Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Action Units to Emotions: A Two-Stage Fine-Tuning Framework with Decision-Level Fusion for Facial Expression Recognition

Rui Tu
Technical University of Munich
Liguo Zhou
liguo.zhou@tum.de
Technical University of Munich
Alois Knoll
Technical University of Munich

Research Article

Keywords:

Posted Date: August 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10578386/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

From Action Units to Emotions: A Two-Stage Fine-Tuning Framework with Decision-Level Fusion for Facial Expression Recognition

Author Rui Tu(rui.tu@tum.de)

Dr. Liguo Zhou(liguo.zhou@tum.de)

Prof. Alois Christian Knoll(knoll@in.tum.de)

Date July 23, 2026 in Munich

Abstract

Facial expressions serve as a primary non-verbal channel through which humans convey emotional states and behavioral intentions, holding significant value across diverse application scenarios including digital mental health, human-computer interaction, educational engagement analysis, and consumer behavior understanding. However, facial expression recognition (FER) technology remains constrained by practical challenges such as illumination variations, pose differences, and scarcity of annotated data, with existing models generally suffering from poor cross-domain adaptability, overfitting, and insufficient interpretability.

To address these issues, this paper proposes a multi-modal large-model enhancement strategy based on two-stage fine-tuning. First, a question-answer paired facial action unit (AU) fine-tuning dataset is constructed from the CAS(ME)³ dataset, and the Qwen3-VL-32B-Instruct multi-modal large model is fine-tuned in the first stage using LoRA to reinforce its AU recognition capability. Subsequently, a second-stage fine-tuning is conducted on the FER dataset to guide the model in achieving emotion classification through AU-based reasoning. On this basis, decision-level joint prediction is further explored by combining the fine-tuned large model with the specialized small model OpenFace 3.0.

Systematic experimental evaluations demonstrate that the proposed two-stage fine-tuned fusion model achieves significant superiority over both the non-fine-tuned Qwen3-VL and other mainstream deep models, and also substantially outperforms the standalone inference results of OpenFace 3.0, fully validating the effectiveness of the two-stage fine-tuning strategy in eliciting discriminative features for subtle facial geometric deformations and low-intensity expressions from large visual models. Meanwhile, the study also reveals that while pure large visual models possess formidable capabilities in high-level semantic understanding, they still exhibit inherent limitations in perceiving subtle facial deformations. The proposed method demonstrates clear efficacy in both improving prediction accuracy and enhancing interpretability, offering a novel technical pathway and empirical foundation for constructing highly robust and interpretable FER systems.

Contents

1 Introduction **1**

2 related works **3**
2.1 Traditional Facial Expression Recognition Methods 3
2.2 Multimodal Large Language Models for Emotion Recognition 3
2.3 AU-Based Fine-Tuning and Inference Strategies 3
2.4 Collaborative Decision-Making 4

3 The proposed Method **5**
3.1 Two-stage Progressive Fine-tuning Framework 5
3.2 Task Fomulation 5
3.3 Data Preprocess 7
3.3.1 Static Image Dataset Processing (FER-2013) 7
3.3.2 Frame-Sequence Dataset Processing (CAS(ME)3) 7
3.3.3 Dataset Splitting 8
3.4 Qwen3-VL Introduction 8
3.5 AU-Detection Fine-Tuning 8
3.5.1 Fine-Tuning Objective 9
3.5.2 LoRA-Based Parameter Efficient Fine-tuning 9
3.5.3 AU-Oriented Instruction Tuning 10
3.5.4 Training Objective 12
3.6 Expression Recognition Fine-tuning 13
3.6.1 Fine-Tuning Objective 13
3.6.2 Fine-Tuning Strategy 13
3.7 Decision-Level Fusion with OpenFace 3.0 14

4 Experiments and Results **17**
4.1 Datasets 17
4.2 Evaluation Metrics 17
4.3 Model Performance 18
4.3.1 Experimental Setup 18
4.3.2 Overall Model Performance 18
4.3.3 Advantages of the Fused Model over the Pure Large Model 20
4.3.4 Effectiveness of Fine-Tuning 20
4.3.5 Advantages of the Fused Model over Pure OpenFace 3.0 21
4.3.6 Comparative Experimental Analysis 21
4.3.7 Confusion Matrix Analysis 22

5 Conclusions and Future Work **25**
5.1 Conclusions 25
5.2 Future Work 25

Bibliography **27**

Chapter 1

Introduction

Facial expressions directly reflect an individual's internal emotions and behavioral intentions. It plays a central role in interpersonal communication. Facial expression recognition (FER) technology has now been widely applied across various domains, including digital mental health, human-computer interaction, educational engagement analysis, and consumer behavior understanding. Automatic facial expression analysis is one of the core research topics in the field of affective computing. Researchers have explored a variety of facial expression recognition methods, with the core idea being to encode emotion-related information from facial visual representations and subsequently decode it for emotion detection and classification. While significant progress has been made through the expansion of datasets and the development of advanced neural architectures, existing Facial Expression Recognition (FER) methods still struggle with robustness and generalization.

Current 2D FER systems face persistent challenges, including illumination variability, pose misalignment, and severe individual differences, which demand models to possess invariance, robustness, and identity-disentanglement capabilities. To mitigate these issues, researchers have explored 3D/4D data fusion for geometric capture, as well as attention mechanisms and Generative Adversarial Networks (GANs) for occlusion handling. However, these approaches are often constrained by their heavy reliance on expensively annotated large-scale data and their failure to fully disentangle identity from expression.

Compounding these technical hurdles is the core bottleneck of data scarcity, resulting from the subjective nature of expression annotation, the requirement for professional expertise, and the rising costs and privacy regulations associated with data acquisition. Early handcrafted methods, such as LBP-TOP [ZP07] and Non-negative Matrix Factorization [Zhi+11], achieved initial success but proved insufficient for capturing the subtle and transient nature of genuine expressions. Subsequently, deep learning paradigms—including CNNs [Zha+21], GNNs [Xie+20], and Transformers [Ind+23]—have pushed performance boundaries by treating facial images as structured data. Nevertheless, while effective in single-domain tasks, these models exhibit limited cross-domain semantic alignment and interactive capability.

The emergence of Multimodal Large Language Models (MLLMs), capable of jointly processing visual and textual modalities, offers a new paradigm for FER. Fine-tuning MLLMs has recently shown promise in enhancing recognition performance. In parallel, the Facial Action Coding System [EF78] and its Action Units (AUs) provide a fine-grained, anatomically grounded taxonomy for describing facial muscle movements (see Figure 1.1). Given their ability to bridge low-level muscle actions and high-level emotions, AUs have been extensively employed as effective discriminative features in FER [ZLZ20].

Despite these advances, end-to-end fine-tuning of MLLMs relying solely on "video-label" supervisory signals remains suboptimal, as it neglects the fine-grained intermediate representations (i.e., AUs) that underlie emotional expressions.

To address this gap, we propose a two-stage fine-tuning strategy using the AU annota-

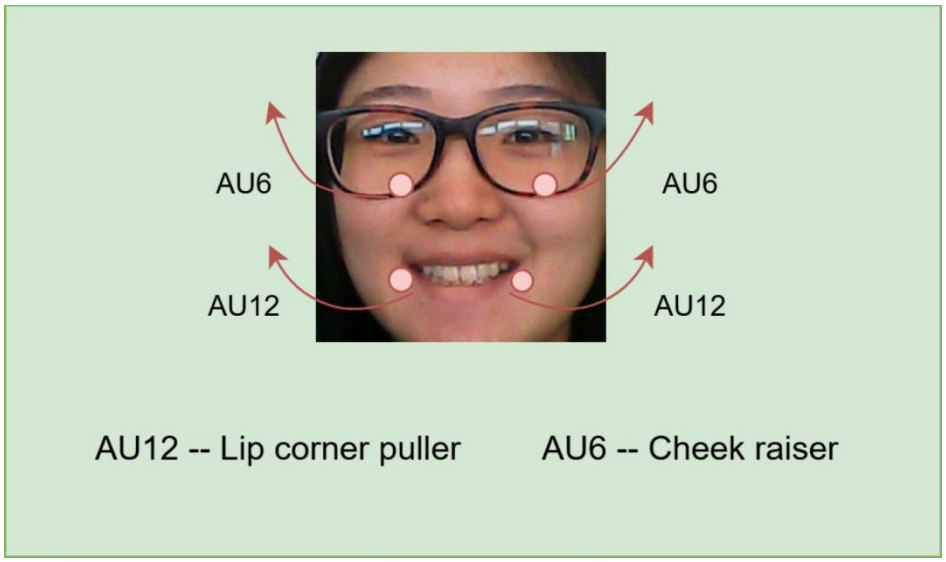


Figure 1.1: An illustrative example of the association between Action Unit (AU) expressions and emotions

tions from the CAS(ME)3 dataset to enhance the Qwen3-VL-32B-Instruct model. Our approach first guides the model to infer AUs before predicting emotions. We systematically evaluate our method on the FER-2013 benchmark, where our fine-tuned model significantly outperforms existing vision-language baselines (e.g., Qwen2.5-VL and QVQ [Qwe24]) across multiple metrics, including macro/micro expression recognition accuracy and recall. Furthermore, we integrate our fine-tuned MLLM with OpenFace, a lightweight CNN-based FER model, via decision-level fusion. The results show that our model substantially boosts OpenFace's performance, highlighting its practical collaborative value.

