

Article

Large Language Model-Assisted Distillation–Fusion Framework for Visual Emotion Recognition

Yujun Ma , Yunjie Zeng , Zhiyuan Chen, Zhiwei Ye , Wen Zhou and Chunli Xiang 

School of Computer Science and Artificial Intelligence, Hubei University of Technology, Wuhan 430068, China; mayujun@hbut.edu.cn (Y.M.); 102411225@hbut.edu.cn (Y.Z.); 102301205@hbut.edu.cn (Z.C.); zw_mmwh@hbut.edu.cn (W.Z.); xiangchunli@hbut.edu.cn (C.X.)

* Correspondence: hgcsyzw@hbut.edu.cn

Abstract

Visual emotion recognition plays a critical role in human–computer interaction and mental health applications. Although existing Vision–Language Models (VLMs) alleviate the limitations of conventional vision models in high-level semantic understanding, they still face three main challenges: limited emotional semantic understanding, insufficient visual emotional perception capability, and high computational costs when deploying both models simultaneously. To address these issues, a large language model-assisted distillation–fusion framework (VERLADF) is proposed, which introduces emotion instruction data generated by GPT to fine-tune a VLM, thereby enhancing its emotional semantic understanding capability. Furthermore, we transfer the visual emotion discrimination knowledge of a conventional vision model into the VLM using a distillation module while keeping the VLM frozen during training, which reduces the computational costs. Following that, we design a fusion and prediction module that adaptively fuses predictions from the instruction-tuned VLM and the distillation module for final emotion recognition. The experimental results on the Abstract, ArtPhoto, Emotion6, and FI datasets demonstrate that VERLADF achieves recognition accuracies of 36.71%, 52.38%, 74.73%, and 79.69%, respectively, significantly outperforming many methods in the literature and demonstrating the effectiveness of the proposed framework.

Keywords: visual emotion recognition; vision-language model; knowledge distillation; proxy fusion; visual instruction tuning



Academic Editors: Qiyue Yin and Frank Werner

Received: 30 May 2026

Revised: 6 August 2026

Accepted: 7 August 2026

Published: 11 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Visual emotion recognition (VER) has been widely applied in areas such as mental health [1], social relationship reasoning [2], and advertising [3]. However, compared with traditional visual recognition tasks, emotion recognition involves higher complexity and uncertainty. Emotional expression is often highly subjective, and its visual cues are usually embedded in multiple visual elements, such as facial expressions, body postures, and surrounding environmental scenes. These emotional cues are scattered and subtle in the visual space, making it difficult for conventional models relying on single visual features to fully capture high-level emotional semantics, thereby limiting further improvement in emotion recognition performance [4,5]. Early convolutional neural network-based methods with attention or multi-scale mechanisms [6,7] improved emotion classification but lacked explicit emotional semantic hierarchies, limiting generalization. Vision Transformer (ViT) enhances the ability to capture long-range dependencies; however, vision models still

struggle to characterize high-level emotional semantics due to subjectivity, visual cue uncertainty, and limited annotated samples [8,9].

In recent years, the emergence of Vision–Language Models (VLMs) has introduced a new research paradigm for visual emotion recognition. By pre-training on large-scale image–text pairs, VLMs are capable of learning alignments between visual content and natural language semantics, thereby enabling cross-modal reasoning and high-level semantic understanding. Representative works such as Flamingo [10] and BLIP-2 [11] combine visual encoders with Large Language Models (LLMs), demonstrating strong generalization capabilities in zero-shot recognition and instruction-following tasks. Building on this, LLaVA [12] and InstructBLIP [13] further incorporate instruction tuning mechanisms, allowing models to better adapt to human linguistic expressions and perform complex visual semantic reasoning tasks. In the context of emotion understanding, some studies have begun exploring the application of VLMs to emotion recognition. For example, EmoVIT [14] employs emotion instruction data generated by GPT-4o [15] to fine-tune visual models via instruction tuning, thereby enhancing the model’s ability to align with emotional semantics. General-purpose VLMs such as Qwen2-VL [16], with their powerful cross-modal reasoning and semantic understanding capabilities, provide strong foundational support for visual emotion analysis. Several efforts [17,18] have extended VLMs to multimodal emotion recognition scenarios by jointly modeling visual, textual, and acoustic information, further improving their ability to comprehend emotional expressions. Despite these positive advances, the visual encoders of VLMs still exhibit insufficient visual emotional perception capability when handling complex visual emotional cues, limiting their ability to discriminate emotions. Furthermore, jointly deploying VLMs and vision models introduces substantial computational overhead, and few studies have investigated a unified framework that effectively integrates the complementary strengths of the strong semantic understanding capability of VLMs and the complex emotion-discriminative capability of vision models.

To address these challenges, we propose a large language model-assisted distillation–fusion framework for visual emotion recognition (VERLADF). Previous instruction-tuning approaches, such as EmoVIT [14], primarily enhance emotional semantic understanding through language supervision. However, visual emotion recognition also relies heavily on the ability to perceive emotion-discriminative visual cues, such as color compositions, scene atmosphere, and object interactions. Such visual perception capability cannot be fully acquired through language supervision alone. In contrast, vision models are effective at capturing visual cues that distinguish different emotional categories, but still struggle to characterize high-level emotional semantics. Therefore, we propose a new solution, called VERLADF, that effectively integrates the semantic reasoning capability of instruction-tuned VLMs with the emotion-discriminative visual perception capability of vision models within a unified and computationally efficient framework. The proposed framework aims to simultaneously enhance the semantic understanding and visual emotional perception capabilities of VLMs by transferring emotion-discriminative knowledge from a vision model into an instruction-tuned VLM through knowledge distillation and adaptively integrating their complementary strengths via a fusion mechanism. Therefore, instead of solely improving emotional semantic understanding, VERLADF explicitly focuses on enhancing the visual emotional perception capability of VLMs. Furthermore, although knowledge distillation has been widely adopted for model compression and teacher–student knowledge transfer, existing approaches typically focus on training a compact student model to approximate or replace the teacher model. In contrast, VERLADF employs knowledge distillation to construct a lightweight emotion-specialized distillation module attached to the frozen visual encoder of an instruction-tuned VLM. Instead of learning an independent student model

from raw image inputs, the distillation module utilizes the visual representations extracted by the frozen VLM encoder and learns emotion-discriminative knowledge transferred from a vision teacher model through logits-level knowledge distillation, introducing only a small number of trainable parameters. The proposed fusion mechanism enables the framework to jointly exploit the complementary strengths of semantic reasoning from the VLM and emotion-discriminative visual perception from the distilled module.

Specifically, our framework first utilizes emotion instruction data generated by GPT-4o to fine-tune the VLM, thereby enhancing its emotional semantic understanding. Subsequently, a vision model is employed as the teacher model to transfer discriminative visual emotional perception knowledge to a lightweight distillation module via knowledge distillation, alleviating insufficient sensitivity to visual emotional cues. The lightweight distillation module is then attached to the visual encoder of the VLM while keeping the VLM frozen during training, effectively reducing the computational cost of jointly deploying two large models. Furthermore, a fusion and prediction module is introduced to adaptively integrate the outputs of the VLM and the distillation module, producing more comprehensive and reliable emotion recognition results.

The main contributions of this paper are summarized as follows:

- We introduce a novel large language model-assisted distillation–fusion framework that effectively integrates the strong semantic understanding capability of VLMs with the emotion-discriminative visual perception capability of vision models.
- Unlike previous methods, we design a knowledge distillation module to transfer the emotion expertise of the vision model into the instruction-tuned VLM, achieving this with only a minimal increase in parameters for training and lower computational cost.
- We propose a fusion and prediction module that adaptively fuses predictions from VLM and distillation module, effectively leveraging the complementary strengths of different models.
- Extensive experiments demonstrate that our approach outperforms many existing VER approaches on four representative visual emotion datasets.

This paper is organized as follows: Section 2 reviews related work on VLMs and distillation. Section 3 presents the proposed VERLADF framework. Section 4 reports the experimental results and discussion. Section 5 concludes this paper.

2. Related Work

2.1. Foundational VLMs for Emotion Recognition

VLMs have introduced a new research paradigm for visual emotion recognition by establishing semantic alignment between visual content and natural language. CLIP [19] is one of the representative works in this field. Through large-scale image–text contrastive learning, CLIP learns transferable cross-modal representations and demonstrates strong generalization ability in various downstream visual understanding tasks. In affective computing, the cross-modal alignment capability of CLIP provides a new foundation for visual emotion recognition by enabling models to associate visual content with semantic descriptions.

Building on CLIP, subsequent studies further explored the integration of visual encoders and language generation modules to enhance multimodal understanding and reasoning capabilities. BLIP [20] and BLIP-2 [11] introduce cross-modal transformer architectures to enable deep interaction between visual features and language representations, allowing the models to support image captioning, visual question answering, and image–text matching simultaneously. Flamingo [10] adopts a frozen visual encoder and language model architecture while injecting visual information through cross-attention modules, demonstrating strong few-shot generalization ability across multiple multimodal tasks. These studies show that VLMs can effectively bridge the gap between visual repre-

sentation learning and semantic understanding, providing important support for high-level affective analysis.

However, most of the prior studies have been limited to object-scene semantic alignment; they often struggle to capture subtle emotional relationships and subjective emotional semantics in visual emotion recognition tasks. These limitations motivate VLMs enhanced with LLMs, which further improve the emotional semantic understanding capability of VLMs.

2.2. LLM-Enhanced VLMs for Emotion Recognition

With the rapid development of LLMs, recent studies have introduced instruction tuning and generative reasoning mechanisms into VLMs, significantly improving their multimodal semantic understanding capability. Models such as LLaVA [12] and InstructBLIP [13] construct large-scale image–text instruction datasets to align visual representations with natural language instructions, enabling models to perform reasoning-oriented multimodal tasks. Compared with traditional VLMs, these instruction-tuned frameworks exhibit stronger capabilities in semantic reasoning, visual dialogue, and complex language generation.

General-purpose multimodal foundation models such as Qwen2-VL [16] and GPT-4V [15] further extend the reasoning and understanding capabilities of VLMs. By jointly training LLMs and high-capacity visual encoders on large-scale multimodal datasets, these models achieve strong performance in image understanding, multimodal reasoning, and language generation. Their powerful semantic modeling capability provides a new technical foundation for high-level cognitive tasks such as emotion understanding and psychological analysis.

