recall, and 80.20% macro-F1. The w/o image result is 2.60 percentage points higher in accuracy and 2.37 points higher in macro-F1 than the visual-only baseline, confirming that the generated descriptions themselves contain strong affective semantics. However, the complete ED-VLAlign model still improves over w/o image by 2.45 percentage points in accuracy and 2.38 points in macro-F1. Thus, the overall gain cannot be attributed to the text branch alone; combining visual evidence with the emotional description provides additional complementary information.

Removing the InfoNCE-based contrastive loss (w/o InfoNCE) results in a noticeable performance decrease (81.36% accuracy). Although both modalities are retained, the lack of explicit cross-modal alignment weakens semantic consistency, particularly for the Neutral class, indicating that effective vision–language alignment is critical for stable multimodal learning.

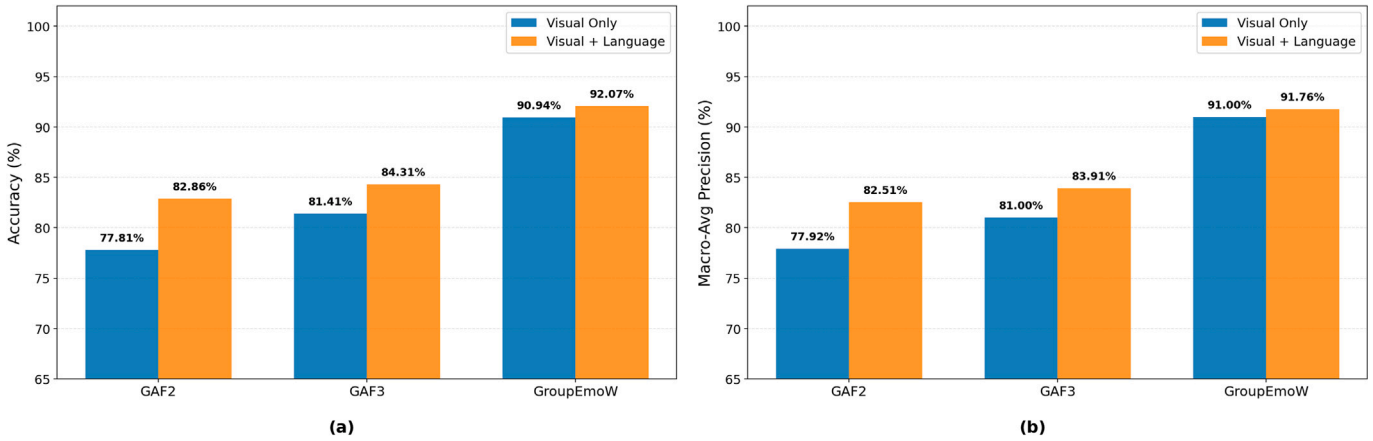


Fig. 3. Cross-dataset comparison between visual-only representations and the integration of visual features with emotional language representations. (a) accuracy and (b) macro-averaged precision on GAF2, GAF3, and GroupEmoW.

Table 6
Prompt-level ablation on the GAF2 dataset. Best results are in bold. (Unit: %).

Methods	Acc.	Precision				Recall				F1			
		Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.
w/o (1)	80.77	91.28	77.09	73.40	80.59	83.96	72.12	87.59	81.22	87.47	74.52	79.87	80.62
w/o (2)	79.08	90.81	73.90	72.85	79.19	79.30	71.57	88.48	79.78	84.67	72.71	79.90	79.10
w/o (3)	80.24	90.60	77.44	72.27	80.10	83.57	69.78	**89.18**	80.85	86.94	73.41	79.84	80.07
w/o (4)	80.73	**91.69**	73.91	**77.22**	80.94	79.95	**77.06**	86.52	81.18	85.42	75.45	**81.61**	80.83
w/o (5)	80.29	91.68	75.21	74.03	80.31	81.24	73.35	87.94	80.85	86.15	74.27	80.39	80.27
ED-VLAlign (Ours)	**82.86**	91.15	**81.16**	75.23	**82.51**	**87.97**	72.80	88.83	**83.20**	**89.53**	**76.76**	81.46	**82.58**

Table 7
Model-level ablation and unimodal baselines on the GAF2 dataset. Best results are in bold. (Unit: %).