Overall, our main contributions are four-fold:

- **Validation of Large Model Baseline Performance:** We systematically evaluated the zero-shot and few-shot performance of Qwen3-VL-32B-Instruct on facial emotion recognition tasks, establishing a strong baseline for this model on this task.
- **Large Model Fine-Tuning Integrated with AU Annotations:** We fine-tuned Qwen3-VL-32B-Instruct using the AU annotations from the CAS(ME)3 dataset, enabling the model to acquire AU inference capabilities. The model was then trained and evaluated on the original train-test split of FER-2013, following a two-stage inference pipeline of "AU inference first, emotion prediction second," fully leveraging its visual perception and reasoning abilities.
- **Decision-Level Fusion of Large and Traditional Models:** We performed decision-level fusion between the fine-tuned Qwen3-VL and OpenFace 3.0, a lightweight convolutional neural network, to explore the feasibility of complementary integration and validate the practical effectiveness of heterogeneous model collaboration.
- **Robust Empirical Performance:** The Qwen3-VL-32B-Instruct model fine-tuned on CAS(ME)3 achieved superior results over existing vision-language baselines on the FER-2013 test set. Moreover, when integrated with traditional deep learning models, its recognition accuracy significantly surpassed that of any single neural network model. These results indicate that large models fine-tuned with AU annotations can substantially improve the detection accuracy of facial affective states while exhibiting strong cross-dataset transferability.

Chapter 2

related works

Facial Expression Recognition (FER) has evolved from handcrafted feature-based pipelines to deep learning and, more recently, multimodal large language models (MLLMs).

2.1 Traditional Facial Expression Recognition Methods

Early FER relied on handcrafted descriptors such as LBP-TOP [ZP07], HOG 3D [PKO09], and optical flow features (e.g., MDMO [LLL18], HOOF [LYZ16]). These methods are computationally efficient and label-lean, but are vulnerable to illumination and pose variations, require extensive domain knowledge, and lack precision for subtle micro-expressions. Deep learning approaches—CNNs, RNNs/LSTMs, C3D, and ensembles—automatically learn hierarchical features and achieve strong performance on large-scale datasets [He+16] [Don+15] [Zha+17]. Nonetheless, they demand large annotated corpora, suffer from poor interpretability, and exhibit limited cross-domain generalization.

2.2 Multimodal Large Language Models for Emotion Recognition

Vision-language models like CLIP enable zero-shot FER via embedding similarity [Zho+24] [Ryu+25], yet operate primarily at the level of feature matching without explicit reasoning. Instruction-tuned MLLMs (e.g., AffectGPT [Lia+25], EmoLA [Li+24], AU-LLaVA [Hu+24]) treat emotion recognition as a generative task, but most still follow an end-to-end "image-to-emotion" mapping, lacking an explicit intermediate reasoning chain.

2.3 AU-Based Fine-Tuning and Inference Strategies

The Facial Action Coding System (FACS) provides an objective link between facial muscle movements and emotions. Several works incorporate AUs as auxiliary cues [Wan+21] [Xie+20] [Sun+20] [Wan+23]. ExpLLM [Wan+26] and AULLM++ [Qwe24] introduce AU-grounded Chain-of-Thought reasoning and achieve state-of-the-art results. However, AU annotations are costly and scarce, and AUs are typically treated as optional supplements rather than mandatory reasoning prerequisites. Existing MLLM-based FER methods have not systematically integrated AU detection as a structured, obligatory intermediate step within the fine-tuning pipeline.

2.4 Collaborative Decision-Making

Recent work in general domains combines large and small models via cascading, routing, or decision fusion to balance performance and efficiency [HGJ+25]. However, such collaboration remains largely unexplored in FER. Lightweight experts like OpenFace 3.0 [Hu+25] capture local motion cues effectively but lack semantic reasoning capacity, while MLLMs offer strong visual understanding but are relatively insensitive to fine-grained micro-movements. Systematic frameworks that leverage the complementary strengths of both paradigms for FER are currently absent.

Chapter 3

The proposed Method

3.1 Two-stage Progressive Fine-tuning Framework

The two-stage progressive fine-tuning strategy adopted in this paper essentially decomposes the complex cognitive task of facial emotion recognition into two sequential sub-problems: AU perception ("what facial changes are observed") and emotion inference ("what emotions these changes signify"). In the first stage, fine-grained AU annotations from CAS(ME)³ video frames are leveraged to train the model to establish robust, domain-agnostic perceptual capabilities for subtle facial micro-movements. In the second stage, built upon this foundation, the model learns to map AU combinations to emotion categories on the target-domain FER data. This design achieves a triple decoupling: the decoupling of perception from reasoning, the decoupling of domain-agnostic knowledge (AUs) from domain-specific knowledge (emotions), and the decoupling of physical observations (muscle movements) from emotion labels. Our algorithm is shown in Algorithm 1. Experimental results demonstrate that this curriculum learning paradigm effectively mitigates the domain discrepancy between video frames and static images, while significantly improving the accuracy and interpretability of cross-domain emotion recognition.

3.2 Task Fomulation

This subsection focuses on the task of Facial Expression Recognition (FER), aiming to endow multimodal large language models with fine-grained perceptual and semantic interpretation capabilities for emotional states in facial images. Unlike conventional FER tasks that merely require models to output discrete emotion category labels, this task further demands that models explicitly reveal the intermediate facial representations—namely, Action Units (AUs)—that underpin their decisions, thereby enabling an interpretable reasoning chain from micro-level facial manifestations to high-level emotional semantics. The core definition of the task is as follows:

Inputs: A static facial image $I \in \mathrm{R}^{H \times W \times 3}$, where H and W denote the height and width of the image, respectively. The image is captured from facial expression scenarios under arbitrary lighting conditions, head poses, and occlusions.

Outputs: A natural language response $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}$, where y_j denotes the j-th generated token. The response is required to simultaneously encompass the following three levels of information:

Intermediate-Level Facial Encoding: Detect and output the set $A = \{\mathrm{AU}_1, \mathrm{AU}_2, \cdots, \mathrm{AU}_K\}$ of active facial action units present in the input image, where K denotes the number of detected action units. These AUs constitute an interpretable physiological bridge linking

Algorithm 1 Two-Stage Progressive Fine-tuning with Decision-Level Fusion for Facial Emotion Recognition

Require Pre-trained VLM: Qwen3-VL-32B-Instruct; OpenFace 3.0;
Datasets: $D_{\text{CAS(ME)}^3}$ (AU labels),
$D_{\text{FER2013-train}}$ (emotion labels);
Test set: $D_{\text{FER2013-test}}$;
Fusion weight: $\alpha \in [0, 1]$
Ensure Final predicted emotion labels for $D_{\text{FER2013-test}}$
Stage 1: AU Recognition Fine-tuning (VLM Branch)
Load pre-trained Qwen-VL as M_{base}
Fine-tune M_{base} on D_{CASME3} with AU detection objective
Obtain M_{AU} with capability to predict Action Unit activations
Stage 2: Emotion Fine-tuning on FER2013 (VLM Branch)
Initialize M_{em} with weights from M_{AU}
Fine-tune M_{em} on $D_{\text{FER2013-train}}$ with emotion classification objective
Obtain VLM-based emotion predictor M_{VLM}
Stage 3: Decision-Level Fusion with OpenFace 3.0
for each test sample **x** in $D_{\text{FER2013-test}}$ **do**
Extract facial image from **x**
Obtain VLM logits: $\mathbf{z}_{\text{VLM}} = M_{\text{VLM}}(\text{image})$
Compute VLM probability: $P_{\text{VLM}} = \text{Softmax}(\mathbf{z}_{\text{VLM}})$
Obtain OpenFace 3.0 logits: $\mathbf{z}_{\text{OF}} = \text{OpenFace3.0}(\text{image})$
Compute OpenFace probability: $P_{\text{OF}} = \text{Softmax}(\mathbf{z}_{\text{OF}})$
Fuse predictions: $P_{\text{final}} = \alpha \cdot P_{\text{VLM}} + (1 - \alpha) \cdot P_{\text{OF}}$
Predict emotion: $\hat{y} = \arg\max_{c \in \{1,\ldots,8\}} P_{\text{final}}(c)$
end for
Evaluation
Compute accuracy and other metrics on $D_{\text{FER2013-test}}$
return Evaluation results

facial deformations to emotional states, laying the foundation for the model's subsequent emotion output.

Final Prediction: Based on the AU representations, the model is required to reason about and output the emotion category label $C \in \mathrm{C}$ expressed in the image, where $C = \{$Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral$\}$ is the seven-class basic emotion space.

Semantic Interpretation: In addition to detecting the set of facial action units in the input image and generating the final recognized facial emotion, the model is also required to produce a natural language description that articulates the logical association between the detected AU combinations and the final emotion judgment. This textual elaboration enhances the interpretability and persuasiveness of the model's output, facilitating user comprehension of the rationale behind the emotion judgment—for example, "detected AU4 (brow lowerer) and AU7 (lid tightener), suggesting that this facial expression is likely Anger."