Building on these advances, several studies have explored the application of LLM-enhanced VLMs to emotion recognition tasks. EmoVIT [14] employs emotion instruction data generated by GPT-4o to fine-tune visual models through instruction tuning, enabling better alignment between visual inputs and emotional semantic labels. Other studies [21–23] further investigate prompt engineering and multimodal reasoning strategies for emotion understanding under few-shot or zero-shot settings.

In spite of recent advances in LLM-enhanced VLMs, their effectiveness in visual emotion recognition remains constrained [24–26]. Therefore, how to effectively integrate the complementary strengths of VLMs and conventional vision models to enhance visual emotion analysis remains a key challenge [27,28].

2.3. Knowledge Distillation

Knowledge distillation is an effective model compression and representation transfer technique that transfers knowledge from a teacher model to a student model [29]. In computer vision, knowledge distillation has been widely adopted to compress convolutional neural networks and Transformer-based models while maintaining competitive performance. Existing studies [30–32] demonstrate that aligning the feature distributions, activation patterns, or semantic structures between teacher and student models can effectively enhance the representation capability and generalization performance of lightweight architectures. Moreover, knowledge distillation has shown effectiveness in challenging scenarios such as few-shot learning, long-tail recognition, and cross-domain transfer.

With the emergence of large-scale multimodal models, recent research has extended knowledge distillation from traditional single-modal compression to multimodal collaborative distillation and cross-modal knowledge transfer. Several studies [32,33] attempt to transfer semantic reasoning capability and cross-modal alignment knowledge from large multimodal foundation models into compact architectures by aligning multimodal embedding spaces, intermediate semantic representations, or output distributions. In emotion

recognition, CDGKD [32] transfers cross-modal interaction knowledge from multimodal teacher models to lightweight student models through multi-level distillation paths and modality rebalancing mechanisms, achieving strong performance on multiple emotion recognition benchmarks. Despite recent progress, the application of knowledge distillation to visual emotion recognition remains relatively underexplored, particularly in challenging emotion recognition scenarios.

3. Proposed Method

In this section, we present the details of the proposed VERLADF framework. The framework is illustrated in Figure 1, which comprises three stages. First, a VLM is instruction-tuned using data constructed with GPT-4o and the fine-tuned VLM is then kept frozen. Secondly, a ViT model is employed as the teacher to transfer emotional knowledge to the distillation module via knowledge distillation. Finally, the fusion and prediction module adaptively integrates the outputs of the distillation module and the VLM to derive the final output.

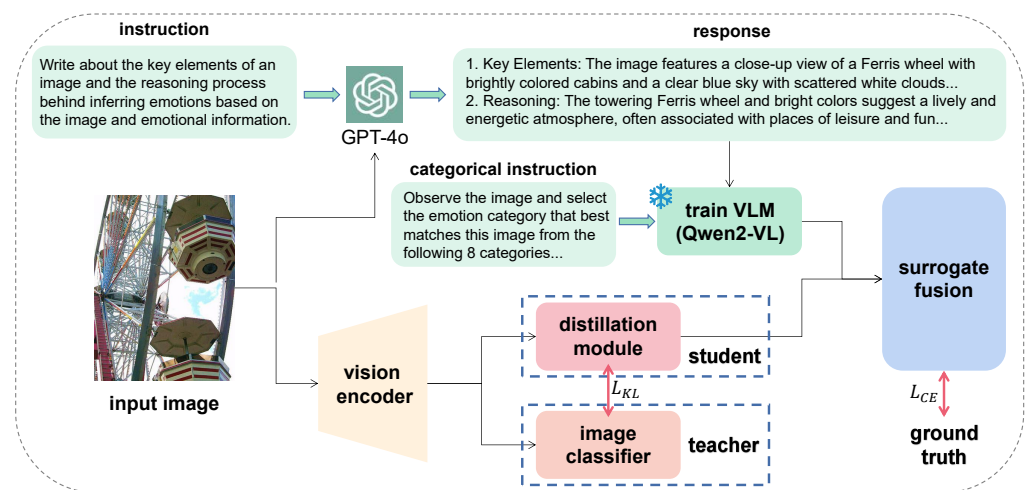


Figure 1. Overview of the proposed large language model-assisted distillation–fusion framework for visual emotion recognition (VERLADF).

3.1. LLM-Assisted Instruction Tuning VLM

Unlike conventional vision models that rely on low-level visual features, VLMs establish semantic connections between vision and language through cross-modal pre-training, thereby enhancing the understanding of emotional expressions. Following the instruction construction strategy of EmoViT [14], we adapt a pre-trained VLM to the visual emotion recognition task via instruction tuning. Specifically, Qwen2-VL-7B [16] is adopted as the base VLM, and an emotion-oriented instruction dataset is constructed. During data construction, GPT-4o is employed to generate descriptive responses based on predefined instructions, as illustrated in Figure 2. The resulting training samples consist of image–instruction–response triplets. The instructions prompt the model to attend to key visual elements and emotional cues, while the responses provide linguistic descriptions of image content and associated emotions. To comprehensively capture different aspects of visual emotion understanding, the instruction dataset includes multiple types. Emotion classification instructions require selecting the most appropriate label from the predefined categories, enabling the evaluation of recognition performance. In addition, conversational and reasoning instructions encourage the model to describe and reason about emotions, thereby enhancing semantic understanding.

To further enhance the model’s ability to understand emotional cues and generate more interpretable descriptions, we further introduce a structured emotion reasoning

prompt template that explicitly guides the model to articulate key visual elements and the associated reasoning process, which is absent in prior instruction-tuning strategies. The template guides the model to attend to salient visual elements and reason about potential emotions from scene-level semantic information. Specifically, the prompt provides the associated emotion category and requires the model to reason over a set of candidate emotions, including amusement, excitement, sadness, fear, contentment, anger, disgust, and awe. Unlike conventional emotion classification tasks, the template does not directly require the model to output an emotion label; instead, it encourages the model to describe key visual elements and the reasoning process that leads to emotion prediction. This design promotes learning explicit associations between visual cues and emotional semantics in the language space.

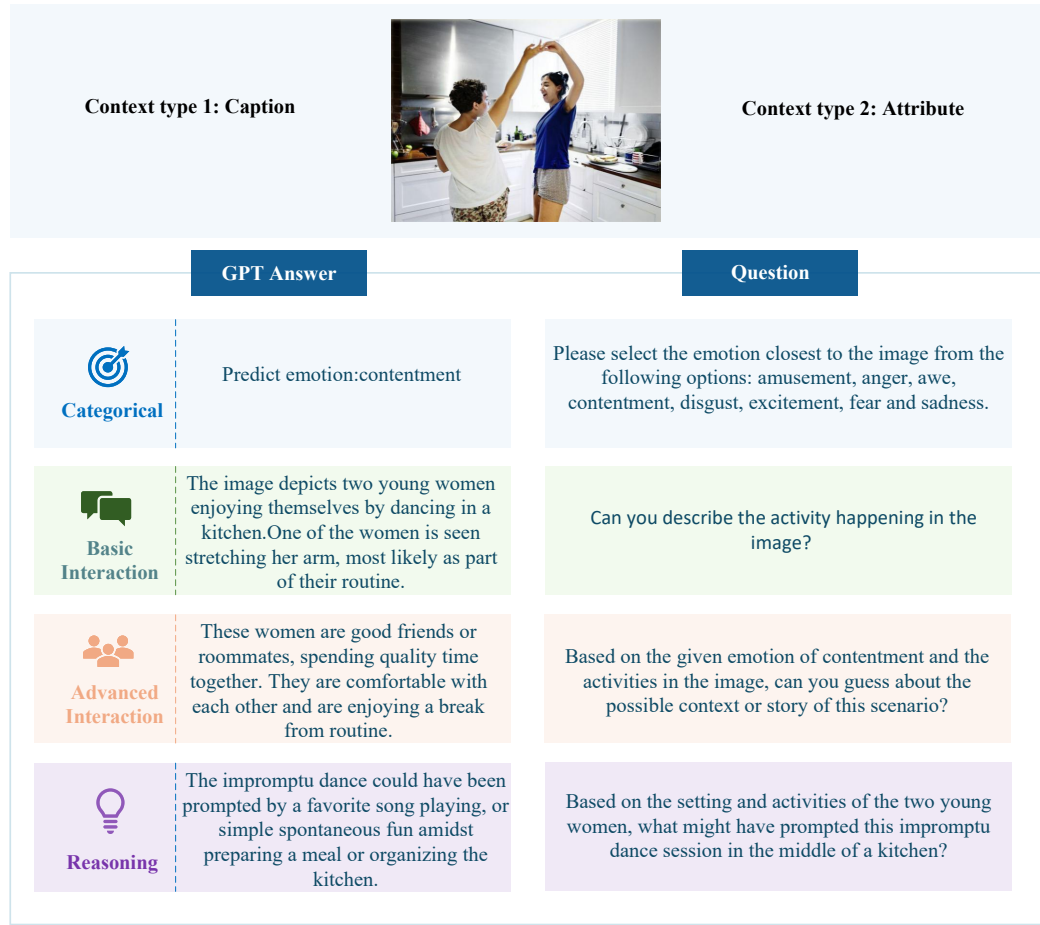


Figure 2. The sample of visual emotion instruction data generated with the assistance of GPT-4o.

The prompt template consists of two components: (1) key elements, which describe emotion-relevant visual cues such as human posture, object attributes, and scene context; and (2) reasoning, which infers potential emotional sources based on these elements to form a complete understanding process. This structured prompting strategy encourages the model to explicitly articulate the reasoning process rather than merely predicting labels, thereby enhancing its semantic expressiveness in visual emotion understanding. After instruction tuning, the VLM parameters are frozen during subsequent training to preserve cross-modal semantic understanding, allowing the model to serve as a stable high-level semantic encoder for downstream knowledge distillation and feature fusion.