Methods	Acc.	Precision				Recall				F1			
		Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.
w/o text	77.81	88.75	72.22	72.80	77.92	78.46	71.03	85.67	78.39	83.24	71.54	78.69	77.83
w/o image	80.41	90.83	76.03	73.57	80.15	85.30	70.38	86.65	80.78	87.96	73.07	79.56	80.20
w/o InfoNCE	81.36	**91.58**	77.33	74.62	81.18	85.41	71.92	87.98	81.77	88.34	74.46	80.72	81.17
w/o attention	81.58	90.28	78.46	**75.25**	81.33	87.50	71.92	85.94	81.79	88.79	74.89	80.17	81.28
w/o residual visual fusion	81.26	91.09	78.60	73.05	80.92	87.32	69.64	87.94	81.64	89.17	73.85	79.81	80.94
ED-VLAlign (Ours)	**82.86**	91.15	**81.16**	75.23	**82.51**	**87.97**	**72.80**	**88.83**	**83.20**	**89.53**	**76.76**	**81.46**	**82.58**

In the w/o attention variant, visual and textual features are directly concatenated without attention-based fusion. While this setting slightly outperforms w/o InfoNCE, it still underperforms the full model, showing that naive fusion fails to capture fine-grained cross-modal interactions. The attention mechanism enables adaptive integration of linguistic cues conditioned on visual context, which is essential for exploiting complementary information.

The w/o residual visual fusion variant removes the addition of the original CLIP visual embedding and directly forwards the cross-attention feature to the classifier. Its accuracy decreases from 82.86% to 81.26%, while macro-F1 drops from 82.58% to 80.94%. The largest degradation appears in Neutral recall (69.64% versus 72.80%) and in Neutral and Negative precision. These results support the residual design in Eq. (9): the original visual embedding retains holistic appearance and scene evidence, whereas the cross-attention branch emphasizes description-conditioned semantics. Their addition preserves complementary information and provides a more stable representation than the cross-attention output alone.

4.5. Generic description comparison

To determine whether the improvement originates from emotion-oriented semantic guidance rather than merely from adding a textual modality, we construct a controlled generic-description variant for GAF2. For every image, GLM-4.1 V-Thinking-Flash receives the prompt: *"Describe the image objectively in no more than 50 English words. Mention visible people, actions, objects, and the scene."* The resulting descriptions focus on observable image content without explicitly requesting emotion categories, affective appraisal, group atmosphere, or socio-cultural interpretation. These generic descriptions replace the emotion-oriented descriptions, while the CLIP encoders, attention module, loss functions, data split, optimization settings, and evaluation protocol remain unchanged.

As shown in Table 8, generic descriptions achieve 78.50% accuracy and a macro-F1 score of 78.36%, whereas the emotion-oriented descriptions used by ED-VLAlign reach 82.86% accuracy and 82.58% macro-F1. The corresponding gains of 4.36 and 4.22 percentage points indicate that the benefit cannot be attributed solely to introducing an additional text stream. The largest precision gap occurs for the Neutral class (81.16% versus 74.96%), suggesting that task-specific descriptions are especially useful when the visible scene is emotionally ambiguous and objective object–action captions do not explicitly encode collective affect. Generic descriptions remain slightly above the visual-only accuracy of 77.81% in Table 7, showing that general scene semantics provide some auxiliary information; however, the substantially stronger full result confirms the additional value of emotion-oriented appraisal, group-atmosphere reasoning, and socio-cultural cues.

Table 8
Comparison between generic descriptions and emotion-oriented descriptions on the GAF2 validation set. Best results are in bold (Unit: %).

Methods	Acc.	Precision				Recall				F1			
		Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.
Generic descriptions	78.50	89.33	74.96	70.91	78.40	81.24	69.09	86.88	79.07	85.09	71.91	78.09	78.36
ED-VLAlign (Ours)	**82.86**	**91.15**	**81.16**	**75.23**	**82.51**	**87.97**	**72.80**	**88.83**	**83.20**	**89.53**	**76.76**	**81.46**	**82.58**

Table 9
Performance with descriptions generated by different VLMs on the GAF2 validation set. Best results are in bold. (Unit: %).

Description generator	Acc.	Precision				Recall				F1			
		Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.
Qwen3-VL-8B-Instruct	81.65	**91.69**	77.07	**75.65**	81.47	84.22	74.31	87.59	82.04	87.80	75.66	81.18	81.55
InternVL3.5-8B	80.73	90.83	78.57	72.33	80.58	84.61	69.51	**89.89**	81.33	87.61	73.76	80.16	80.51
Idefics3-8B-Llama3	79.95	91.53	74.25	74.47	80.08	79.69	**74.86**	86.88	80.48	85.20	74.56	80.20	79.98
GLM-4.1 V-Thinking-Flash (Ours)	**82.86**	91.15	**81.16**	75.23	**82.51**	**87.97**	72.80	88.83	**83.20**	**89.53**	**76.76**	**81.46**	**82.58**

Table 10
Zero-shot classification results of GLM-4.1 V-Thinking-Flash on the GAF2 validation set. Best results are in bold. (Unit: %).