To address the aforementioned task, this paper constructs a two-stage progressive fine-tuning framework based on the open-source multimodal large model Qwen3-VL, which decomposes the target task into two sequential learning stages. In the first stage—AU perception fine-tuning—the model is trained on static frames extracted from the CAS(ME)^3 video dataset, along with their fine-grained AU annotations, to prioritize learning the facial action unit detection task. This enables the model to establish, at the parameter level, a high

sensitivity to subtle facial micro-movements. The second stage is emotion reasoning fine-tuning, where the model parameters obtained from the first stage are loaded and further fine-tuned on the training set of the target-domain FER dataset, guiding the model to learn the joint mapping from "AU detection" to "emotion classification." This effectively achieves the transformation from mid-level facial encoding to high-level emotional semantics. Finally, the model's facial emotion recognition performance is systematically evaluated on the test set of the FER dataset using standardized metrics.

3.3 Data Preprocess

The two datasets employed in this study—CAS(ME)^3 and FER-2013—both provide pre-cropped facial images, thus obviating the need for additional face detection and alignment operations, allowing them to be directly fed into the model for training and evaluation. In light of the distinct characteristics of each dataset, we design differentiated data processing strategies, as detailed below.

3.3.1 Static Image Dataset Processing (FER-2013)

FER-2013 is a static image dataset, where each image corresponds to a single independent emotion label. Specifically, FER-2013 comprises 35,887 grayscale images of 48×48 pixels, covering seven basic emotion categories: anger, disgust, fear, happiness, neutral, sadness, and surprise.

For the static image dataset, we adopt the following preprocessing pipeline:

Size Normalization The 48×48 grayscale images from FER-2013 are upsampled to 224×224 pixels via bicubic interpolation and replicated into three channels to meet the input requirements of the visual encoder; the video frames from CAS(ME)^3 are uniformly adjusted to 112×112 pixels to accommodate the input specifications of OpenFace.

Pixel Value Normalization All input images undergo pixel value normalization. For the OpenFace model, the input images are automatically scaled by a factor of $1/255$ prior to inference. For large models such as Qwen3-VL, the built-in image processor adopts a normalization scheme with a mean of 0.5 and a standard deviation of 0.5, i.e., $(pixel/255 - 0.5)/0.5$, to standardize the input data distribution and accelerate model convergence.

Data Augmentation To enhance the model's generalization performance, we employ online data augmentation strategies including random horizontal flipping (with a probability of 0.5) and random rotation within $\pm 15^\circ$.

3.3.2 Frame-Sequence Dataset Processing (CAS(ME)3)

For the video sequences in the CAS(ME)^3 dataset, we leverage their frame-level expression annotations and design a sampling strategy based on the Apex frame. Specifically, we crop each video clip around the Apex frame (i.e., the frame at which the expression intensity reaches its peak, denoted as I_t), taking the 15 frames preceding the Apex frame and the 14

frames following it to construct a continuous sequence of 30 frames. The sampling scheme is formulated as follows:

$$S = \{I_{t-15}, I_{t-14}, \ldots, I_t, \ldots, I_{t+14}\} \tag{3.1}$$

where S denotes the sampled frame sequence and I_t denotes the Apex frame. For clips with insufficient frames, we pad the sequence to a fixed length by repeating the last frame until the sequence length reaches $L = 30$. This sampling strategy effectively captures the complete dynamic process of expression evolution—from onset to peak to offset—while maintaining the sequence length within a computationally feasible range.

3.3.3 Dataset Splitting

To ensure the reproducibility and fairness of the experiments, we partition each dataset according to the following strategies:

Dataset	Training Set	Validation Set	Test Set
FER2013	28,709	3,589	3,589
CAS(ME)^3	80%	10%	10%

Table 3.1: Dataset partitioning strategies.

For both the training and validation sets, we adopt a stratified sampling strategy to ensure that each class maintains a proportion distribution consistent with the original dataset after partitioning, thereby avoiding bias caused by class imbalance during training. The validation set is used for hyperparameter tuning and early stopping, while the test set is reserved exclusively for final model performance evaluation.

3.4 Qwen3-VL Introduction

Against the backdrop of the rapid advancement of artificial intelligence technologies, multimodal large models are progressively emerging as a core bridge connecting the physical world with digital intelligence. The Qwen3-VL series, introduced by the Qwen team, represents a state-of-the-art contribution to the multimodal domain. Among this series, Qwen3-VL-32B-Instruct, as the largest Dense model variant without reasoning-specialized optimization, achieves exceptional performance across a wide range of multimodal benchmarks while maintaining relatively manageable computational resource consumption. The Qwen3-VL model offers three core advantages: (1) significantly enhanced pure-text understanding capabilities, outperforming comparable pure-text models across multiple evaluations; (2) support for processing text and interleaved multimodal inputs through a proprietary 256K token context window, enabling precise information extraction, retrieval, and cross-document correlation in long documents and videos; and (3) advanced multimodal reasoning capabilities across single-image, multi-image, and video tasks, achieving leading performance on MMMU as well as visual mathematics benchmarks including MathVista and MathVision.

3.5 AU-Detection Fine-Tuning

3.5.1 Fine-Tuning Objective

This stage employs Facial Action Units (AUs) as supervisory signals to train the Qwen3-VL vision encoder for precise perception of subtle facial micro-movements. AUs are objective physical descriptors (e.g., "eyebrow lowering"), not subjective emotion labels, so the acquired knowledge forms a domain-agnostic facial representation with cross-dataset generalizability.

CAS(ME)³ provides comprehensive AU annotations suitable for this purpose, whereas standard FER datasets lack fine-grained AU labels, making them unsuitable for direct AU supervision. End-to-end joint training on both CAS(ME)³ and FER data would introduce domain confusion at the visual feature level, undermining feature transferability.

To address this, a two-stage strategy is adopted: Stage 1 trains exclusively on CAS(ME)³ for AU prediction. Since AUs denote objective facial morphology, the acquired detection capability transfers zero-shot to the FER domain without requiring target-domain AU annotations. Stage 2 performs lightweight fine-tuning of only the AU-to-emotion mapping head using a small set of labeled FER samples, enabling emotion recognition with minimal target supervision.

This paradigm offers three advantages. Interpretability improves as final decisions trace back to specific AU activations, providing physiological grounding. Parameter efficiency and reusability are enhanced because the LoRA adapters from Stage 1 encapsulate generic facial perception knowledge, reusable directly in Stage 2 without retraining the vision encoder. Overall, the strategy reduces reliance on fine-grained target annotations while ensuring cross-domain generalization, balancing performance, efficiency, and interpretability.

3.5.2 LoRA-Based Parameter Efficient Fine-tuning

Since the selected multimodal large model, Qwen3-VL-32B, comprises approximately 32 billion parameters, adopting full fine-tuning would require updating all parameters of the model, which not only incurs enormous computational resource consumption but also poses a high risk of overfitting on small-scale datasets. For the CAS(ME)3 micro-expression dataset, the number of samples is far smaller than the massive vision-language data used during the pre-training phase of Qwen3-VL; therefore, directly updating all model parameters is not only computationally expensive but also prone to disrupting the general visual-semantic knowledge that the model has already acquired. To address this, we adopt the Low-Rank Adaptation (LoRA) strategy, a parameter-efficient fine-tuning (PEFT) approach, to fine-tune the Qwen3-VL model, optimizing only a small number of additional parameters while enabling the model to learn discriminative capabilities for facial action units, all while preserving the original vision-language understanding of the pre-trained model.

For a pre-trained weight matrix $W_0 \in \mathrm{R}^{d \times d}$, LoRA constrains its update manner by representing the incremental parameter matrix ΔW from full fine-tuning as a low-rank approximation of two smaller matrices $B \in \mathrm{R}^{d \times r}$ and $A \in \mathrm{R}^{r \times d}$:

$$W_0 + \Delta W = W_0 + BA \tag{3.2}$$

where the rank r is much smaller than d.

For a given input $x \in \mathrm{R}^d$, the output $h \in \mathrm{R}^d$ after incorporating LoRA is formulated as:

$$h = (W_0 + \Delta W)x = W_0 x + BAx \tag{3.3}$$

During training, the original parameter matrix W_0 is frozen, meaning that although W_0 participates in both forward propagation and backpropagation, its corresponding gradients are not computed, and its parameters are not updated.

At the start of training, the matrix B is initialized to zero, i.e., $B = 0$. The matrix A is initialized via a Gaussian function:

$$\alpha_i \sim N(0, \sigma_\alpha^2) \tag{3.4}$$

This ensures that before training begins, $BA = 0$, so the fine-tuning process starts from the pre-trained weights W_0. The backpropagation process of LoRA is illustrated in Figure 3.1. During training, the pre-trained weights W_0 are frozen, and only the low-rank matrices B and A are optimized. Consequently, after LoRA training, only the parameters of the low-rank matrices B and A need to be saved. Similar to rating matrix factorization in recommender systems, non-negative matrix factorization in text processing, and singular value decomposition, the rank r of the low-rank decomposition approximation matrices B and A determines their fitting capacity. Ideally, we seek a rank r such that the low-rank approximation structure BA can approximate the expressive power of the full fine-tuning incremental matrix ΔW, with closer being better. As the rank r increases, the number of trainable parameters grows accordingly, and the low-rank adaptation capability of LoRA gradually improves, potentially leading to overfitting. The selection of the rank r should be comprehensively determined by the difficulty of the downstream task, the capacity of the baseline model, and the scale of the training data.