3.2. Lightweight Distillation Module

We propose a lightweight distillation module built upon the frozen visual encoder of Qwen2-VL-7B. By reusing visual features extracted from the frozen visual encoder, the distillation module benefits from rich general-purpose visual representations without introducing additional feature extraction costs. However, these representations are not optimized for emotion recognition. In contrast, the fine-tuned ViT teacher learns emotion-discriminative knowledge from emotion recognition datasets. Therefore, to compensate for the limited emotion discrimination capability of the frozen visual representations, we design a response-based distillation strategy, where the distillation module aligns its output distribution with that of the teacher through distillation losses. By matching the teacher's output distribution, the distillation head inherits emotion-sensitive decision patterns learned by the ViT teacher, enabling it to more effectively associate visual patterns with emotion categories. Specifically, frozen Qwen2-VL-7B visual encoder extracts token-level visual representations of the image, which are aggregated by the distillation module via global average pooling into a fixed-dimensional feature vector. A trainable MLP distillation head then maps this vector to emotion logits. Training minimizes a joint loss function combination of (1) the temperature-scaled KL divergence loss between the distillation-head logits and those of the ViT teacher, and (2) a cross-entropy loss with respect to the ground-truth emotion labels. Thus, the MLP head is aligned with the ViT teacher through distribution matching while being directly supervised by the annotated labels. Notably, the teacher model is only required during training and is completely removed during inference, ensuring that no additional inference-time computational overhead is introduced.

Architecture of the lightweight distillation module: The proposed distillation module consists of two components: (1) a frozen VLM visual encoder, and (2) a lightweight MLP-based distillation head. Importantly, the visual encoder is not part of the trainable module. It serves as a shared frozen feature extractor.

Given an input image x , the frozen VLM visual encoder extracts patch-level visual representations:

$$\mathbf{F} = \mathcal{F}_{\text{VLM}}(x) \in \mathbb{R}^{L \times 3584}, \quad (1)$$

where L denotes the number of visual tokens and 3584 is the hidden dimension. Here, x is the input RGB image; $\mathcal{F}_{\text{VLM}}(\cdot)$ denotes the frozen Qwen2-VL visual encoder; L is determined by the image resolution and patch grid; and 3584 is the fixed output channel dimension of the Qwen2-VL vision merger. To obtain a global representation, we apply mean pooling:

$$\mathbf{f} = \frac{1}{L} \sum_{i=1}^L \mathbf{F}_i \in \mathbb{R}^{3584}. \quad (2)$$

where $\mathbf{F}_i \in \mathbb{R}^{3584}$ is the feature vector of the i -th visual token, and $\mathbf{f}$ is the pooled global representation fed to the distillation head. The pooled feature is then fed into a lightweight multi-layer perceptron (MLP), producing emotion logits:

$$\mathbf{z}_s = \text{MLP}(\mathbf{f}) \in \mathbb{R}^C. \quad (3)$$

where $\mathbf{z}_s = [z_s^{(1)}, \dots, z_s^{(C)}]^\top$ denotes the emotion logits produced by the distillation module over C emotion classes, and C depends on the dataset (e.g., $C = 6$ for Emotion6).

Joint loss function: The lightweight distillation module is optimized using a joint loss function that consists of knowledge distillation loss and cross-entropy loss. To enable the distillation module to learn the prediction behavior of the teacher model, a knowledge distillation loss is introduced to constrain the output distribution of the MLP-based distillation

head against that of the teacher model. This loss minimizes the KL divergence between the two distributions, allowing the MLP head to gradually approach the prediction pattern of the teacher during training, thereby effectively inheriting the emotion-discriminative knowledge embedded in the teacher model. The KL divergence loss is defined as follows:

$$L_{KD} = \tau^2 \sum_{i=1}^C p_t^\tau(i) \log \frac{p_t^\tau(i)}{p_s^\tau(i)}, \quad (4)$$

where $p_t^\tau(i)$ and $p_s^\tau(i)$ denote the temperature-scaled probability distributions of the teacher model and distillation module. More explicitly:

$$p_t^\tau(i) = \frac{\exp(z_t^{(i)} / \tau)}{\sum_{j=1}^C \exp(z_t^{(j)} / \tau)}, \quad (5)$$

$$p_s^\tau(i) = \frac{\exp(z_s^{(i)} / \tau)}{\sum_{j=1}^C \exp(z_s^{(j)} / \tau)}, \quad (6)$$

where z_t denotes the logits produced by the ViT teacher model, and z_s denotes the logits produced by the distillation module; $i \in \{1, \dots, C\}$ indexes emotion classes, and the default temperature is $\tau = 2.0$.

On this basis, the cross-entropy loss [34] is introduced as a supervised learning objective. The cross-entropy measures the discrepancy between the model's predicted probability distribution and the ground-truth label distribution. By penalizing incorrect predictions, it guides the model to optimize toward the correct class direction. It is defined as follows:

$$L_{CE} = - \sum_{i=1} y_i \log p_s(i) \quad (7)$$

where y_i denotes the one-hot encoding of the ground-truth label for the i emotion category, and $p_s(i)$ represents the probability predicted by the model for that category.

Finally, to combine the advantages of distillation learning and supervised learning, the overall training objective is defined as a weighted combination of the KL divergence loss and the cross-entropy loss.

$$L_{\text{total}} = \alpha L_{KD} + (1 - \alpha) L_{CE} \quad (8)$$

where α is a weighting factor used to balance the proportion of the two losses. Through multiple comparative experiments, it is determined that the model achieves better performance when $\alpha = 0.5$.

MLP design: The distillation head adopts a configurable architecture. In the default setting (depth = 1), it consists of a single hidden layer:

$$3584 \rightarrow 1024 \rightarrow C, \quad (9)$$

where the hidden layer is followed by GELU activation and dropout ($p = 0.1$). The final layer outputs logits. Specifically, depth = 1 means one hidden block (Linear + GELU + Dropout), the hidden width is 1024, dropout rate is $p = 0.1$, and the output dimension is C .

Only the parameters of the MLP head are trainable, while both the VLM visual encoder and the ViT teacher model remain frozen during training. This results in a parameter-efficient design with approximately 3.7 M trainable parameters under the default configuration. Compared with the VLM branch, which relies on autoregressive language generation,

the proposed distillation module directly produces vision-driven emotion logits in a single forward pass, enabling fast and efficient inference without text decoding.

3.3. Fusion and Prediction Module

To fully leverage the complementary advantages of the distillation module and the VLM in emotion recognition, we design a fusion and prediction module to jointly model and integrate emotion predictions from different branches. During inference, the input image, together with the corresponding text prompt, is fed into the VLM framework. The image is first processed by the visual encoder to extract visual features, which are then passed to both the VLM and the distillation module for emotion reasoning. The VLM performs high-level semantic inference based on its cross-modal understanding capability. Unlike conventional classifiers, the instruction-tuned VLM generates emotion categories in an autoregressive manner rather than directly outputting a probability distribution over all emotion classes. Meanwhile, the distillation module generates a probability distribution over emotion categories based on visual features, providing complementary information from a discriminative visual perspective. These two prediction branches capture emotional information at different levels.

To effectively integrate these predictions, the proposed framework adopts a proxy fusion strategy. The proxy fusion strategy adopts a lightweight Multi-Layer Perceptron (MLP) architecture to learn the mapping relationships between different prediction signals for adaptive feature fusion. Specifically, unlike conventional classifiers, the instruction-tuned Qwen2-VL-7B does not directly produce a probability distribution over predefined emotion categories. Instead, it predicts emotion categories through autoregressive text generation. Therefore, we represent the output of Qwen2-VL-7B as a one-hot semantic decision vector and use it as an explicit semantic cue for fusion. Furthermore, the objective of the fusion module is not to re-model the complete semantic reasoning process already performed by the instruction-tuned VLM, but to effectively combine the semantic decision produced by the VLM with the emotion-discriminative visual information generated by the distillation module. Therefore, the one-hot representation serves as a compact semantic decision cue, simplifying the learning objective of the fusion module while maintaining effective complementarity between semantic reasoning and visual discrimination. Subsequently, the outputs from the lightweight distillation module and the VLM are concatenated to form a joint representation, which is then fed into the MLP. Through non-linear transformations, the model learns the latent relationships between signals, resulting in a more robust emotion representation. The non-linear transformation is defined as follows:

$$F_{\text{fusion}} = \text{MLP}(P_{\text{VLM}} \| P_{\text{distill}}) \quad (10)$$

where $P_{\text{VLM}} \in \mathbb{R}^C$ is the one-hot prediction from the VLM, $P_{\text{distill}} \in \mathbb{R}^C$ is the probability output from the distillation module, $\|$ denotes vector concatenation, $\text{MLP}(\cdot)$ is a lightweight multi-layer perceptron, and F_{fusion} is the resulting fused emotion representation.

Unlike linear gating mechanisms, the MLP-based proxy fusion introduces non-linear interaction capabilities, enabling the model to learn more complex fusion strategies and automatically determine the fusion weights among different information sources in a data-driven manner. This design improves fusion flexibility while maintaining a lightweight model architecture.

From a complementary perspective, the frozen VLM obtained after LLM-assisted instruction tuning and the lightweight distillation module address fundamentally different aspects of visual emotion recognition. The VLM, benefiting from large-scale vision–language pre-training and emotion-specific instruction tuning, excels at high-level semantic reasoning, enabling it to understand the contextual narrative and abstract emotional implications

of a scene. However, visual emotion recognition is not solely a semantic understanding problem. Unlike general visual recognition tasks that primarily depend on object identity, emotion recognition relies heavily on subtle and fine-grained visual cues, such as facial expressions, color harmony, lighting conditions, compositional structures, and scene atmosphere. These cues often determine emotional categories even when semantic content remains similar. To address this limitation, the distillation module is trained under the supervision of a vision teacher model via knowledge distillation and specializes in capturing emotion-discriminative visual patterns. Since emotion categories are frequently distinguished by subtle visual characteristics rather than explicit semantic concepts, transferring emotion-discriminative knowledge from a vision model specifically optimized for emotion recognition enables the distillation module to become highly sensitive to these visual emotional cues. As a result, the distillation module provides complementary visual perception capability that cannot be fully acquired through instruction tuning alone.