Methods	Acc.	Precision				Recall				F1			
		Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.	Pos.	Neu.	Neg.	Ave.
GLM	73.26	73.73	68.97	**80.23**	74.31	**94.52**	66.24	52.40	71.05	82.84	67.58	63.39	71.27
ED-VLAlign (Ours)	**82.86**	**91.15**	**81.16**	75.23	**82.51**	87.97	**72.80**	**88.83**	**83.20**	**89.53**	**76.76**	**81.46**	**82.58**

4.6. Sensitivity to the description-generation VLM

To examine whether ED-VLAlign depends on a particular description-generation model, we replace the primary GLM-based describer with three open vision–language models: Qwen3-VL-8B-Instruct [45], InternVL3.5-8B [46], and Idefics3-8B-Llama3 [47]. Each model receives the same emotion-oriented prompt used in the primary pipeline and generates one description for every GAF2 image. The generated descriptions are then used to train and evaluate ED-VLAlign under the same data split, CLIP encoders, optimization settings, attention module, and loss functions. Thus, the description-generation VLM is the only changed component.

Table 9 shows that the framework remains effective when descriptions are produced by different VLM families. Qwen3-VL-8B-Instruct reaches 81.65% accuracy, 1.21 percentage points below the primary GLM-based result, while InternVL3.5-8B and Idefics3-8B-Llama3 obtain 80.73% and 79.95%, respectively. All three variants remain above the 77.81% visual-only result in Table 7, confirming that the benefit is not restricted to one description generator. The class-wise results also reveal model-dependent differences: Qwen3-VL provides the highest Positive and Negative precision among the evaluated describers, Idefics3-8B-Llama3 yields the highest Neutral recall, and InternVL3.5 yields the highest Negative recall. The primary GLM-based configuration retains the best overall accuracy, macro-precision, macro-recall, and macro-F1, indicating that description quality affects performance even though the alignment framework generalizes across VLMs.

4.7. Zero-shot recognition with vision–language models

Recent VLMs have demonstrated strong transfer across visual understanding tasks; CLIP provides a representative zero-shot image-classification example, while Flamingo demonstrates few-shot vision–language adaptation [11,20]. A key question is whether such models can be directly applied to group-level emotion recognition without task-specific training. To investigate this and ensure a fair comparison, we adopt the general-purpose vision–language model GLM-4.1 V-Thinking-Flash as a zero-shot baseline for vision-language semantic description generation. Since this model is not specifically designed for group emotion recognition, this experiment provides a direct zero-shot reference for a general-purpose VLM under the same input and prompt information.

Table 11
Cross-dataset recognition accuracy on the GAF2 and GroupEmoW datasets. Best results are in bold. (Unit: %).

Methods	GAF2→GroupEmoW	GroupEmoW→GAF2
Li et al. [36]	81.30	82.41
ED-VLAlign (Ours)	**88.74**	**82.57**

To maintain strict consistency with our proposed method, the model is provided with the same input images and emotion-related textual description prompts as in Section 3.1. The only difference is that the model is instructed to output predictions in a structured JSON format, including a three-class group emotion label (positive, neutral, or negative) and auxiliary semantic attributes. The predicted emotion label is extracted from the generated JSON and compared with ground-truth annotations. All experiments are conducted on the GAF2 dataset, and accuracy, precision, recall, and F1-score are reported for each emotion category.

Although the VLM achieves reasonable zero-shot performance, Table 10 reveals several limitations. First, the model exhibits high precision for negative emotions but relatively lower precision for positive and neutral categories. Second, while recall is high for positive emotions, it drops substantially for neutral and negative emotions, suggesting difficulty in capturing subtle or implicit group affect that is context-dependent and distributed across individuals. These findings align with prior observations that zero-shot VLMs favor globally salient or linguistically frequent emotions while struggling with fine-grained collective affect reasoning. This performance gap highlights the importance of emotion-aware vision–language alignment optimized jointly through supervised classification and image–text contrastive learning for group-level emotion recognition.

4.8. Cross-dataset testing

To further evaluate generalization under dataset shifts, we conduct cross-dataset experiments, training models on one dataset and directly testing on another without fine-tuning.

4.8.1. Cross-dataset results on GAF2 and GroupEmoW

Table 11 reports cross-dataset results between GAF2 and GroupEmoW. When trained on GAF2 and tested on GroupEmoW,

Table 12
Cross-dataset recognition accuracy on the GAF3 and GroupEmoW datasets. Best results are in bold. (Unit: %).

Methods	GAF3→GroupEmoW	GroupEmoW→GAF3
Wang et al. [8]	77.86	70.00
ED-VLAlign (Ours)	**88.69**	**81.55**

our method achieves 88.74%, compared with 81.30% for Li et al. [36], showing a clear advantage for ED-VLAlign under this specific transfer direction without implying a general limitation of all visual-only approaches. Conversely, training on GroupEmoW and testing on GAF2, our approach attains 82.57%, slightly higher than Li et al. [36] (82.41%), showing more stable performance across dataset shifts. This asymmetry reflects the higher semantic diversity and richer scene coverage in GroupEmoW, which better supports transferable emotional language representations.