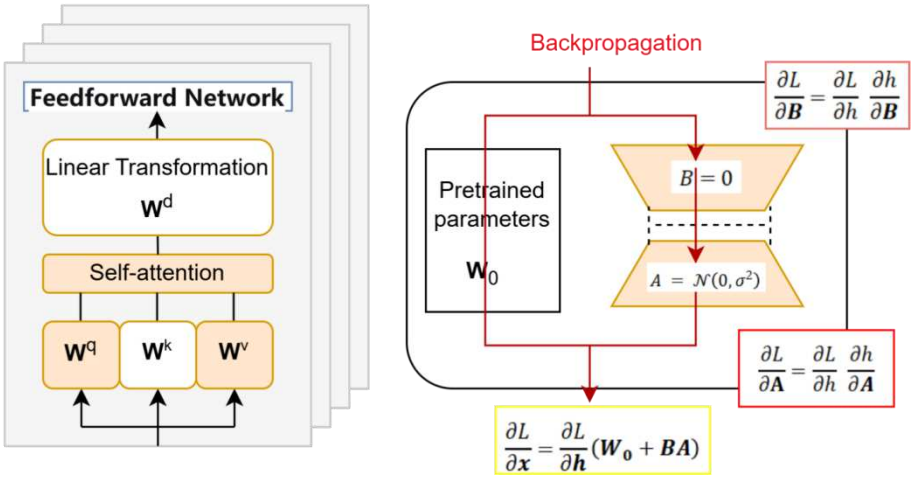


Figure 3.1: The backpropagation process of LoRA

3.5.3 AU-Oriented Instruction Tuning

To enable Qwen3-VL to fully learn the fine-grained facial muscle movement patterns in micro-expression images, this paper does not adopt the traditional classification label supervision approach, but instead constructs a multimodal instruction tuning dataset oriented toward Action Units (AUs). By converting the video sequences from the CAS(ME)3 dataset and their corresponding AU annotations into a Visual Question Answering (VQA) format, the model is enabled to learn the mapping relationships among facial action units within a unified framework of visual understanding and natural language reasoning, thereby enhancing the model's perception of local deformations in micro-expressions.

This subsection employs CAS(ME)3 as the data source for model fine-tuning. CAS(ME)3 is one of the largest publicly available high-frame-rate micro-expression databases, capturing facial video sequences under genuinely induced scenarios using high-speed camera equipment, and is capable of comprehensively recording the dynamic muscle movements

during micro-expression generation. Compared with traditional static expression datasets, CAS(ME)3 provides continuous video frames along with corresponding Action Unit (AU) annotations, which better aligns with the practical requirements of facial action analysis tasks. Therefore, this paper selects CAS(ME)3 as the training data for the AU recognition task, providing fine-grained physiologically meaningful feature representations for subsequent expression recognition.

Qwen3-VL is a vision-language large model whose training objective is not traditional image classification but rather natural language generation based on visual information. Therefore, directly adopting the "image–category label" supervised learning paradigm fails to fully leverage the model's multimodal understanding capabilities. To address this, we reorganize the CASME3 dataset into an Instruction Following data format, converting the original video sequences and their corresponding Action Unit (AU) annotations into visual question-answering samples. This enables the model to learn the AU recognition task in a "observe image–comprehend question–generate answer" manner. The instruction primarily consists of three components:

Task Description: The model is explicitly required to analyze continuous facial image sequences and identify the action units occurring therein.

Prior Knowledge: To reduce the model's difficulty in understanding the FACS framework, we further incorporate into the instruction the major AU numbers along with their corresponding muscular movement semantics—for instance, AU1 denotes inner brow raiser, AU12 denotes lip corner puller, and AU24 denotes lip presser—thereby enabling the model to establish correspondences between muscular movements and action units based on visual information.

Output Constraint: To ensure consistency in the model's output format, the instruction further stipulates that the model must generate AU combinations in a fixed format.

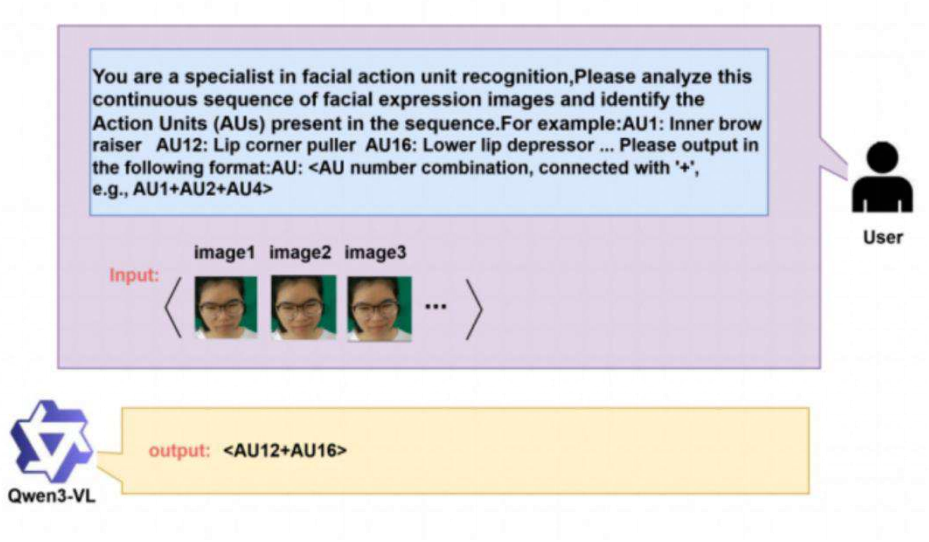


Figure 3.2: An Example of First-stage Fine-tuning Instruction

Unlike traditional AU recognition methods that directly employ classification networks to predict action units, this paper reformulates the AU recognition task as a visual question-answering task, enabling Qwen3-VL to fully leverage its cross-modal understanding capabilities, as shown in Figure 3.2. The continuous image sequences not only preserve the dynamic temporal information inherent in the micro-expression generation process, but the task description, prior knowledge of action units, and output format constraints provided in the instruction further reduce the difficulty of task comprehension for the model, allowing it to

gradually establish semantic correspondences between facial muscle movements and action units.

Moreover, the AU representations learned via Instruction Tuning are not only highly interpretable but can also serve as intermediate semantic representations for subsequent expression recognition tasks. This enables a hierarchical modeling pipeline from "facial muscle movements" to "emotional expressions," thereby enhancing the model's generalizability and interpretability in micro-expression analysis tasks.

3.5.4 Training Objective

In this study, we aim to guide the multimodal large model Qwen3-VL, via instruction tuning, to accurately recognize Action Units (AUs) from sequential facial expression image frames. To achieve this goal, we define a clear training objective function and optimization strategy to balance the modeling of spatiotemporal features within the sequence and the precise prediction of AU labels.

We formulate the AU recognition task as a sequence-to-sequence generation problem and adopt the Supervised Fine-Tuning (SFT) paradigm, with the objective of minimizing the conditional language modeling loss of the model given multi-frame image sequences and textual instructions. Given a sequence of T facial expression image frames along with a human instruction I, the model is required to generate a response Y, i.e., a combination of AU labels. The training objective is defined as maximizing the log-likelihood of the correct response sequence:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{t=1}^{L} \log p_\theta \left(y_t \mid V, \mathcal{I}\right) \tag{3.5}$$

where L is the length of the response sequence, Y_t is the t-th output token, and θ denotes the trainable parameters of the model. This objective function forces the model, at each generation step, to leverage visual features, instruction semantics, and previously generated contextual tokens to progressively reason out the complete set of AUs.

Since the input consists of multiple image frames, we first extract facial representations from each frame using a visual encoder (e.g., ViT), and then aggregate them into a unified visual feature Z_V through a temporal aggregation module. Subsequently, Z_V is concatenated with the text instruction embedding Z_t and fed into the decoder of the large language model. In loss computation, we only calculate the cross-entropy loss on the tokens corresponding to the response part, so as to focus on the accuracy of AU prediction. Formally, the cross-entropy loss function can be expanded as:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{t=1}^{L} \log p_\theta \left(y_t^* \mid V, \mathcal{I}\right) \tag{3.6}$$

where Y_t^* is the probability that the model predicts token c at the t-th position, and c is a token in the vocabulary C. This loss function directly optimizes the model's predictions of AU numbers and their delimiters, ensuring that the sequence format strictly complies with the requirements.

The optimization process during training adopts the AdamW optimizer with a learning rate of $\eta = 5 \times 10^{-5}$, coupled with a Cosine Annealing schedule for learning rate decay. The batch size is set to 2 to balance GPU memory consumption and gradient stability. Although the cross-entropy loss effectively improves generation accuracy, given the stringent requirements of the AU recognition task regarding the rigor and precision of outputs, we introduce a format penalty term into the loss function. During the validation phase, rewards or penalties

are assigned based on regular expression matching; however, in the training phase, cross-entropy remains the primary loss to maintain optimization stability. Overall, this training objective effectively integrates multimodal feature fusion, parameter-efficient fine-tuning, and autoregressive generation, providing a solid optimization foundation for AU recognition.