The proxy fusion module subsequently reconciles these complementary signals, allowing the framework to draw upon semantic context when visual cues are ambiguous and rely on discriminative visual features when semantic reasoning alone is insufficient. This synergy is particularly important for visual emotion recognition, where emotional expressions span a wide spectrum from explicit cues (e.g., facial expressions) to implicit cues (e.g., abstract art). Relying solely on semantic reasoning may overlook subtle visual cues, while relying solely on visual patterns may fail to capture contextual emotional meanings. By combining both through the proposed fusion strategy, VERLADF is able to achieve more robust and discriminative emotion representations. A quantitative comparison of different fusion mechanisms is provided in Section 4.6.

4. Experiments and Discussion

4.1. Datasets

We employed four benchmark datasets widely used in VEA, namely Abstract [35], ArtPhoto [35], Emotion6 [36], and Flickr and Instagram (FI) [37], to train and evaluate the proposed framework. The statistical details of each dataset are summarized in Table 1. These datasets exhibit significant differences in emotion category definitions and image styles, enabling a comprehensive evaluation of the model across varying hierarchical structures of emotional expression and semantic complexity, thereby validating the robustness and generalization capability of the proposed method.

Table 1. Dataset statistics.

Dataset	Number of Images	Emotion Classes	Characteristics
FI	21,824	8	Large-scale social media
Emotion6	1980	6	Standard benchmark
ArtPhoto	806	8	Strong emotions
Abstract	228	8	Abstract art, implicit emotions

First, we adopt the FI dataset [37], which is annotated based on Mikels et al.’s eight basic emotion categories [38] (i.e., amusement, anger, awe, contentment, disgust, excitement, fear, and sadness) and contains 21,824 images. This dataset provides a comprehensive benchmark for evaluating the model’s ability to recognize diverse emotions. Second, the Emotion6 dataset [36] consists of 1980 images annotated according to Ekman’s six basic emotions (anger, disgust, fear, happiness, sadness, and surprise). This dataset has a moderate size and a well-defined category structure, making it widely used for evaluating model stability and classification performance under a standard emotion category. Third, the ArtPhoto dataset [35] contains high-quality images collected from professional photog-

raphy websites. It covers various artistic styles, including natural landscapes, portraits, architecture, and abstract compositions, and is annotated with eight emotion categories. Due to its complex compositions and more abstract emotional expressions, it is suitable for evaluating the model's ability to capture emotion cues in complex semantic scenes. Finally, the Abstract dataset [35] contains abstract art images that lack recognizable real-world objects or human figures. Emotions are conveyed through low-level visual features such as color, geometry, and composition. This dataset is used to assess the model's ability to recognize implicit emotional expressions, thereby posing greater challenges for model generalization and robustness. For the FI dataset, the data are divided into training, validation, and test sets. For Emotion6, ArtPhoto, and Abstract, the training and test sets are first constructed, after which validation sets are created using stratified sampling with a random seed of 42. The proportions of training images used for validation are 9%, 9%, and 8% for Emotion6, ArtPhoto, and Abstract, respectively, and 8.5% for FI. The resulting dataset statistics are summarized in Table 2.

In summary, these four datasets differ significantly in image style, emotion taxonomy, and semantic complexity, enabling a comprehensive evaluation of the proposed model across diverse visual emotion recognition scenarios.

Table 2. Dataset splits used in our experiments. The FI dataset is divided into training, validation, and test sets, with 8.5% of the training images held out as the validation set. For Emotion6, ArtPhoto, and Abstract, 9%, 9%, and 8% of the training images are held out as the validation set, respectively, via stratified sampling with a random seed of 42.

Dataset	Train	Val	Test	Total
Abstract	184	23	21	228
ArtPhoto	650	72	84	806
Emotion6	1604	178	198	1980
FI	17,459	2039	2326	21,824

4.2. Experimental Settings

Detailed training configurations are listed in Table 3, and accuracy is used as an evaluation metric.

Table 3. Hyperparameter settings.

Model/Module	Epoch	Batch Size	Learning Rate
Vision Transformer Model	10	10	0.00003
Distillation Module	5	1	0.00001
Qwen2-VL Model	5	1	0.0001
Fusion and Prediction Module	10	1	0.0001

As the base VLM, we used Qwen2-VL-7B [16]. Pre-trained on large-scale multimodal data, Qwen2-VL-7B demonstrates strong performance across a variety of vision–language understanding and reasoning tasks. To enhance its ability in capturing emotion-related semantics, the model is fine-tuned on the constructed emotion instruction dataset using the AdamW optimizer, with a learning rate of 1×10^{-4} . After fine-tuning, the model parameters are frozen during subsequent training to preserve its cross-modal reasoning ability and mitigate the risk of overfitting.

In the knowledge distillation stage, a ViT model is employed as the teacher model. Compared with traditional neural networks, ViT exhibits superior capability in visual representation learning and modeling global dependencies, thereby providing higher-quality supervision signals for distillation. Each ViT teacher model is independently fine-tuned

on the corresponding dataset to better capture domain-specific emotion characteristics. The distillation module learns emotion-discriminative knowledge from the ViT teacher through response-based knowledge distillation. The learning rate is set to 1×10^{-5} . During distillation training, only the parameters of the MLP head are updated, while all other VLM parameters remain frozen. This design ensures improved visual emotional perception capability while maintaining low computational complexity and stable inference performance. In addition, our VERLADF is implemented in PyTorch 2.5.1 and all experiments are conducted on NVIDIA RTX 3090 GPU.

4.3. Results

The quantitative results on four visual emotion recognition datasets, namely Abstract, ArtPhoto, Emotion6, and FI, are summarized in Table 4. Figure 3 provides a visual comparison of different methods across these datasets, where three representative results are selected from Table 4 for visualization. To ensure result stability and reproducibility, all reported values are averaged over multiple independent runs. Overall, the proposed VERLADF achieves the best performance across all four visual emotion recognition datasets, consistently outperforming existing mainstream VLMs. The results of SEMP and AIMER were obtained under the same experimental protocol. These results demonstrate the effectiveness of the proposed distillation–fusion framework and suggest that combining instruction-tuned semantic reasoning with discriminative visual feature learning provides a robust solution for visual emotion recognition.

Table 4. Comparison of accuracy of different methods (%). For VERLADF, mean accuracy over three independent runs is reported.

Method	Abstract	ArtPhoto	Emotion6	FI
Flamingo [10]	3.57	17.50	21.67	14.91
BLIP2 [13]	28.57	36.25	50.00	57.72
LLaVA [12]	19.54	36.25	49.44	56.04
InstructBLIP [13]	21.42	26.25	46.11	37.97
Qwen2-VL-7B [16]	32.14	40.48	38.55	53.54
SEPM [39]	26.43	42.39	54.21	71.35
AIMER [22]	34.82	46.72	62.33	79.63
EmoVIT [14]	32.34	44.90	57.81	68.09
VERLADF (Ours)	36.71	52.38	74.73	79.69

To better understand the behavior of the proposed method under different levels of semantic complexity, we first analyze its performance on Abstract and ArtPhoto, which mainly involve weak or implicit emotional cues. On the Abstract dataset, VERLADF achieves 36.71% accuracy, surpassing EmoVIT by 4.37 percentage points. Since this dataset contains abstract artistic images without explicit semantic objects, emotion recognition relies heavily on low-level visual patterns such as color and composition. The improvement demonstrates that combining high-level semantic reasoning from the VLM with discriminative features learned through the distillation module effectively enhances emotion perception in highly abstract scenarios. Similarly, on the ArtPhoto dataset, VERLADF achieves 52.38% accuracy, outperforming EmoVIT by 7.48 percentage points. Due to the complex artistic compositions and implicit emotional expressions, conventional vision models often struggle to capture emotional semantics. The proposed framework benefits from cross-modal semantic understanding, enabling more accurate modeling of emotionally relevant visual cues in artistic scenes.

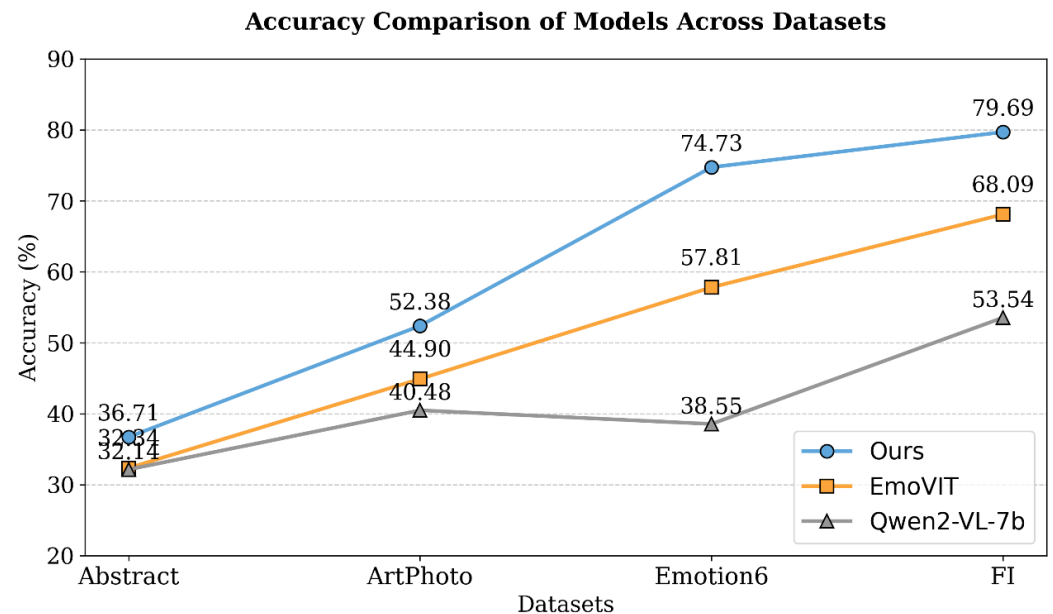


Figure 3. Visual comparison of the performance of different methods across datasets. The three representative results displayed in this figure are selected from the quantitative results reported in Table 4.

We further evaluate VERLADF on Emotion6 and FI, which contain more structured labels and more realistic visual content. On Emotion6, VERLADF achieves 74.73% accuracy, significantly improving over EmoVIT by 16.92 percentage points, representing the largest gain among all datasets. This improvement suggests that the proposed framework effectively alleviates data scarcity and enhances the learning of stable emotion representations under limited supervision. On the large-scale FI dataset, VERLADF achieves 79.69% accuracy, outperforming EmoVIT by 11.60 percentage points. Since FI consists of real-world social media images with complex backgrounds and diverse semantic contexts, these results further demonstrate the robustness and scalability of the proposed method in realistic scenarios.