4.8.2. Cross-dataset results on GAF3 and GroupEmoW

A similar trend appears in cross-dataset experiments between GAF3 and GroupEmoW (Table 12). Training on GAF3 and testing on GroupEmoW, our method reaches 88.69%, outperforming Wang et al. [8] (77.86%), indicating stronger transfer performance than Wang et al. for this dataset pair. When reversing the datasets, our approach maintains 81.55% accuracy, whereas Wang et al. [8] obtain 70.00%, again indicating stronger cross-dataset transfer relative to Wang et al. in the reverse setting.

4.9. Visualization analysis

To further examine the behavior of our vision–language framework and compare it with representative vision-driven methods, we visualize both non-normalized and row-normalized confusion matrices on GAF2, GAF3, and GroupEmoW. The non-normalized matrices expose the absolute number of correct and incorrect predictions, whereas the normalized matrices reveal class-conditional error rates independently of class frequency. We compare our approach with two strong baselines: Wang et al. [8], which use graph-based visual interaction modeling, and Zhu et al. [28], which apply hierarchical relation inference with key-role guidance.

Fig. 4(a) complements the percentage view by showing the absolute scale of each error pattern. For ED-VLAlign on GAF2, 160 Neutral samples are assigned to Negative, whereas 35 Negative samples are assigned to Neutral. The corresponding normalized rates in Fig. 4(b) are approximately 22% and 6%, respectively, confirming that Neutral-to-Negative confusion is the dominant residual error rather than an artifact of class frequency. At the same time, the proposed model correctly classifies 680 Positive, 530 Neutral, and 501 Negative GAF2 samples.

The normalized matrices show that the proposed method generally produces strong diagonal responses, but the improvement is not uniform across every class and dataset. On GAF2, the Positive and Negative recalls reach 87.97% and 88.83%, and the Negative-to-Neutral error falls to 6.21%, compared with 20.11% for Wang et al. and 17.73% for Zhu et al. However, 21.98% of Neutral samples are still predicted as Negative. This residual confusion indicates that low-arousal or contextually ambiguous group scenes can be interpreted as negative when the textual description emphasizes concern, conflict, or subdued behavior. Thus, the class-separability benefit should be understood as dataset- and class-dependent rather than universal.

On GAF3, ED-VLAlign improves Positive recall to 91.30% and achieves a more balanced diagonal pattern than Wang et al., although Neutral and Negative confusion remain. The count matrix shows 156 Neutral-to-Negative and 179 Negative-to-Neutral errors, consistent with the bidirectional ambiguity visible in the normalized panel. On GroupEmoW, the diagonal values of 95.63%, 86.23%, and 93.27% show the most consistent separation among the three classes, corresponding to 635, 426, and 402 correctly classified samples. Overall, emotional descriptions reduce several visually driven confusions, especially Negative-to-Neutral errors, while the GAF2 Neutral-to-Negative pattern identifies a clear limitation of the current semantic alignment strategy.

4.10. Efficiency analysis

We profile the parameter scale, floating-point operations, and inference latency of the complete two-stage pipeline. Fig. 5 reports on