3.6 Expression Recognition Fine-tuning

After AU recognition fine-tuning on CAS(ME)³, the model acquires AU detection capability. The second stage transfers this knowledge to emotion recognition, as FER datasets provide only emotion labels without AU annotations.

3.6.1 Fine-Tuning Objective

Three objectives guide this stage. First, mapping AU space to emotion space: emotion categories correspond to combinations of multiple AUs; the model must learn a nonlinear mapping from AU activation patterns to high-level affective semantics. Second, alleviating sparse supervision: direct end-to-end training on emotion labels risks learning shallow visual correlations and ignoring facial action structures. Pretrained AU knowledge provides structured intermediate representations, enabling the model to leverage limited emotion labels to reinforce the AU-to-emotion reasoning pathway. Third, task adaptation with knowledge preservation: while the first-stage model excels at AU description, its output format is biased toward AU detection. The second stage adjusts outputs to emotion recognition format without forgetting AU knowledge, leveraging the model's inherent reasoning for improved interpretability.

3.6.2 Fine-Tuning Strategy

Lightweight Parameter Updates Based on LoRA. Consistent with Stage 1, LoRA is applied to Query and Value matrices in attention modules, freezing all original parameters. This reduces trainable parameters (0.1% of total), mitigating overfitting on small FER datasets, and preserves Stage 1 AU knowledge by restricting the update subspace.

Continuation of Task Encapsulation in the Visual Question Answering Format. To ensure task coherence between the two stages of fine-tuning, this stage still formulates the expression recognition task in a visual question answering format. Given an input facial image, a question text following a fixed template is constructed. Differing from the first stage, the model's output now requires not only the recognized AU categories but also the final predicted emotion category, accompanied by a reasoning explanation based on the AUs. This design maintains consistency between the second-stage fine-tuning and the first stage in terms of data format and interactive paradigm, reducing the disturbance of task switching on the model's internal representations, while facilitating subsequent evaluation of the model's joint capabilities within a unified framework. The specific fine-tuning instructions are illustrated in Figure 3.3.

Single-Stage Fine-Tuning and Freezing Strategy. Only newly initialized LoRA parameters for emotion recognition are optimized; the visual encoder, language backbone, and Stage 1

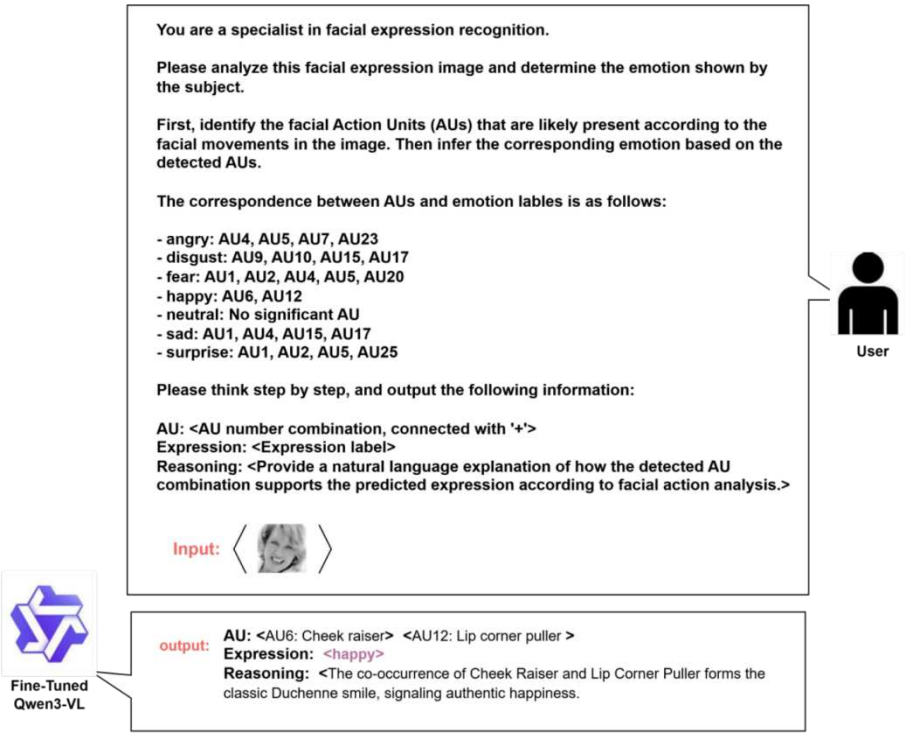


Figure 3.3: Second-stage Fine-tuning Strategy

LoRA adapters remain frozen. This decouples the parameter spaces of the two tasks, preserving a clear AU-to-emotion reasoning chain and enabling independent ablation studies.

Prompt Engineering and Output Format Constraints. The system prompt restricts output to a single emotion category word (neutral, happiness, sadness, surprise, fear, disgust, anger), aligned with the label space, preventing redundant or non-standard responses.

3.7 Decision-Level Fusion with OpenFace 3.0

In the emotion recognition module, to enhance the robustness and accuracy of classification results, this paper proposes a decision-level weighted fusion strategy that integrates the logits output from the OpenFace 3.0 vision model with the predictions from the large language model (LLM).

Specifically, let the emotion category probability distribution output by OpenFace 3.0 be denoted as $P_{OF} \in R^7$, which is obtained through a single fully-connected mapping of the "unified facial features." Meanwhile, let the probability distribution output by the large language model based on non-visual information (e.g., dialogue text or context) be denoted as $P_{LLM} \in R^7$. To avoid misclassification caused by a single modality under specific noisy conditions, we employ weighted linear interpolation to fuse the two distributions. The final predicted probability P_{final} is computed as follows:

$$P_{final} = \alpha \cdot P_{OF} + (1 - \alpha) \cdot P_{LLM} \tag{3.7}$$

where the weight coefficient $\alpha \in [0, 1]$ denotes the contribution of the visual modality. This weight can be either a hyperparameter tuned on the validation set to find the optimal

ratio (e.g., set to a fixed value $\alpha = 0.6$), or designed as an adaptive weight that is dynamically adjusted based on face detection confidence or emotional context consistency.

Finally, the fused emotion label $\hat{y}$ is determined by selecting the category with the highest probability:

$$\hat{y} = \arg\max_{c \in \{1,\ldots,7\}} P_{\text{final}}(c) \tag{3.8}$$

The advantages of this fusion strategy are twofold: on the one hand, the visual features (facial muscle movements) provided by OpenFace 3.0 and the semantic contextual features captured by the large language model are inherently complementary; on the other hand, decision-level fusion offers better interpretability compared to feature-level fusion, is insensitive to adjustments in model architecture, and is easy to deploy and scale in practical systems.

Chapter 4

Experiments and Results

4.1 Datasets

In our work, we primarily employ two datasets: CAS(ME)^3 and FER2013. CAS(ME)^3 provides fine-grained AU temporal annotations (onset frame, apex frame, and offset frame) along with a large sample size, enabling the model to receive precise supervisory signals for facial micro-movements. This effectively drives the model to learn the spatiotemporal dynamic features of micro-expressions and enhances its reasoning capability for subtle facial actions. FER2013 is used to validate whether the model can effectively transfer the micro-movement representations learned from fine-grained micro-expression data to conventional expression recognition tasks with different data distributions, thereby comprehensively evaluating the model's generalizable representation capability and domain adaptability.

4.2 Evaluation Metrics

To comprehensively evaluate the performance of our proposed two-stage fine-tuning framework with decision-level fusion, we adopt a suite of standard and robust metrics for multi-class emotion classification on the FER2013 test set.

Specifically, we report Accuracy (ACC) as the primary metric, which reflects the overall proportion of correctly classified samples. It is defined as:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4.1}$$

To account for class-wise performance, we calculate Precision and Recall for each emotion category, capturing the model's ability to avoid false positives and false negatives, respectively. They are defined as:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4.2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.3}$$

Furthermore, we report the Macro F1-Score and Weighted F1-Score as complementary metrics. The Macro F1-Score averages the F1 scores across all emotion classes equally, providing a balanced assessment regardless of class frequency, while the Weighted F1-Score weights each class's contribution by its support, offering a class-distribution-aware evaluation. These two F1 variants are particularly important given the inherent class imbalance in

FER2013, where certain emotion categories (e.g., disgust) have fewer samples than others (e.g., happiness, neutrality). They are defined as:

$$\text{Macro F1} = \frac{1}{N}\sum_{i=1}^{N}\text{F1}_i \tag{4.4}$$

$$\text{Weighted F1} = \sum_{i=1}^{N} w_i \cdot \text{F1}_i, \quad w_i = \frac{n_i}{\sum_{j=1}^{N} n_j} \tag{4.5}$$

In addition to the scalar metrics, we visualize the Confusion Matrix to qualitatively analyze the model's classification behavior, especially the confusions between semantically similar emotion pairs (e.g., fear vs. surprise, sadness vs. disgust). This allows us to identify systematic misclassification patterns and understand the model's perceptual ambiguities.