To further validate the effectiveness of VERLADF, we compare it with several general-purpose VLMs, including Flamingo, BLIP-2, LLaVA, and InstructBLIP. Although these models show strong capabilities in multimodal understanding and reasoning tasks, their performance on emotion recognition remains limited without task-specific adaptation. This indicates that general semantic understanding alone is insufficient for capturing subtle emotional cues in visual content. In contrast, the proposed emotion-aware instruction tuning combined with visual knowledge distillation effectively enhances emotion-sensitive representation learning and improves sensitivity to discriminative visual features.

The trends shown in Figure 3 further corroborate the quantitative results. VERLADF consistently maintains the highest performance across all datasets, with particularly large improvements in Emotion6 (+16.92%) and FI (+11.60%). These substantial gains indicate that the proposed framework is especially effective in handling complex and diverse emotional distributions. Moreover, the consistent improvements in both abstract and real-world datasets demonstrate that VERLADF not only enhances emotion discrimination but also maintains strong generalization ability across different visual domains. To further evaluate the classification performance from a class-balanced perspective, we additionally report the Macro-F1 and Weighted-F1 scores. On the Emotion6 dataset, VERLADF achieves a Macro-F1 and Weighted-F1 of 73.20%, while on the more challenging FI dataset, it obtains a Macro-F1 of 75.80% and a Weighted-F1 of 78.67%. The relatively high Weighted-F1 on

FI indicates that the proposed method maintains strong overall recognition performance across categories with different sample distributions.

In summary, the experimental results demonstrate that integrating high-level semantic reasoning from VLMs with a lightweight visual distillation module significantly improves visual emotion recognition performance. More importantly, the proposed framework exhibits strong robustness and generalization across datasets with varying semantic complexity, ranging from abstract artistic images to real-world social media scenarios, thereby validating its effectiveness and practical value for visual emotion recognition.

4.4. Feature Representation Analysis

To provide a more intuitive understanding of the learned emotion representations, t-SNE is employed to visualize the feature distributions in a two-dimensional space. t-SNE projects high-dimensional features into a low-dimensional manifold while preserving local neighborhood relationships as much as possible, thereby facilitating intuitive observation of the separability among different emotion categories. Considering that the proposed VERLADF achieves the best recognition performance on the Emotion6 and FI datasets among all evaluated datasets, these two datasets are selected for detailed visualization analysis. The superior performance achieved on these datasets indicates that the learned representations exhibit stronger discriminative capability, making them particularly suitable for analyzing the effectiveness of the proposed fusion framework.

Figures 4 and 5 illustrate the t-SNE visualizations on the Emotion6 and FI test sets, respectively. In each figure, the left subfigure shows the original feature distribution, while the right subfigure presents the feature distribution after applying VERLADF. The results on the Emotion6 dataset are illustrated in Figure 4. It can be observed that, in the original visual representations, samples from different emotion categories are widely dispersed and exhibit significant overlap. For instance, categories such as “fear,” “joy,” and “surprise” are heavily intermingled in the feature space, making it difficult to form clear class boundaries. This indicates that relying solely on original visual features leads to limited discriminative capability. In contrast, after applying the proposed framework, samples from different emotion categories gradually form more compact clusters in the two-dimensional space, with clearer inter-class boundaries. For example, categories such as “disgust” and “anger” form more distinct clusters, and the overlap between categories is significantly reduced, implying that VERLADF effectively enhances the discriminability of emotion features.

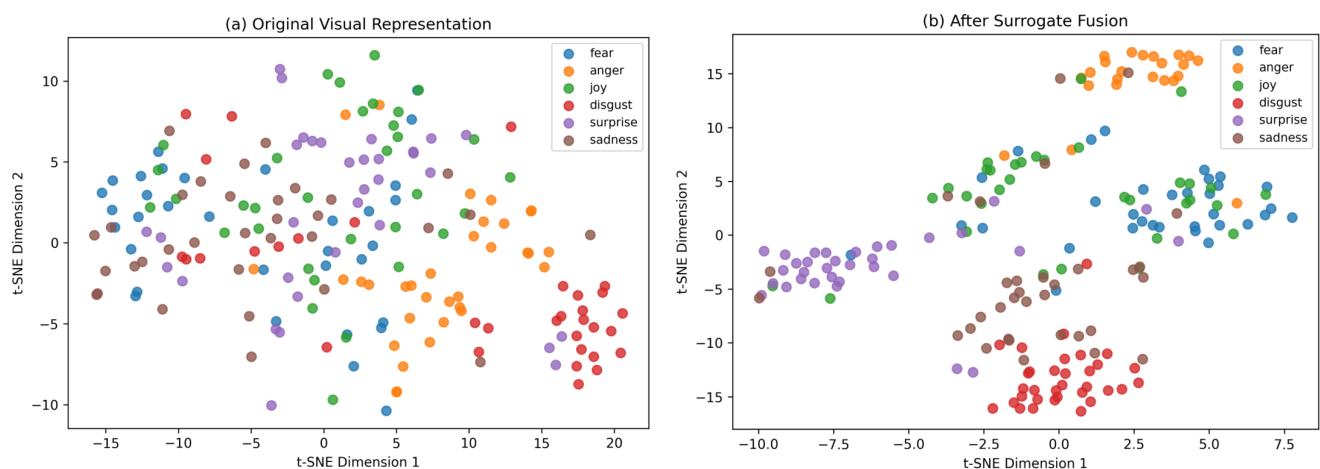


Figure 4. t-SNE visualization on the test set of the Emotion6 dataset.

A similar trend can be observed on the FI dataset, as illustrated in Figure 5. The original visual features exhibit a highly mixed distribution in the two-dimensional space,

with substantial overlap across different emotion categories. In particular, semantically similar categories such as “awe,” “contentment,” and “amusement” are densely distributed without clear separation. After applying the fusion model, samples from different categories form more distinct clusters. For instance, categories such as “excitement,” “fear,” and “amusement” exhibit relatively independent cluster structures, and the inter-class distances increase significantly, indicating that the model learns more discriminative emotion representations.

Overall, the visualization results on both datasets demonstrate that the proposed VERLADF framework effectively refines the distribution of original visual features, making samples within the same category more compact while increasing the separation between different categories. This improvement contributes to enhanced emotion recognition performance, which is consistent with the quantitative results reported in previous experiments and further validates the effectiveness of VERLADF in learning discriminative emotion representations.

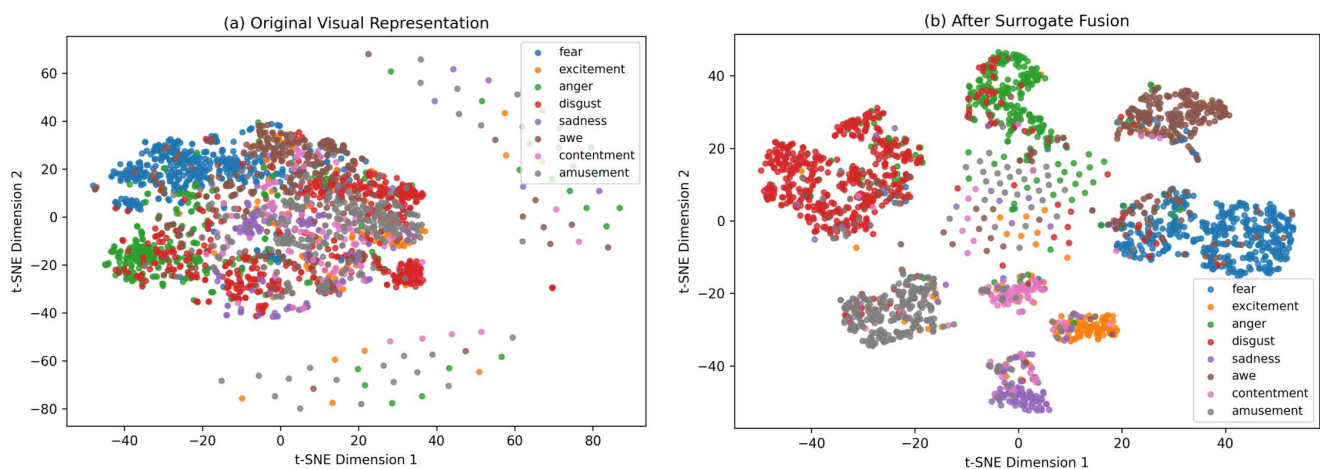


Figure 5. t-SNE visualization on the test set of the FI dataset.

4.5. Confusion Matrix Analysis

Confusion matrices are employed to analyze the category-level recognition performance on the Emotion6 and FI test sets, as shown in Figures 6 and 7. Similar to the t-SNE analysis, these two datasets are selected because VERLADF achieves the highest overall recognition accuracy on them, allowing the class-wise prediction behavior of the proposed framework to be more clearly illustrated. The confusion matrix provides an intuitive representation of the prediction performance across different categories, where the diagonal elements indicate correctly classified samples and the off-diagonal elements represent inter-class misclassifications.

Figure 6 illustrates the confusion matrix obtained on the Emotion6 dataset. It can be observed that most emotion categories exhibit relatively high values along the diagonal, indicating that VERLADF achieves overall strong recognition performance. Among them, the categories “disgust,” “joy,” and “sadness” demonstrate particularly strong performance, with significantly higher numbers of correctly classified samples compared with other categories, suggesting that the model effectively captures their discriminative visual characteristics. In contrast, the “fear” category is relatively more challenging to recognize, as some of its samples are misclassified as “anger” or “surprise.” This phenomenon may be attributed to the visual similarity among these emotions, such as shared features related to tension or surprise, which can overlap across categories and lead to confusion in feature discrimination. In addition, a small number of “surprise” samples are misclassified as “joy” or “fear,” indicating that surprise is often semantically associated with excitement or fear at

the visual level. Such inter-class confusion is common in emotion recognition tasks and reflects the inherent continuity and ambiguity of emotional expressions.