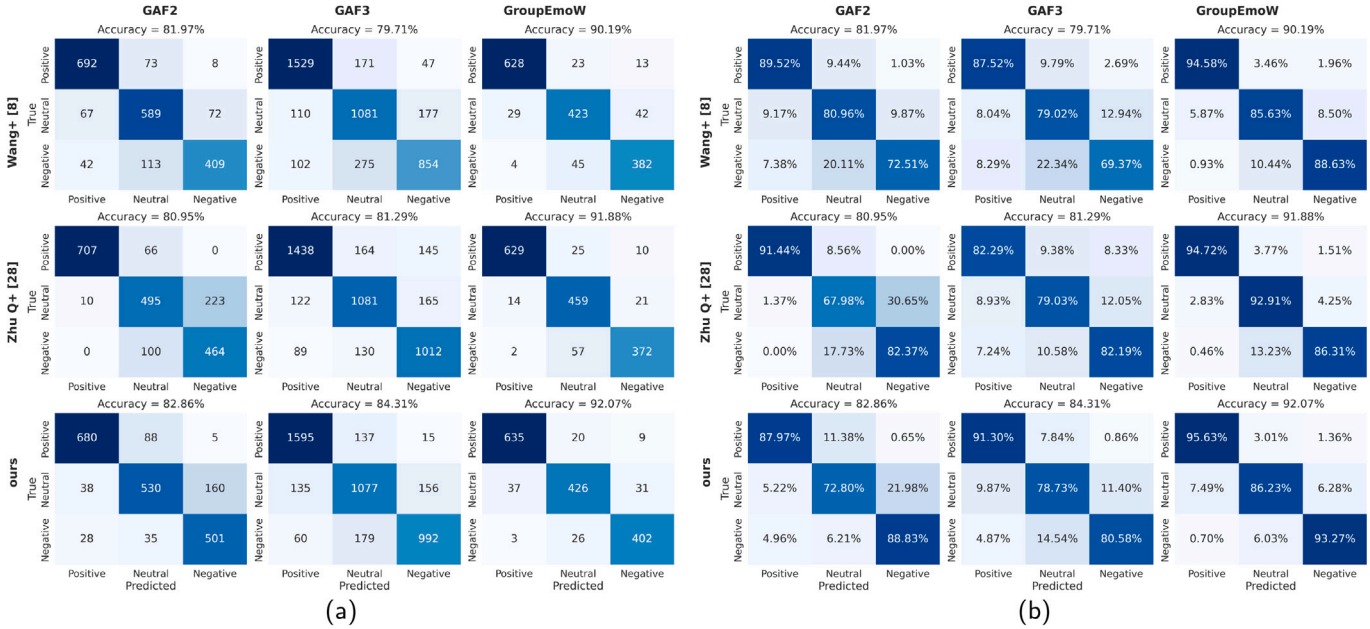


Fig. 4. Non-normalized and normalized confusion matrices on GAF2, GAF3, and GroupEmoW. (a) each cell reports the number of samples. (b) each row is normalized by its ground-truth class and is reported as a percentage. Rows denote ground-truth classes and columns denote predicted classes. Citation numbers are given in the surrounding text rather than embedded in the figure to ensure consistency with the reference list.

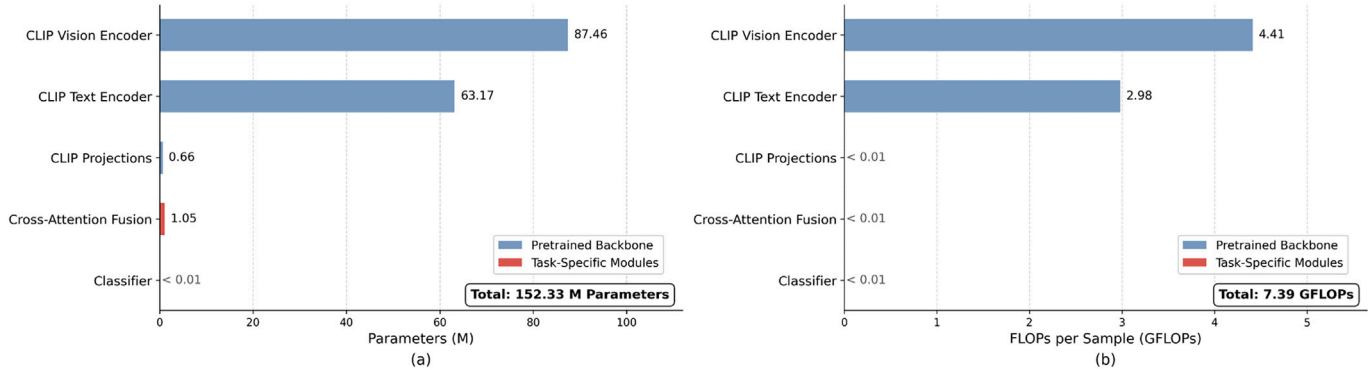


Fig. 5. Parameter and computation profiles of the downstream ED-VLAlign recognition model. (a) parameter count by component. (b) FLOPs per image–text sample. The hosted GLM-4.1 V-Thinking-Flash description generator is excluded because its internal parameter count and FLOP profile are not disclosed.

Table 13
Parameter scale, computational cost, and inference latency of the two-stage ED-VLAlign pipeline. Latency is reported as mean±SD over 100 runs. The symbol "–" denotes an unavailable or undisclosed value.

Stage	Parameters	FLOPs per sample	Inference latency
GLM-4.1 V-Thinking-Flash description generation	–	–	3544.6 ± 1227.8 ms
Downstream ED-VLAlign recognition	152.33 M	7.39 G	26.9 ± 0.7 ms
Task-specific fusion and classifier	≈1.05 M	<0.02 G	Included above
Sequential end-to-end pipeline	–	–	≈3571.5 ms

the downstream ED-VLAlign recognition model, excluding the hosted GLM description generator whose internal architecture and serving configuration are not disclosed. The downstream model contains 152.33 M parameters and requires 7.39 GFLOPs for one image–text sample. The pretrained CLIP vision and text encoders dominate this cost, with 87.46 M and 63.17 M parameters and 4.41 and 2.98 GFLOPs, respectively. The task-specific cross-attention fusion and classifier add approximately 1.05 M parameters, less than 0.7% of the downstream total, and each contributes less than 0.01 GFLOPs.