All reported results are computed using the widely-adopted scikit-learn implementation, ensuring reproducibility and fair comparison with prior studies.

Specifically, for the multi-class classification task in this paper, let TP_i, FP_i, and FN_i denote the true positives, false positives, and false negatives for the *i*-th class in the confusion matrix, respectively. Then:

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \quad \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}, \quad \text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{4.6}$$

The Macro F1-Score and Weighted F1-Score are respectively computed as:

$$\text{Macro F1} = \frac{1}{N}\sum_{i=1}^{N}\text{F1}_i, \quad \text{Weighted F1} = \sum_{i=1}^{N}\frac{n_i}{\sum_{j=1}^{N} n_j}\cdot \text{F1}_i \tag{4.7}$$

where $N = 8$ is the total number of emotion categories, and n_i is the number of samples belonging to the *i*-th class in the test set.

4.3 Model Performance

4.3.1 Experimental Setup

All experiments are conducted on 4 × NVIDIA A100 (40GB) GPUs for supervised fine-tuning (SFT) of the Qwen3-VL-32B-Instruct model. To maximize GPU memory utilization, we do not employ quantization strategies, thereby preserving the full precision of the model. To improve efficiency for long-context multimodal inputs, we enable FlashAttention. The batch size is set to 2 and the maximum sequence length to 8196 to avoid exceeding the memory limit during training. The learning rate is set to 5×10^{-5}, a common choice for fine-tuning large models, in conjunction with a cosine annealing scheduler to prevent overfitting. The number of training epochs is set to 15 to fully exploit the fitting potential of both the model and the dataset. The detailed training hyperparameters are presented in Table 4.1.

During the inference stage, we employ the vLLM framework on 2 NVIDIA A100 GPUs to reduce computational cost.

4.3.2 Overall Model Performance

We evaluate the performance of the fine-tuned Qwen3-VL model, the fine-tuned Qwen3-VL model with decision-level fusion with OpenFace 3.0, the non-fine-tuned Qwen3-VL model,

Category	Parameter	Value	Description
Model	Base Model	Qwen3-VL-32B-Instruct	Pre-trained multimodal LLM
	Fine-Tuning Method	LoRA	Parameter-efficient fine-tuning
	LoRA Rank	8	Low-rank matrix dimension
	LoRA Alpha	16	Scaling coefficient
	LoRA Target	all	Applied to all linear layers
	Quantization	None	Full-precision training
Training	Batch Size (per device)	2	Samples per step per GPU
	Gradient Accumulation Steps	8	Effective batch size = $2 \times 4 \times 8 = 64$
	Learning Rate	5×10^{-5}	AdamW optimizer
	Learning Rate Scheduler	Cosine	Cosine annealing
	Warmup Steps	0	—
	Training Epochs	15	—
	Max Gradient Norm	1.0	Gradient clipping threshold
Data	Max Sequence Length	8192	Upper limit of input tokens
	Image Max Pixels	768	Upper limit of pixels per image
	Image Min Pixels	32	Lower limit of pixels per image
Distributed	Parallelism Strategy	DeepSpeed ZeRO-3	Parameter sharding
	CPU Offload	Enabled	Offload optimizer states to CPU
Precision	Data Type	FP32	Single-precision training
	Flash Attention	Enabled	Accelerate attention computation

Table 4.1: Key hyperparameter configurations for fine-tuning experiments.

and the OpenFace 3.0 model on the FER-2013 dataset. In addition, we compare their performance against several models that have demonstrated strong results on FER-2013, including ResNet50, VGG16, DenseNet, and GoogLeNet.

The experimental results are presented in Table 4.2

Model	ACC	MACRO-Precision	Weighted Precision	MACRO-RECALL	Macro F1-score	Weighted Recall	Weighted F1
Fine-tuned QWEN	66.73%	67.21%	68.89%	58.64%	57.29%	66.73%	65.80%
Fused model	67.37%	64.87%	69.47%	62.71%	59.09%	67.37%	66.52%
Non-fine-tuned QWEN	58.73%	51.80%	60.64%	52.76%	47.45%	58.73%	55.95%
OpenFace 3.0	48.02%	44.69%	56.31%	49.38%	43.29%	48.02%	49.73%
ResNet50	69.33%	69.88%	69.84%	65.92%	67.49%	69.33%	69.35%
VGG-16	68.75%	69.28%	68.53%	66.86%	67.89%	68.75%	68.58%
DenseNet	69.38%	69.86%	69.31%	67.23%	68.18%	69.38%	69.10%
GoogLeNet	67.04%	69.56%	67.70%	64.18%	66.05%	67.04%	67.04%

Table 4.2: Performance comparison of different models on the FER-2013 test set.

4.3.3 Advantages of the Fused Model over the Pure Large Model

The experimental results demonstrate that the fused model achieves an overall accuracy of 67.37%, representing an improvement of approximately 0.64 percentage points over the non-fused large model (66.73%). Although the improvement is modest, the fusion strategy exhibits clear advantages in fine-grained metrics.

In terms of recall, the fused model achieves a Macro-Recall of 62.71%, substantially higher than the baseline model's 58.64% (an improvement of approximately 4.07%). This indicates that after incorporating the geometric features from OpenFace 3.0 at the decision level, the model's ability to capture minority-class samples (e.g., disgust, fear) has been significantly enhanced, effectively alleviating the bias toward head classes (e.g., happiness, neutral) that is inherent in a single multimodal large model when trained on imbalanced data. Meanwhile, the Weighted Recall increases from 66.73% to 67.37%, consistent with the trend in accuracy, further validating the effectiveness of the fusion strategy on the overall sample set.

The decision-level fusion of a multimodal large model and a specialized geometric feature extraction model offers significant complementary advantages. While the pure vision-language model (Qwen3-VL) possesses powerful high-level semantic understanding capabilities, it exhibits limitations in capturing subtle facial geometric deformations, especially for low-intensity expressions. OpenFace 3.0, with its facial landmark-based geometric features, effectively compensates for this shortcoming. Their joint reasoning not only improves overall classification accuracy but also significantly enhances the model's robustness toward easily confused and rare emotion categories, endowing the system with stronger generalization capability in emotion recognition tasks that more closely resemble real-world scenarios.

4.3.4 Effectiveness of Fine-Tuning

As clearly shown in Table 4.2, the model after two-stage fine-tuning significantly outperforms the non-fine-tuned baseline model across all evaluation metrics. Specifically, the fine-tuned model achieves an overall accuracy (ACC) of 66.73%, representing an improvement of 8 percentage points over the baseline model's 58.73%, corresponding to a relative improvement of 13.6%. This result indicates that the proposed two-stage fine-tuning strategy effectively

unlocks the potential of the Qwen3-VL model on facial expression recognition tasks, enabling it to better capture emotion-related discriminative features from facial images. The significant improvements across all evaluation dimensions fully demonstrate the rationality and superiority of this training paradigm.

4.3.5 Advantages of the Fused Model over Pure OpenFace 3.0

As shown in Table 4.2, the joint reasoning framework proposed in this paper significantly outperforms OpenFace 3.0 standalone inference across all evaluation metrics. Specifically, the joint reasoning framework achieves an accuracy of 67.37%, an improvement of 19.35 percentage points over OpenFace 3.0 standalone inference (48.02%), corresponding to a relative improvement of 40.3%. This result indicates that relying solely on the specialized facial Action Unit analysis model OpenFace 3.0 is insufficient for fully capturing the subtle semantic differences and contextual information in facial expressions, as its reasoning capability is constrained by predefined geometric and motion feature representations. The experimental results fully validate the effectiveness of the proposed two-stage fine-tuning strategy and decision-level joint reasoning framework for facial expression recognition, achieving comprehensive and significant performance improvements over the specialized model alone, and demonstrating the potential and advantages of collaborative reasoning between multimodal large models and specialized models.

4.3.6 Comparative Experimental Analysis

This section compares our proposed method, which combines the Qwen3-VL multimodal large model with OpenFace 3.0 through decision-level joint reasoning, against four classic deep convolutional neural networks (ResNet-50, VGG-16, DenseNet, and GoogLeNet).

As shown in Table 4.2, all four traditional CNN models achieve relatively stable performance on the FER-2013 dataset, with DenseNet performing the best, achieving the highest accuracy (69.38%) and Macro-F1 (68.18%), while GoogLeNet achieves the lowest accuracy (67.04%). In comparison, our proposed joint reasoning model achieves an accuracy of 67.37%, slightly outperforming GoogLeNet but falling behind ResNet-50, VGG-16, and DenseNet.

Although our model has not yet surpassed the optimal specialized CNN models (DenseNet, 69.38%) in terms of overall accuracy, a deeper analysis of fine-grained metrics reveals that our joint reasoning framework exhibits several significant and noteworthy advantages. First, in terms of prediction reliability and confidence, our model performs exceptionally well. Its Weighted Precision reaches 69.47%, almost on par with the best-performing model among all comparators, ResNet-50 (69.84%), while significantly outperforming VGG-16 (68.53%) and GoogLeNet (67.70%). This result indicates that when our model predicts a particular emotion category, its prediction is highly trustworthy—the “precision-first” reasoning characteristic enables the model to provide more reliable outputs in critical decision-making scenarios (e.g., key-frame judgments in human-computer interaction), effectively reducing the risk of false positives. Compared to CNN models, which may incur increased misclassifications when pursuing high recall, our model exhibits a more cautious and well-defined decision boundary.