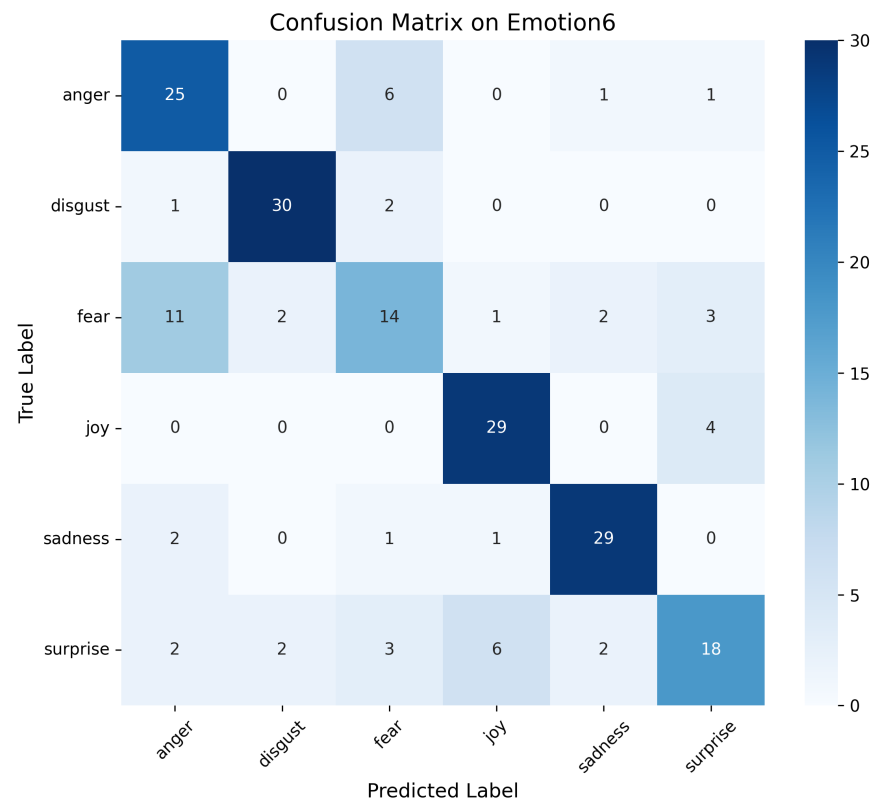


Figure 6. Confusion matrix on the test set of the Emotion6 dataset.

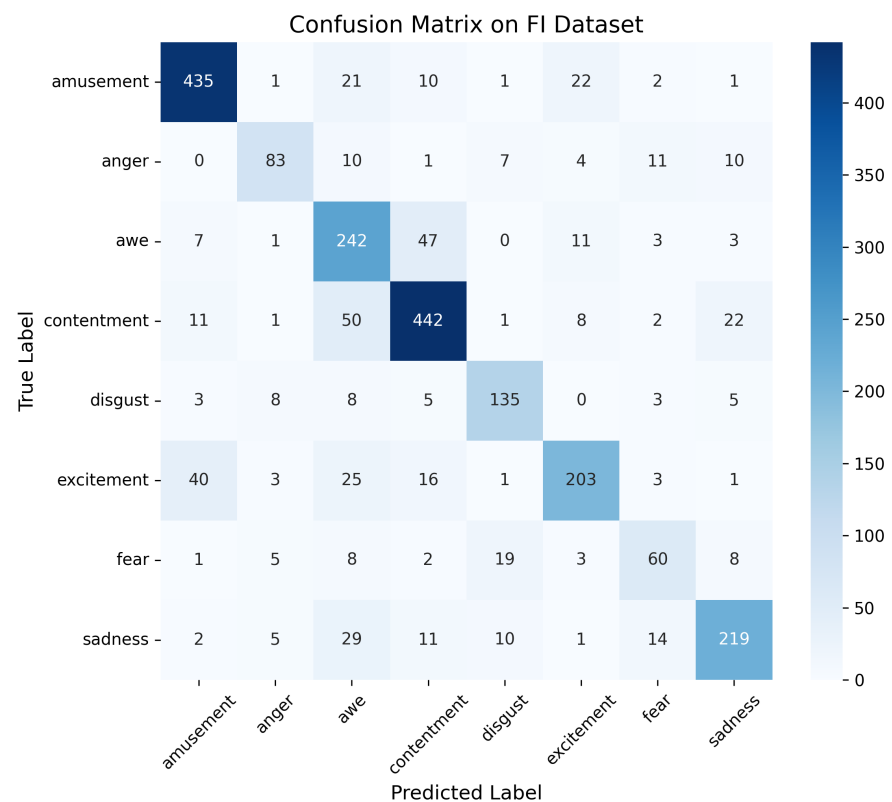


Figure 7. Confusion matrix on the test set of the FI dataset.

The confusion matrix results on the FI dataset are shown in Figure 7. Most samples are concentrated along the diagonal, indicating that the model achieves relatively accurate recognition across the majority of emotion categories. Specifically, categories such as “amusement,” “contentment,” “awe,” and “sadness” exhibit high numbers of correctly classified samples along the diagonal. For example, the “amusement” category includes 435 correctly classified samples, while the “contentment” category contains 442, demonstrating that the model effectively captures their discriminative visual features. Nevertheless, several semantically related categories still exhibit moderate confusion. For instance, some “excitement” samples are misclassified as “amusement,” while a small number of “fear” samples are misclassified as “sadness.” These results suggest that although the proposed framework substantially improves feature discriminability, distinguishing visually and semantically similar emotions remains a challenging problem.

Based on the above observations, the remaining recognition errors are primarily concentrated on emotion categories with subtle semantic boundaries and highly similar visual characteristics. Representative examples include the confusion between fear and anger in Emotion6, as well as the bidirectional confusion between awe and contentment and the confusion between excitement and amusement in FI. These emotion pairs often share similar affective attributes and contextual visual patterns, making their discrimination highly dependent on fine-grained emotional cues rather than global scene semantics. These observations suggest that, although VERLADF significantly improves recognition performance, modeling subtle emotional differences remains an open challenge. Future work may benefit from incorporating fine-grained emotion reasoning, emotion-aware visual representation learning, or richer multimodal semantic interactions to further enhance the discrimination of visually and semantically similar emotion categories.

Overall, the confusion matrix analysis demonstrates that VERLADF effectively learns discriminative representations across different emotion categories and achieves strong recognition performance on multiple datasets. Meanwhile, the remaining confusion among semantically similar categories also highlights potential future directions, such as incorporating richer emotional priors or enhancing multimodal semantic interaction to further improve the discrimination of subtle emotional differences.

4.6. Ablation Study

4.6.1. Contribution Analysis of Different Components

Ablation experiments are conducted on four benchmark datasets, including Abstract, ArtPhoto, Emotion6, and FI, to evaluate the contribution of each component in VERLADF. The quantitative results are reported in Table 5. Specifically, we evaluate five configurations: (1) the instruction-tuned VLM only; (2) a standalone ViT-based emotion recognition model; (3) the proposed lightweight distillation module without fusion with the VLM; (4) the complete VERLADF framework; and (5) VLM + Distillation Module with a simple linear fusion, denoted as VERLADF-Linear.

Table 5. Ablation study results (%). Mean accuracy over three independent runs is reported with standard deviation.

Method	Abstract	ArtPhoto	Emotion6	FI
Instruction-tuned VLM	32.54 ± 1.12	48.81 ± 0.83	63.09 ± 0.52	73.08 ± 0.35
Vision Transformer	33.92 ± 1.25	45.24 ± 0.96	67.17 ± 0.61	68.86 ± 0.42
Distillation Module	35.71 ± 1.08	49.86 ± 0.79	72.04 ± 0.48	77.39 ± 0.32
VERLADF-Linear	35.86 ± 1.05	51.19 ± 0.76	73.28 ± 0.45	78.57 ± 0.30
VERLADF	36.71 ± 1.02	52.38 ± 0.72	74.73 ± 0.41	79.69 ± 0.28

As shown in Table 5, compared with the instruction-tuned VLM, VERLADF improves the accuracy by 4.57%, 3.57%, 4.64%, and 6.61% on Abstract, ArtPhoto, Emotion6, and FI, respectively. These consistent improvements demonstrate the overall effectiveness of the proposed framework in enhancing visual emotion recognition performance. To further evaluate the effectiveness of the proposed fusion strategy, we compare VERLADF with VERLADF-Linear, where the MLP-based proxy fusion is replaced by a simple linear fusion mechanism. The results show that the proposed MLP-based proxy fusion consistently outperforms linear fusion across all datasets, with improvements of 0.85%, 1.19%, 1.45%, and 1.12% on Abstract, ArtPhoto, Emotion6, and FI, respectively. This confirms that the non-linear interaction capability introduced by the MLP architecture provides measurable benefits over linear gating. Finally, the complete VERLADF framework achieves the best performance across all datasets. These results indicate that the proposed fusion and prediction module successfully integrates the complementary strengths of semantic reasoning from the VLM and emotion-discriminative visual perception from the distilled visual branch.

To gain further insight into the source of these performance improvements, we quantitatively analyze the contribution of each component. Taking the FI dataset as an example, the total improvement of 6.61% over the instruction-tuned VLM baseline can be decomposed into two stages. First, the distillation module improves the accuracy from 73.08% to 77.39%, yielding a gain of 4.31%, which demonstrates the effectiveness of transferring emotion-discriminative knowledge from the vision teacher model. Second, the fusion framework further improves the accuracy from 77.39% to 79.69%, contributing an additional 2.30% gain by integrating the complementary strengths of the VLM and the distilled visual branch. Notably, the contribution of the fusion framework is larger than the performance gap between VERLADF-Linear and VERLADF (1.12%). This observation suggests that the overall gain introduced by fusion arises not only from the non-linear fusion architecture itself, but also from the effective exploitation of complementary information provided by semantic reasoning and emotion-discriminative visual perception. Therefore, both the knowledge distillation branch and the proposed fusion strategy play essential and complementary roles in the overall performance improvement in VERLADF.