The latency results in Table 13 separate description generation from downstream recognition. GLM-4.1 V-Thinking-Flash was measured through its hosted API for 100 requests, yielding 3544.6 ± 1227.8 ms per description; this value includes service and network variability. The downstream ED-VLAlign model was measured with CUDA acceleration over 100 runs and requires 26.9 ± 0.7 ms per sample. Adding the two stage means gives an approximate sequential latency of 3571.5 ms, of which description generation accounts for 99.25%. When descriptions are generated and cached offline, the downstream recognition stage can operate at approximately 37 samples per second; online end-to-end use is instead constrained by the hosted VLM.

This profile clarifies the performance–efficiency trade-off. The accuracy and macro-precision gains in Fig. 3 are obtained with little task-specific overhead beyond the pretrained CLIP backbone, but the semantic description stage introduces a substantial latency cost relative to vision-only recognition. The current system is therefore most suitable for offline annotation, multimedia indexing, or applications in which descriptions can be precomputed and reused. Real-time deployment would require a smaller locally hosted VLM, description caching, distillation, or other parameter-efficient alternatives. The undisclosed parameter and FLOP scale of the hosted generator are reported as unavailable rather than estimated.

5. Conclusion

This paper proposes ED-VLAlign, a semantic emotion-guided framework for group emotion recognition that integrates emotional language modeling with visual representation learning. By using VLMs to generate group-level emotional descriptions and reformulating group emotion recognition as an image/video–text alignment task, the proposed method captures high-level semantic information beyond surface-level visual cues. Experiments on four datasets, including the video-based VGAF benchmark, show that the framework remains effective for both static images and temporally pooled video representations. Visualization analyses indicate reduced confusion under several ambiguous conditions, while also revealing persistent Neutral-to-Negative confusion on GAF2. Controlled comparisons further show that emotion-oriented descriptions outperform generic image descriptions, residual visual fusion preserves complementary evidence beyond the cross-attention output, and the framework remains competitive with descriptions produced by three alternative VLM families. Efficiency profiling shows that the task-specific fusion adds little overhead to the CLIP backbone, whereas hosted description generation dominates sequential inference latency, making the current implementation most suitable for offline or cached-description use. The repeated-run significance analysis on GAF2 further shows that the accuracy advantage over the reproduced Wang et al. baseline is statistically significant across Welch's t-test, the Mann–Whitney U test, and a permutation test. These results suggest that vision–language alignment is a promising direction for advancing group emotion recognition. Future work will explore finer-grained emotional categories, stronger temporal modeling beyond uniform frame averaging, robustness to inaccurate or semantically noisy descriptions, and smaller locally deployable description generators.

CRediT authorship contribution statement

Hongxi Yi: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Dong Zhang:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Yufei Xie:** Writing – review & editing, Resources, Formal analysis, Data curation. **Wenxing Liu:** Methodology, Conceptualization. **Ming Li:** Supervision, Funding acquisition. **Dah-Jye Lee:** Writing – review & editing, Supervision.

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.

Data availability

Data will be made available upon request.