Second, in terms of inter-class discriminative ability, our model’s Weighted Precision (69.47%) is substantially higher than its Macro Precision (64.87%), with a gap of approximately 4.6 percentage points. This indicates that the model demonstrates strong recognition focus and representational capability on the majority emotion categories with larger sample sizes. This characteristic is highly valuable in practical application scenarios—real-world

facial expression data often exhibit a pronounced long-tail distribution, and our model's ability to prioritize recognition accuracy on high-frequency categories may translate to superior practical performance in real-world deployment compared to laboratory-balanced test sets.

Third, the multimodal decision-level fusion architecture adopted by our model itself possesses important methodological advantages. Unlike "black-box" end-to-end CNNs, our framework explicitly combines Qwen3-VL's deep semantic understanding capabilities with OpenFace 3.0's geometric structure priors, forming a more interpretable reasoning chain. This design allows researchers to clearly trace the source of errors when model predictions deviate (whether due to visual semantic bias or geometric feature bias), providing a clear direction for subsequent targeted optimization. In contrast, performance improvements in pure CNN models typically rely on brute-force trial-and-error through network architecture search or data augmentation, lacking the modular iterative space offered by our framework.

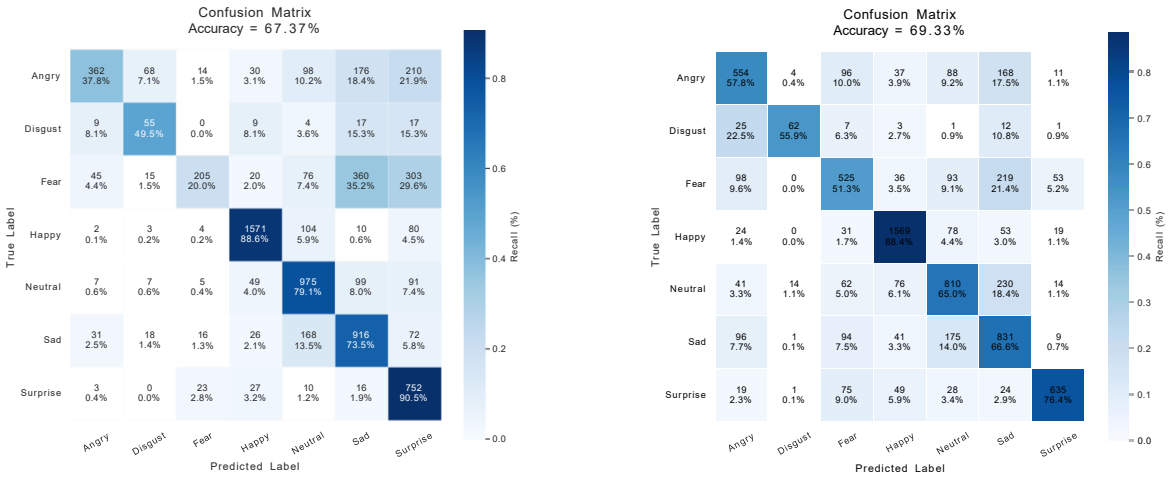
Fourth, from the perspective of feature complementarity, the Weighted Precision advantage of our model suggests the potential gains of multimodal feature fusion. The geometric information provided by OpenFace 3.0—such as head pose and facial Action Unit parameters—together with the contextual semantic knowledge acquired by the large language model through large-scale pre-training, forms effective heterogeneous information complementarity at the decision level. Although the current simple decision fusion strategy has not yet fully unleashed this complementary potential, the Weighted Precision of 69.47% already demonstrates the feasibility of this technical approach—the two information sources reach a consensus in determining "positive predictions," providing a solid experimental foundation for the subsequent introduction of more sophisticated attention mechanisms or adaptive fusion weights.

Finally, it is worth emphasizing that as an initial exploration of synergizing a general-purpose multimodal large model with specialized tools, our model's performance has already approached or even partially surpassed (in terms of Weighted Precision) fully optimized specialized CNN models. Considering that Qwen3-VL was not specifically designed for FER tasks and that our model achieved performance comparable to specialized models after only two-stage fine-tuning, this result itself demonstrates the enormous application potential of large models in the field of affective computing. With further optimization of the fine-tuning strategy and deeper design of the multimodal fusion mechanism, our framework is expected to maintain its high precision advantage while addressing the recall shortfall, ultimately achieving comprehensive superiority.

In summary, although the accuracy metric has not yet reached the optimal level, our model demonstrates clear superiority and promising development prospects across multiple dimensions, including prediction confidence, high-frequency category focus, architectural interpretability, feature complementarity validation, and large model application potential. These findings lay a solid and inspiring foundation for subsequent improvements in this research.

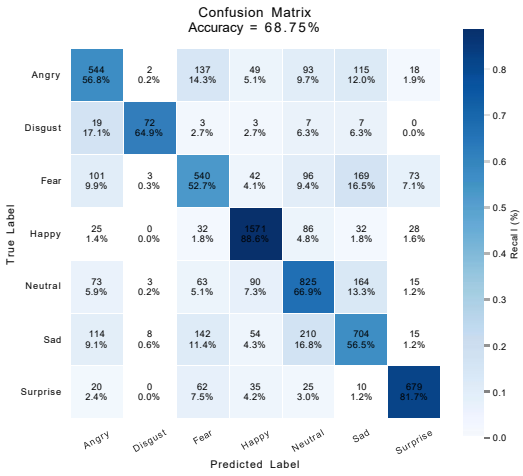
4.3.7 Confusion Matrix Analysis

To further analyze the classification capabilities of different models in fine-grained expression recognition tasks, we compare the confusion matrix results of the Qwen3-VL and OpenFace 3.0 decision-level fusion model against VGG-16, ResNet50, DenseNet, and GoogLeNet on the FER-2013 test set, as shown in Figure 4.1. Although the fusion model does not completely surpass traditional CNN models in single-category recognition counts for certain categories (e.g., angry and disgust), it demonstrates clear advantages in neutral, happy, surprise, and holistic emotional semantic understanding. Compared to CNN models that rely solely on local visual texture features, the decision-level fusion of Qwen3-VL and OpenFace 3.0 combines

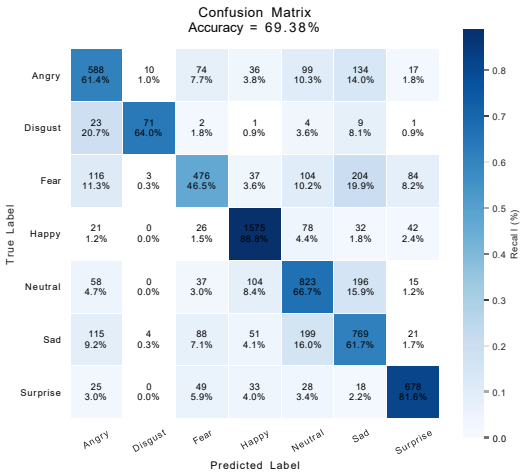


(a) Fused Model

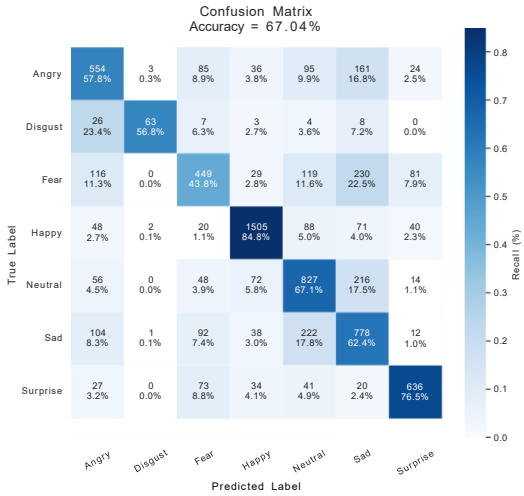
(b) ResNet50



(c) VGG-16



(d) DenseNet



(e) GoogLeNet

Figure 4.1: Confusion Matrices of Different Models

deep visual semantic information with facial dynamic features, achieving complementary advantages and enhancing the model's ability to recognize complex expression variations and weakly expressed emotions. More importantly, the introduction of AU knowledge brings irreplaceable benefits across multiple dimensions: first, it provides interpretable biomechanical constraints through AU activation/inhibition patterns, reducing misclassifications caused by visual ambiguity; second, it strengthens discriminative evidence for subtle muscle differences across categories, particularly effective for categories like neutral and disgust that rely on specific AU combinations; third, under conditions of data imbalance and weak annotation, the AU prior serves as an effective regularization mechanism, improving the model's robustness on few-shot and difficult-to-classify categories. The confusion matrix results further validate the effectiveness of the proposed fusion method on the FER-2013 facial expression recognition task, and reveal the core value of the AU-guided fine-tuning strategy in enhancing fine-grained expression discrimination capability.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This paper proposes a multi-modal large-model enhancement framework based on a two-stage fine-tuning strategy. In the first stage, AU question-answer pairs are constructed using the CAS(ME)³ dataset, and the Qwen3-VL-32B-Instruct model is fine-tuned with LoRA to acquire facial action unit recognition capabilities. In the second stage, the model is further fine-tuned on the FER dataset to enable reasoning mapping from AUs to emotions, thereby effectively bridging the semantic gap between low-level facial deformation perception and high-level emotional semantic understanding. Experimental results demonstrate that the two-stage fine-tuned model outperforms the non-fine-tuned baseline in overall accuracy, and significantly surpasses both the standalone inference of the specialized lightweight model OpenFace 3.0 and other mainstream deep models, confirming the effectiveness of the proposed strategy in eliciting discriminative features for low-intensity expressions from large visual models.