4.6.2. Effect of Prompt Template

To evaluate the effectiveness of the introduced prompt template, we compare the original instruction construction strategy adopted in EmoViT with our enhanced template incorporating Key Elements and Reasoning. Specifically, the EmoViT setting uses the original emotion instruction design, while our method further guides the model to identify emotion-related visual cues and generate structured reasoning for emotion inference, as shown in Figure 8. The proposed template consistently improves performance across all datasets, with particularly notable gains on Emotion6 and FI. As shown in Table 6, these results indicate that explicitly encouraging the model to focus on emotion-relevant visual evidence and reasoning processes enhances the emotion understanding capability of the VLM.

Table 6. Ablation study on prompt template.

Instruction Configuration	Abstract	ArtPhoto	Emotion6	FI
EmoViT	32.34	44.90	57.81	68.09
VLM + Key Elements + Reasoning (Ours)	32.54	48.81	63.09	73.08

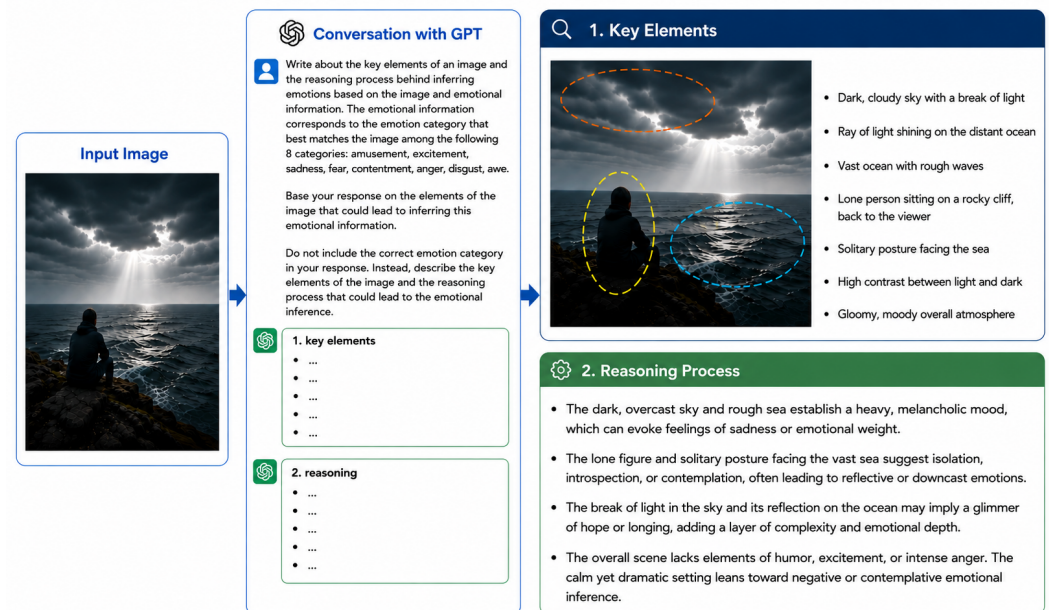


Figure 8. Detailed illustration of the prompt templates comprising “Key Elements” and “Reasoning”.

4.6.3. Lightweight Distillation Evaluation

The inference latency was measured on an NVIDIA RTX 3090 GPU with a batch size of 1 under the same input resolution and prompt setting, excluding data loading and initialization overhead. The reported parameter number corresponds to the parameters involved in each configuration. For the proposed framework, the Qwen2-VL-7B parameters are frozen during training, and only the lightweight distillation module and fusion module are optimized. As shown in Table 7, the proposed distillation module is extremely lightweight in terms of trainable parameters, containing only 3.68 M parameters. Compared with jointly deploying Qwen2-VL-7B and ViT with full fine-tuning (which requires an additional 86M ViT parameters beyond Qwen2-VL-7B itself), VERLADF only needs to optimize the newly added marginal parameters during training, significantly reducing the optimization difficulty. In terms of inference efficiency, VERLADF avoids the extra computational overhead caused by independent parallel inference of ViT and VLM. Owing to the knowledge distillation strategy, our framework does not require deploying the ViT teacher network at inference time. Instead, it reuses the shared features extracted by the VLM visual encoder and performs lightweight mapping through the distillation module. Consequently, compared with the parallel inference scheme (latency 754 ms), the inference latency of VERLADF is almost on par with that of the single Qwen2-VL-7B model, demonstrating its efficiency in deployment.

Table 7. Comparison of distillation module in inference latency and trainable parameters.

Model Configuration	Trainable Parameters	Latency
Qwen2-VL-7B (baseline)	8296.42 M	731 ms
ViT + Qwen2-VL-7B	8382.52 M	754 ms
Distillation module	3.68 M	252 ms
Distillation module + Qwen2-VL-7B (frozen)	3.68 M	733 ms

4.6.4. Fusion Mechanism Comparison

To further justify the choice of the MLP-based proxy fusion strategy, we compare it with three alternative fusion mechanisms commonly adopted in multimodal learning on the Emotion6 and FI datasets. The compared mechanisms include: (1) Linear, which concate-

rates the VLM prediction and distillation output and passes them through a single linear layer [40]; (2) Bilinear Pooling [41], which applies a bilinear operation $v_1^T W v_2$ followed by a linear transformation; and (3) Cross-Gating [42], which uses sigmoid-gated cross-modal signals to dynamically regulate information flow from each module. All variants share the same pre-trained VLM and distillation module, differing only in the fusion architecture.

As shown in Table 8, the proposed MLP-based proxy fusion consistently outperforms all alternative fusion mechanisms on both datasets. On Emotion6, the MLP-based fusion surpasses the alternative linear fusion by 1.45%, while on FI, the improvement is 1.12%. On Abstract and ArtPhoto, the improvements are 0.85% and 1.19%, respectively. These results suggest that the MLP-based proxy fusion enables more effective interactions between the semantic decision cue provided by the VLM and the emotion-discriminative predictions generated by the distillation module. Consequently, the fusion module can better exploit the complementary strengths of semantic reasoning and visual emotion perception, leading to improved recognition performance.

Table 8. Comparison of different fusion mechanisms (%).

Fusion Mechanism	Abstract	ArtPhoto	Emotion6	FI
Linear [40]	35.86	51.19	73.28	78.57
Bilinear Pooling [41]	35.14	50.47	72.61	77.89
Cross-Gating [42]	35.42	50.80	72.84	78.10
MLP-based Proxy Fusion (Ours)	36.71	52.38	74.73	79.69

4.6.5. Statistical Significance Analysis.

We further evaluate the statistical significance of VERLADF over EmoViT using the continuity-corrected McNemar's test on the 198 Emotion6 test samples and the 2326 FI test samples. As shown in Table 9, b denotes the number of samples correctly classified by VERLADF but misclassified by EmoViT, whereas c denotes the number of samples correctly classified by EmoViT but misclassified by VERLADF. On Emotion6, VERLADF correctly classifies 46 samples misclassified by EmoViT, whereas EmoViT correctly classifies only 12 samples misclassified by VERLADF. Similarly, on FI, the corresponding numbers are 502 and 248, respectively. The resulting p -values are both below the significance threshold of 0.05. These results provide strong evidence that the observed performance improvements in VERLADF are statistically significant and consistently achieved across different test samples, further demonstrating the effectiveness and reliability of the proposed framework.

Table 9. Statistical significance test results between VERLADF and EmoViT on two benchmark datasets. The continuity-corrected McNemar's χ^2 test ($df = 1$) was conducted using paired prediction outcomes of VERLADF and EmoViT on the same test images.

Method	Emotion6 (198)					FI (2326)				
	b	c	p	b	c	p	b	c	p	
EmoViT	×	46	✓	12	$p < 10^{-4}$	×	502	✓	248	$p < 1.5 \times 10^{-15}$
VERLADF	✓	×	×	×	×	×	✓	×	×	×

4.6.6. Hyperparameter Sensitivity Analysis

We further investigate the sensitivity of VERLADF to two key hyperparameters: the loss balancing weight α and the distillation temperature τ . During the sensitivity analysis, one hyperparameter is changed, while all other hyperparameters are fixed to their default values.

Loss-balancing weight α : The hyperparameter α controls the relative importance of the KL divergence loss L_{KD} and the cross-entropy loss L_{CE} . We evaluate $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ on the FI and Emotion6 datasets. The results are reported in Figure 9. The proposed method consistently achieves the best performance at $\alpha = 0.5$ on both the FI and Emotion6 datasets. When α is either too small or too large, the performance gradually declines. Nevertheless, the relatively small performance variations across different α values demonstrate that the proposed framework is not overly sensitive to the choice of this hyperparameter.

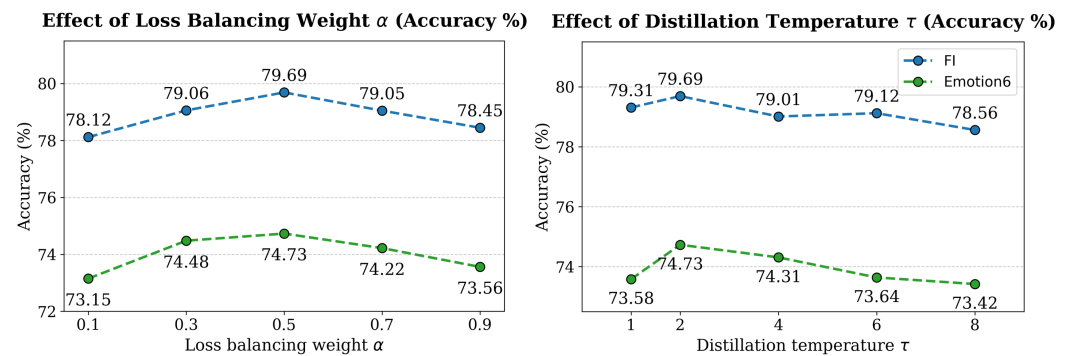


Figure 9. Sensitivity analysis of the loss balancing weight α and the distillation temperature τ on the FI and Emotion6 datasets.

Distillation temperature τ : the best performance is obtained at $\tau = 2$ on both datasets. As τ deviates from this value, the performance decreases slightly, suggesting that a moderate temperature provides the most informative soft targets for knowledge distillation. Overall, the stable performance over a wide range of α and τ values demonstrates the robustness of the proposed framework to hyperparameter selection.