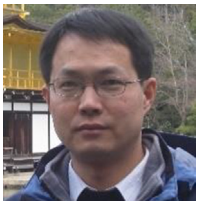
References

[1] T. Kuppens, V.Y. Yzerbyt, Group-based emotions: the impact of social identity on appraisals, emotions, and behaviors, Basic Appl. Soc. Psychol. 34 (1) (2012) 20–33.
[2] A. Goldenberg, D. Garcia, E. Halperin, J.J. Gross, Collective emotions, Curr. Dir. Psychol. Sci. 29 (2) (2020) 154–160.
[3] A. Dhall, R. Goecke, S. Ghosh, et al., From individual to group-level emotion recognition: EmotiW 5.0, in: Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), 2017, pp. 524–528.
[4] P.C. Ellsworth, K.R. Scherer, Appraisal processes in emotion, in: Handbook of Affective Sciences, Oxford University Press, Oxford, U.K., 2003.
[5] A. Abbas, S.K. Chalup, Group emotion recognition in the wild by combining deep neural networks for facial expression classification and scene-context analysis, in: Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), 2017, pp. 561–568.
[6] L. Tan, K. Zhang, K. Wang, et al., Group emotion recognition with individual facial emotion CNNs and global image based CNNs, in: Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), 2017, pp. 549–552.
[7] R. Kosti, J.M. Alvarez, A. Recasens, et al., Emotion recognition in context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1667–1675.
[8] X. Wang, D. Zhang, D.-J. Lee, Implementing the affective mechanism for group emotion recognition with a new graph convolutional network architecture, IEEE Trans. Affect. Comput. 15 (3) (2023) 1104–1115.
[9] X. Tang, Y. Gu, Influence of leaders' emotional labor and its perceived appropriateness on employees' emotional labor, Behav. Sci. 14 (5) (2024) 413.
[10] Q. Zhu, Q. Mao, J. Zhang, et al., Towards a robust group-level emotion recognition via uncertainty-aware learning, IEEE Trans. Affect. Comput. 16 (3) (2025) 1970–1983.
[11] J.-B. Alayrac, J. Donahue, P. Luc, et al., Flamingo: a visual language model for few-shot learning, in: Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 23716–23736.
[12] Z. Lian, L. Sun, H. Sun, et al., GPT-4V with emotion: a zero-shot benchmark for generalized emotion recognition, Inf. Fusion 108 (2024) 102367.
[13] K. Fujii, D. Sugimura, T. Hamamoto, Hierarchical group-level emotion recognition in the wild, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2019, pp. 1–5.
[14] A.S. Khan, Z. Li, J. Cai, et al., Regional attention networks with context-aware fusion for group emotion recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1150–1159.
[15] H. Xie, M.X. Lee, T.J. Chen, et al., Most important person-guided dual-branch cross-patch attention for group affect recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20598–20608.
[16] W. Guo, Q. Zhu, Q. Mao, Joint multi-scale contextual and noise suppression for group emotion recognition, in: ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5.
[17] K.G. Quach, N. Le, C.N. Duong, et al., Non-volume preserving-based fusion to group-level emotion recognition on crowd videos, Pattern Recognit. 128 (2022) 108646.
[18] J. Li, D. Li, C. Xiong, et al., BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation, in: Proceedings of the International Conference on Machine Learning (ICML), 2022, pp. 12888–12900.
[19] W. Hong, W. Yu, X. Gu, et al., GLM-4.1 V-Thinking: towards versatile multimodal reasoning with scalable reinforcement learning, arXiv:2507.01006, 2025.
[20] A. Radford, J.W. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, in: Proceedings of the International Conference on Machine Learning (ICML), 2021, pp. 8748–8763.
[21] J. Li, R.R. Selvaraju, A. Gotmare, et al., Align before fuse: vision and language representation learning with momentum distillation, in: Advances in Neural Information Processing Systems (NeurIPS), vol. 34, 2021, pp. 9694–9705.
[22] Q. Xu, Y. Wei, S. Yuan, et al., Learning emotional prompt features with multiple views for visual emotion analysis, Inf. Fusion 108 (2024) 102366.
[23] S. Wu, D. He, X. Wang, et al., Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, 2025, pp. 1601–1609.
[24] B. Pan, K. Hirota, Z. Jia, Y. Dai, A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods, Neurocomputing 561 (2023) 126866, https://doi.org/10.1016/j.neucom.2023.126866
[25] A. Li, L. Xu, C. Ling, J. Zhang, P. Wang, EmoVerse: enhancing multimodal large language models for affective computing via multitask learning, Neurocomputing 650 (2025) 130810, https://doi.org/10.1016/j.neucom.2025.130810
[26] C. Cai, K. Zhang, Z. Hu, X. Lin, Z. Pan, Prompt-based hybrid supervised contrastive learning for emotion recognition in conversation, Neurocomputing 647 (2025) 130453, https://doi.org/10.1016/j.neucom.2025.130453
[27] P. Kumar, S. Malik, B. Raman, et al., VISTANet: visual spoken textual additive net for interpretable multimodal emotion recognition, IEEE Trans. Affect. Comput. 16 (4) (2025) 2881–2893.
[28] Q. Zhu, Q. Mao, W. Dong, et al., Adaptive key role guided hierarchical relation inference for enhanced group-level emotion recognition, IEEE Trans. Affect. Comput. 17 (1) (2026) 366–378.
[29] X. Guo, B. Zhu, L.F. Polanía, et al., Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions, in: Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), 2018, pp. 635–639.
[30] G.A. Van Kleef, S. Côté, The social effects of emotions, Annu. Rev. Psychol. 73 (2022) 629–658.
[31] A. Dhall, A. Kaur, R. Goecke, et al., EmotiW 2018: audio-video, student engagement and group-level affect prediction, in: Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), 2018, pp. 653–656.
[32] X. Guo, L. Polanía, B. Zhu, et al., Graph neural networks for image understanding based on multiple cues: group emotion recognition and event recognition as use cases, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2921–2930.
[33] G. Sharma, A. Dhall, J. Cai, Audio-visual automatic group affect analysis, IEEE Trans. Affect. Comput. 14 (2) (2021) 1056–1069.
[34] W. Gong, Y. Wang, Y. Wu, et al., A hybrid fusion model for group-level emotion recognition in complex scenarios, Inf. Sci. 704 (2025) 121968.
[35] K. Fujii, D. Sugimura, T. Hamamoto, Hierarchical group-level emotion recognition, IEEE Trans. Multimed. 23 (2020) 3892–3906.
[36] C. Li, W. Huang, X. Yu, et al., Group emotion recognition using fuzzy emotional hierarchical shared space, IEEE Trans. Fuzzy Syst. 33 (11) (2025) 4083–4096.
[37] J. Xu, X. Huang, Group-level emotion recognition using hierarchical dual-branch cross transformer with semi-supervised learning, in: 2024 IEEE 4th International Conference on Software Engineering and Artificial Intelligence (SEAI), 2024, pp. 252–256.
[38] J.R. Pinto, T. Goncalves, C. Pinto, et al., Audiovisual classification of group emotion Valence using activity recognition networks, in: 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), 2020, pp. 114–119.
[39] M. Sun, J. Li, H. Feng, et al., Multi-modal fusion using spatio-temporal and static features for group emotion recognition, in: Proceedings of the 2020 International Conference on Multimodal Interaction, 2020, pp. 835–840.
[40] A. Petrova, D. Vaufreydaz, P. Dessus, Group-level emotion recognition using a unimodal privacy-safe non-individual approach, in: Proceedings of the 2020 International Conference on Multimodal Interaction, 2020, pp. 813–820.
[41] L. Evtodienko, Multimodal end-to-end group emotion recognition using cross-modal attention, arXiv:2111.05890, 2021.
[42] A.V. Savchenko, L.V. Savchenko, I. Makarov, Classifying emotions and engagement in online learning based on a single facial expression recognition neural network, IEEE Trans. Affect. Comput. 13 (4) (2022) 2132–2143.
[43] X. Wang, T. Chen, D. Zhang, A spatial-temporal graph convolutional network for video-based group emotion recognition, in: International Conference on Pattern Recognition, Springer Nature Switzerland, 2024, pp. 339–354.
[44] X. Huang, J. Xu, Multimodal spatiotemporal semisupervised transformer network for video-based group-level emotion recognition, IEEE Trans. Comput. Soc. Syst. 13 (2) (2026) 1719–1733.
[45] S. Bai, Y. Cai, R. Chen, K. Chen, et al., Qwen3-VL technical report, arXiv:2511.21631, 2025.
[46] W. Wang, Z. Gao, L. Gu, H. Pu, et al., InternVL3.5: advancing open-source multimodal models in versatility, reasoning, and efficiency, arXiv:2508.18265, 2025.
[47] H. Laurençon, A. Marafioti, V. Sanh, L. Tronchon, Building and better understanding vision–language models: insights and future directions, arXiv:2408.12637, 2024.