This paper further explores a decision-level fusion mechanism between the fine-tuned large model and the specialized small model OpenFace 3.0, revealing a clear complementarity in their feature representations—the large model excels at complex contextual semantic understanding and cross-scenario generalization, while the small model is more sensitive to local responses to facial geometric deformations. The joint prediction performance after fusion not only surpasses the best standalone inference results of each model, but also significantly alleviates the deficiency of pure large visual models in perceiving subtle deformations due to the lack of fine-grained spatial priors. Meanwhile, by employing AUs as interpretable intermediate representations during inference, the framework enhances the traceability and trustworthiness of model outputs, providing a feasible technical paradigm for robust deployment of FER systems in real-world open scenarios.

5.2 Future Work

Although the proposed method achieves competitive performance on public datasets, several limitations and future directions warrant further exploration. First, the two-stage fine-tuning relies heavily on the quality and coverage of AU-annotated data; the current CAS(ME)³ dataset remains limited in terms of emotion categories and sample diversity. Future work may introduce larger-scale, multi-cultural AU annotation databases to enhance the model's cross-demographic generalization capability. Second, while decision-level fusion achieves a degree of complementary advantages, the fusion strategy remains relatively simplistic. Subsequent research could investigate attention-based dynamic weighted fusion or early-late multi-level

fusion to more fully exploit the synergistic potential among heterogeneous models. Third, the current experiments are primarily conducted on static images or video clips under controlled laboratory conditions; future efforts should validate online inference on real-world in-the-wild long video streams, and incorporate lightweight pruning and distillation under computational resource constraints to facilitate practical deployment on mobile and edge devices. Finally, how to introduce causal reasoning or uncertainty estimation mechanisms within the large-model framework, enabling the FER system to provide more reasonable confidence expressions and rejection strategies when faced with ambiguous or occluded faces, constitutes another important direction worthy of in-depth investigation. In summary, this work provides valuable empirical evidence and theoretical references for the fine-grained adaptation of multi-modal large models in the field of affective computing, and is expected to contribute novel technical support to application scenarios such as digital mental health, intelligent education, and human-machine empathic interaction.

Bibliography

[Don+15] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. “Long-term recurrent convolutional networks for visual recognition and description”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 2625–2634. DOI: 10.1109/CVPR.2015.7298878.

[EF78] Ekman, P. and Friesen, W. V. “Facial action coding system”. In: *Environmental Psychology and Nonverbal Behavior* (1978).

[HGJ+25] Han, S., Gao, M., Jiang, M., et al. “Uncertainty-Aware Collaborative System of Large and Small Models for Multimodal Sentiment Analysis”. In: *arXiv preprint* (2025). arXiv: 2509.04459 [cs.CL].

[He+16] He, K., Zhang, X., Ren, S., and Sun, J. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[Hu+24] Hu, G., Lan, X., Jiang, H., Wang, S., et al. “Towards unified facial action unit recognition framework by large language models”. In: *arXiv preprint* (2024). arXiv: 2409.08444 [cs.CV].

[Hu+25] Hu, J., Mathur, L., Liang, P. P., and Morency, L.-P. “OpenFace 3.0: A Lightweight Multitask System for Comprehensive Facial Behavior Analysis”. In: *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*. May 2025, pp. 1–11. DOI: 10.1109/FG.2025.11099277.

[Ind+23] Indolia, S., Nigam, S., Singh, R., Singh, V. K., and Singh, M. K. “Micro Expression Recognition Using Convolution Patch in Vision Transformer”. In: *IEEE Access* 11 (2023), pp. 100495–100507. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3314797.

[LYZ16] Li, X., Yu, J., and Zhan, S. “Spontaneous facial micro-expression detection based on deep learning”. In: *Proceedings of the 2016 IEEE 13th International Conference on Signal Processing (ICSP)*. Chengdu, China, Nov. 2016, pp. 1130–1134.

[Li+24] Li, Y., Dao, A., Bao, W., Wang, Y., et al. “Facial affective behavior analysis with instruction tuning”. In: *European Conference on Computer Vision (ECCV)*. Cham: Springer, 2024, pp. 165–186.

[Lia+25] Lian, Z., Chen, H., Chen, L., Sun, L., Chen, Y., Wang, J., Li, Z., and Tao, J. “AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models”. In: *arXiv preprint* (2025). arXiv: 2501.16566 [cs.CL].

[LLL18] Liu, Y. J., Li, B. J., and Lai, Y. K. “Sparse MDMO: Learning a discriminative feature for micro-expression recognition”. In: *IEEE Transactions on Affective Computing* 12.1 (2018), pp. 254–261. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2018.2873887.

[PKO09] Polikovsky, S., Kameda, Y., and Ohta, Y. “Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor”. In: *Proceedings of the 3rd International Conference on Crime Detection and Prevention (ICDP 2009)*. London, UK: IET, Dec. 2009.

[Qwe24] Qwen Team. *QVQ: To See the World with Wisdom*. Hugging Face. Model page. 2024. URL: https://huggingface.co/Qwen/QVQ-72B-Preview.

[Ryu+25] Ryumina, E., Markitantov, M., Axyonov, A., and Karpov, A. “Zero-Shot Multimodal Compound Expression Recognition Approach using Off-the-Shelf Large Visual-Language Models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025, pp. 71–79.

[Sun+20] Sun, B., Cao, S., Li, D., He, J., and Yu, L. “Dynamic Micro-Expression Recognition Using Knowledge Distillation”. In: *IEEE Transactions on Affective Computing* 13.2 (2020), pp. 1037–1043. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2020.2988262.

[Wan+26] Wang, S.-J., Liu, X., Li, J., See, J., Yap, M. H., Li, X., Hong, X., Fan, X., Wang, Y., Fang, R., et al. *CAS(ME)3: A Third Generation Facial Spontaneous Micro-Expression Database With Depth Information and High Ecological Validity*. Manuscript in preparation / Under review. 2026.

[Wan+21] Wang, S., Li, B., Liu, Y., Yan, W., Ou, X., Huang, D., Fu, X., and Zhou, C. “Auxiliary Information Guided Micro-Expression Recognition”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 5621–5633. DOI: 10.1109/TIP.2021.3083777.

[Wan+23] Wang, Y., Zheng, S., Sun, X., Guo, D., and Lang, J. “Micro-expression recognition with attention mechanism and region enhancement”. In: *Multimedia Systems* 29.5 (2023), pp. 3095–3103. ISSN: 0942-4962. DOI: 10.1007/s00530-023-01148-0.

[Xie+20] Xie, H.-X., Lo, L., Shuai, H.-H., and Cheng, W.-H. “AU-assisted graph attention convolutional network for micro-expression recognition”. In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 2871–2880. DOI: 10.1145/3394171.3413928.

[Zha+17] Zhang, K., Huang, Y., Du, Y., and Wang, L. “Facial expression recognition based on deep evolutional spatial-temporal networks”. In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4193–4203. ISSN: 1057-7149. DOI: 10.1109/TIP.2017.2710078.

[ZP07] Zhao, G. and Pietikainen, M. “Dynamic texture recognition using local binary patterns with an application to facial expressions”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.6 (June 2007), pp. 915–928. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.1110.

[Zha+21] Zhao, S., Tao, H., Zhang, Y., Xu, T., Zhang, K., Hao, Z., and Chen, E. “A two-stage 3D CNN based learning method for spontaneous micro-expression recognition”. In: *Neurocomputing* 448 (2021), pp. 276–289. DOI: 10.1016/j.neucom.2021.03.089.

[Zhi+11] Zhi, R., Flierl, M., Ruan, Q., and Kleijn, W. B. “Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41.1 (Feb. 2011), pp. 38–52. ISSN: 1083-4419. DOI: 10.1109/TSMCB.2010.2044782.

[ZLZ20] Zhi, R., Liu, M., and Zhang, D. “A comprehensive survey on automatic facial action unit analysis”. In: *The Visual Computer* 36.5 (2020), pp. 1067–1093. ISSN: 0178-2789. DOI: 10.1007/s00371-019-01757-x.

[Zho+24] Zhong, Y., Wei, H., Yang, P., Wang, Y., Li, H., and Liu, J. “ExpCLIP: Bridging text and facial expressions via semantic alignment”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 7. 2024, pp. 7614–7622. DOI: 10.1609/aaai.v38i7.28562.