5. Conclusions

This paper proposes a visual emotion recognition framework based on VLM distillation and fusion, called VERLADF, aiming to fully leverage the advantages of VLMs in high-level semantic modeling while maintaining low computational overhead for practical deployment. By combining an emotion instruction-tuned VLM with a lightweight visual distillation module, the proposed approach achieves effective complementarity between high-level semantic understanding and complex visual emotional perception. Specifically, the VLM is first fine-tuned with emotion instructions to enhance its focus on emotion-related visual semantic information. Subsequently, a ViT teacher model is introduced to transfer its learned discriminative visual emotion knowledge to a distillation module, thereby strengthening the model's ability to capture visual cues. Finally, the predictions of the distillation module are adaptively fused with the outputs of the frozen VLM to achieve stable and robust emotion predictions. Extensive experimental results on four benchmark datasets, namely Abstract, ArtPhoto, Emotion6, and FI, demonstrate that the proposed method significantly outperforms existing strong baseline models across datasets with diverse visual styles and semantic complexity, validating the effectiveness and generalization capability of the framework for the visual emotion recognition task. Future work will further explore the extensibility of this method to more complex multimodal emotion understanding tasks, and investigate more efficient mechanisms for large-model compression and lightweight distillation strategies to support large-scale practical deployment in resource-constrained environments.

Author Contributions: Conceptualization, Y.Z., Z.C. and Z.Y.; methodology, Y.M., Y.Z. and Z.C.; software, Y.M., Y.Z. and Z.C.; validation, Y.M., Y.Z. and Z.C.; formal analysis, Y.M., Y.Z. and Z.C.; investigation, Z.Y., W.Z. and C.X.; resources, Y.M., Z.Y., Z.C. and Z.Y.; data curation, Y.M., Y.Z. and Z.C.; writing—original draft preparation, Y.Z. and Z.C.; writing—review and editing, Y.M., Y.Z. and Z.Y.; visualization, Y.Z. and Z.C.; supervision, Y.M., Y.Z. and Z.C.; project administration, Y.M. and Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant Nos. U23A20318, 62376089, 62302153, 62302154) and the Young and Middle-aged Scientific and Technological Innovation Team Plan in Higher Education Institutions in Hubei Province, China (Grant No. T2023007).

Data Availability Statement: The data in this study are available at <http://anonymous.4open.science/r/VERLADF-4E57> (accessed on 6 August 2026).

Conflicts of Interest: The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Hadjar, H.; Vu, B.; Hemmje, M. TheraSense: Deep learning for facial emotion analysis in mental health teleconsultation. *Electronics* **2025**, *14*, 422. [CrossRef]
2. Moosavi, J.; Resch, A.; Lecchi, A.; Sokolov, A.N.; Fallgatter, A.J.; Pavlova, M.A. Reading language of the eyes in female depression. *Cereb. Cortex* **2024**, *34*, bhae253. [CrossRef]
3. Guo, R.; Guo, H.; Wang, L.; Chen, M.; Yang, D.; Li, B. Development and application of emotion recognition technology—A systematic literature review. *BMC Psychol.* **2024**, *12*, 95. [PubMed]
4. Zhou, Z.; Zhai, Z.; Chen, H.; Lu, S. Object-scene semantics correlation analysis for image emotion classification. *Front. Neurosci.* **2025**, *19*, 1657562. [CrossRef] [PubMed]
5. Wu, D.; Yang, D.; Zhou, Y.; Ma, C. Bridging Visual Affective Gap: Borrowing Textual Knowledge by Learning from Noisy Image-Text Pairs. In *Proceedings of the 32nd ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2024; pp. 602–611.
6. Huang, X.; Hong, X.; Mao, Q.; Zheng, W.; Dhall, A. A Survey on Deep Learning for Group-Level Emotion Recognition. *IEEE Trans. Comput. Soc. Syst.* **2025**, *13*, 2475–2500.
7. Lan, J.; Jiang, X.; Lin, G.; Zhou, X.; You, S.; Liao, Z.; Fan, Y. Expression recognition based on multi-regional coordinate attention residuals. *IEEE Access* **2023**, *11*, 63863–63873. [CrossRef]
8. Al-Shamasneh, A.R.; Karim, F.K.; Wang, Y. A Robust Transformer-Residual Hybrid Framework with Soft Thresholding for High-Performance Image Emotion Classification. *Int. J. Comput. Intell. Syst.* **2025**, *18*, 244. [CrossRef]
9. Kus, I.; Kocak, C.; Keles, A. A systematic review of vision transformer and explainable AI advances in multimodal facial expression recognition. *Intell. Syst. Appl.* **2025**, *29*, 200615.
10. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23716–23736. [CrossRef]
11. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2023; pp. 19730–19742.
12. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 34892–34916. [CrossRef]
13. Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.N.; Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 49250–49267. [CrossRef]
14. Xie, H.; Peng, C.J.; Tseng, Y.W.; Chen, H.J.; Hsu, C.F.; Shuai, H.H.; Cheng, W.H. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 26596–26605.
15. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
16. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv* **2024**, arXiv:2409.12191.
17. Huang, D.; Li, Q.; Yan, C.; Cheng, Z.; Huang, Y.; Li, X.; Li, B.; Wang, X.; Lian, Z.; Peng, X. Emotion-Qwen: Training Hybrid Experts for Unified Emotion and General Vision-Language Understanding. *arXiv* **2025**, arXiv:2505.06685.

18. Guo, Y.; Siddiqui, F.; Zhao, Y.; Chellappa, R.; Lo, S.Y. Stimuvar: Spatiotemporal stimuli-aware video affective reasoning with multimodal large language models. *Int. J. Comput. Vis.* **2025**, *133*, 6456–6472. [[CrossRef](#)]
19. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*; PmLR: Cambridge, MA, USA, 2021; pp. 8748–8763.
20. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2022; pp. 12888–12900.
21. Lee, S.; Lee, Y.; Park, E.; Chae, W. MASP: Multi-Aspect Guided Emotion Reasoning with Soft Prompt Tuning In Vision-Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*; PKP: Burnaby, BC, Canada, 2026; Volume 40, pp. 1882–1890.
22. Lee, Y.; Park, S.; Park, J.; Park, E. AIMER: Affective Intention-guided Multimodal Emotion Reasoner for Visual Emotion Analysis in Social Media. In *Proceedings of the ACM Web Conference 2026*; ACM: New York, NY, USA, 2026; pp. 8361–8364.
23. Huang, J.; Lin, F.; Yang, X.; Feng, C.; Zhu, K.; Yang, X.; Chen, Z. Nano-EmoX: Unifying Multimodal Emotional Intelligence from Perception to Empathy. *arXiv* **2026**, arXiv:2603.02123.
24. Yang, Q.; Ye, M.; Tao, D. Synergy of Sight and Semantics: Holistic Visual Understanding with CLIP. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *48*, 3119–3136. [[CrossRef](#)] [[PubMed](#)]
25. Chakraborty, T.; Caplan, E.; Goldwasser, D. VIBE: Can a VLM Read the Room? *arXiv* **2025**, arXiv:2506.11162.
26. Agarwal, M.; Tsaftaris, S.A.; Sevilla-Lara, L.; McDonagh, S. Why Do Vision Language Models Struggle To Recognize Human Emotions? *arXiv* **2026**, arXiv:2604.15280.
27. Shinde, G.; Ravi, A.; Dey, E.; Sakib, S.; Rampure, M.; Roy, N. A Survey on Efficient Vision-Language Models. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2025**, *15*, e70036. [[CrossRef](#)]
28. Ahmed, S.I.; Hasan, M.Z.; Niloy, A.J.; Monsur, M.S.; Albert, M.V. Scaling Down, Powering Up: A Survey on the Advancements of Small Vision-Language Models. *Inf. Fusion* **2025**, *127*, 103805. [[CrossRef](#)]
29. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
30. Win, K.C.; Akhtar, Z.; Mohan, C.K. CE-KD: Class-Wise Expert-Based Knowledge Distillation for Facial Expression Recognition. In *Proceedings of the 2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*; IEEE: New York, NY, USA, 2025; pp. 1–6.
31. Wang, N.; Wang, Q. Modality-uncertainty-aware knowledge distillation framework for multimodal sentiment analysis. *Complex Intell. Syst.* **2026**, *12*, 14. [[CrossRef](#)]
32. Wu, S.; Liang, H.; Zhang, Y.; Chen, Y.; Jia, Z. A cross-modal densely guided knowledge distillation based on modality rebalancing strategy for enhanced unimodal emotion recognition. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, QC, Canada, 16–22 August 2025*; pp. 4236–4244.
33. Li, Y.; Wang, Y.; Ding, Y.; Zhang, S.; Lu, K.; Guan, C. Decoupled Hierarchical Distillation for Multimodal Emotion Recognition. *arXiv* **2026**, arXiv:2602.04260.
34. Mao, A.; Mohri, M.; Zhong, Y. Cross-entropy loss functions: Theoretical analysis and applications. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2023; pp. 23803–23828.
35. Machajdik, J.; Hanbury, A. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*; ACM: New York, NY, USA, 2010; pp. 83–92.
36. Peng, K.C.; Chen, T.; Sadovnik, A.; Gallagher, A.C. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2015; pp. 860–868.
37. You, Q.; Luo, J.; Jin, H.; Yang, J. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*; ACM: New York, NY, USA, 2016; Volume 30.
38. Mikels, J.A.; Fredrickson, B.L.; Larkin, G.R.; Lindberg, C.M.; Maglio, S.J.; Reuter-Lorenz, P.A. Emotional category data on images from the international affective picture system. *Behav. Res. Methods* **2005**, *37*, 626–630. [[CrossRef](#)] [[PubMed](#)]
39. Fang, Y.; Liang, J.; Huang, W.; Li, H.; Su, K.; Ye, M. Catch Your Emotion: Sharpening Emotion Perception in Multimodal Large Language Models. In *Proceedings of the International Conference on Machine Learning*; ACM: New York, NY, USA, 2025.
40. Li, S.; Tang, H. Multimodal alignment and fusion: A survey. *arXiv* **2024**, arXiv:2411.17040.
41. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; ACL: Kerrville, TX, USA, 2016; pp. 457–468.
42. Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; González, F.A. Gated multimodal units for information fusion. In *Proceedings of the International Conference on Learning Representations Workshop, Toulon, France, 24–26 April 2017*.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.