Author biography



Hongxi Yi received the BS degree from Sun Yat-sen University, China, in 2024. He is currently working toward the graduate degree at a School of Electronics and Information Technology, Sun Yat-sen University. His research interests include affective computing and computer vision.



Dong Zhang received the BSEE and MS degrees from Nanjing University, in 1999 and 2003, respectively, and the PhD degree from Sun Yat-sen University, in 2009. He is currently an associate professor at the School of Electronics and Information Technology, Sun Yat-sen University. His research interests include image processing, computer vision, affective computing, and information hiding.



Yufei Xie received an M.A. degree in Automation from Wuhan University of Technology, Hubei, China. He is currently pursuing a Ph.D. degree at the School of Electronics and Information Technology, Sun Yat-sen University, China. His research interests include sentiment analysis, large language models, and Vision-Language models.



Wenxing Liu received a master's degree in computer science from Chongqing University of Technology, China. He is currently pursuing a Ph.D. degree in computer science at Wuhan University, China. His research interests include autism spectrum disorder (ASD) diagnosis, video understanding and multimodal learning with large language models.



Ming Li received his Ph.D. in Electrical Engineering from the University of Southern California in 2013. He is a Full Professor in the Department of Artificial Intelligence at the Chinese University of Hong Kong, Shenzhen. He is also an Adjunct Professor in the School of Artificial Intelligence at Wuhan University. His research interests are in the areas of audio, speech and language processing as well as multimodal behavioral signal analysis and interpretation.



Dah-Jye Lee (Senior Member, IEEE) received the MS and PhD degrees in electrical engineering from Texas Tech University, Lubbock, TX, USA, in 1987 and 1990, respectively, and the MBA degree from Shenandoah University, Winchester, VA, USA, in 1999. He worked in the machine vision industry for 11 years before joining Brigham Young University (BYU), Provo, UT, USA, in 2001. He is currently a professor with the Department of Electrical and Computer Engineering, BYU. His research interests include artificial intelligence, robotic vision, highperformance visual computing, and visual inspection automation.