

# Robust Stage-Wise LVLM Adaptation: Multi-Phase Prompt Lora Fine-tuning for Compound Expression Recognition

Xilong Lu, Jun Yu\*, Yunxiang Zhang, Lingsi Zhu, Yang Zheng, Yongqi Wang, Qiang Ling  
 University of Science and Technology of China  
 Hefei, Anhui, China

luxilong@mail.ustc.edu.cn,

harryjun@ustc.edu.cn,

{mesa, ls-zhu24, zhengyang, wangyongqi}@mail.ustc.edu.cn,

qling@ustc.edu.cn

## Abstract

*Compound Expression Recognition (CER) is crucial for understanding human emotions and improving human-computer interaction. However, CER faces challenges due to the complexity of facial expressions and the difficulty of capturing subtle emotional cues. To surmount these obstacles, we present a novel approach that harnesses the power of Large Vision-Language Models (LVLMs). Our methodology incorporates a two-stage fine-tuning process, complemented by the design of exclusive prompts. In the first stage, pre-trained LVLMs are fine-tuned on basic facial expressions to establish fundamental patterns. Subsequently, in the second stage, the model is further optimized on a compound-expression dataset to refine the interactions between compound expressions. Our approach has achieved remarkable results. It has attained advanced accuracy on the RAF-DB dataset and demonstrated robust zero-shot generalization on the C-EXPR-DB dataset. Notably, in the 8th ABAW Compound Expression Recognition Challenge, our method secured the first place with an F1 score of 0.5723, highlighting its great potential for real-world applications in emotion analysis and human-computer interaction.*

## 1. Introduction

In the rapidly evolving field of artificial intelligence, emotion recognition has emerged as a critical research area, serving as a cornerstone for enabling natural human-computer interaction and deepening our understanding of human psychology [28]. Among its subfields, Compound Expression Recognition (CER), which focuses on identi-

fying multiple co-occurring facial emotions, occupies a unique position. Human facial expressions rarely convey a single emotion; instead, they often reflect a blend of emotions, a complexity that CER aims to capture. This pursuit is not only academically intriguing but also holds significant practical relevance. However, the inherent complexity of compound expressions, characterized by subtle nuances and intricate combinations of facial muscle movements, poses significant challenges, making CER a focal point of research.

Recent advancements in Large Vision-Language Models (LVLMs) have revolutionized multimodal data processing. These models are designed to seamlessly integrate visual data, such as facial features, with language data, such as emotion-related textual descriptions [3]. This unique capability bridges the gap between vision and language, opening new avenues for CER. By learning from both modalities simultaneously, LVLMs can uncover previously hidden emotional patterns. For example, an LVLM might correlate the visual feature of a tightened jaw (associated with anger or stress) with the textual description "feeling pressured" to more accurately classify a compound emotion [29].

In the context of the Affective Behavior Analysis in-the-Wild (ABAW) competitions [10, 11, 15–17, 21, 23–25], significant progress has been made in CER and related tasks. The ABAW competitions have consistently pushed the boundaries of emotion recognition by introducing large-scale datasets and challenging tasks such as multi-label compound expression recognition, valence-arousal estimation, and action unit detection [13, 14, 18–20, 22, 30]. These competitions have highlighted the importance of multi-task learning and the integration of diverse modalities for robust emotion recognition. For instance, the 7th ABAW competition focused on multi-task learning and compound expression recognition, emphasizing the need for models

---

\*Corresponding author

that can simultaneously handle multiple emotion-related tasks [24]. Similarly, the 6th ABAW competition demonstrated the effectiveness of distribution matching techniques for multi-task learning, which can be applied to CER to improve generalization across different emotional categories [23].

In this paper, we present a novel stage-wise LoRA (Low-Rank Adaptation) fine-tuning approach to optimize pre-trained models for CER, especially in resource-constrained scenarios. LoRA, which reduces the number of trainable parameters by introducing low-rank matrices [9], effectively overcomes the computational hurdles of fine-tuning large models. Our stage-wise strategy allows the model to learn task-specific patterns more efficiently. By adjusting the model's parameters in a phased manner, according to the data characteristics and task demands at each stage, the model can better capture the complex patterns of compound expressions, thus enhancing its generalization ability. We also emphasize the significance of well-designed context prompts for LVLMs in accurate CER. These prompts offer background knowledge and semantic guidance, enabling the model to concentrate on crucial features during recognition.

In summary, by integrating the advantages of LVLMs, the stage-wise LoRA fine-tuning strategy, and innovative context prompt design, we have developed an efficient and highly accurate CER method. Our proposed method has achieved remarkable results. In the 8th ABAW Compound Expression Recognition Challenge, it ranked first with an F1 score of 0.5723. This work aims to resolve long-standing challenges in this field, promoting the practical application of emotion recognition technology and facilitating more intelligent and empathetic human-machine interactions.

## 2. Related Work

In this section, we review existing research in Compound Expression Recognition (CER) and fine-tuning methods for Large Vision-Language Models (LVLMs). This review provides a comprehensive understanding of the current state of research and positions our proposed approach within the broader context.

### 2.1. Compound Expression Recognition

Compound Expression Recognition focuses on identifying multiple emotional states simultaneously expressed on the face. This research direction is of significant importance in artificial intelligence, as it captures the complexity of human emotions more accurately than single-emotion recognition. However, the high complexity and diversity of compound expressions, characterized by subtle variations and intricate combinations of facial muscle movements, pose substantial challenges for accurate recognition.

Early approaches to CER relied on handcrafted features. For example, geometric features of facial landmarks, such as the distance between eyebrows, the curvature of the mouth corners, and the degree of eye opening, were extracted and fed into traditional classifiers like Support Vector Machines (SVMs) for emotion classification [2, 28]. However, these handcrafted features often fail to capture the nuanced and complex information inherent in compound expressions.

With the advent of deep learning, Convolutional Neural Networks (CNNs) have become widely adopted in CER. CNNs can automatically learn hierarchical features from facial images, significantly improving performance over traditional methods. Pre-trained architectures such as VGG-16 and ResNet have been directly applied to CER tasks [31]. Despite their success, these models still struggle with the high complexity of compound expressions. To address this, more advanced architectures have been proposed. For instance, multi-stream CNN models process different facial regions separately to capture local-global relationships in expressions [32].

Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory networks (LSTMs), have also been employed in CER. Given the temporal nature of facial expressions, RNN-based models can analyze dynamic changes over time, making them particularly effective for video-based CER tasks. These models infer potential emotions by analyzing sequences of facial changes, providing a richer understanding of compound expressions [5].

### 2.2. Large Vision-Language Models (LVLMs)

Large Vision-Language Models (LVLMs) have recently emerged as a transformative force in artificial intelligence [27], excelling in multimodal tasks by effectively integrating visual information (e.g., facial features) and language information (e.g., emotion-related text). LVLMs leverage large-scale data to build a deep understanding of both modalities, offering new opportunities for addressing complex tasks like CER.

Fine-tuning is a common strategy for adapting LVLMs to specific tasks. The most straightforward approach is full fine-tuning, where all parameters of the pre-trained LVLM are adjusted during training on the target task [4]. While this method can achieve high performance, it is computationally expensive and resource-intensive, particularly for large-scale LVLMs.

To mitigate these challenges, partial fine-tuning methods have been developed. One common strategy is to freeze the early layers of the LVLM and fine-tune only the later layers, as early layers typically capture general features while later layers are more task-specific [8]. This approach significantly reduces the number of trainable parameters and computational costs.

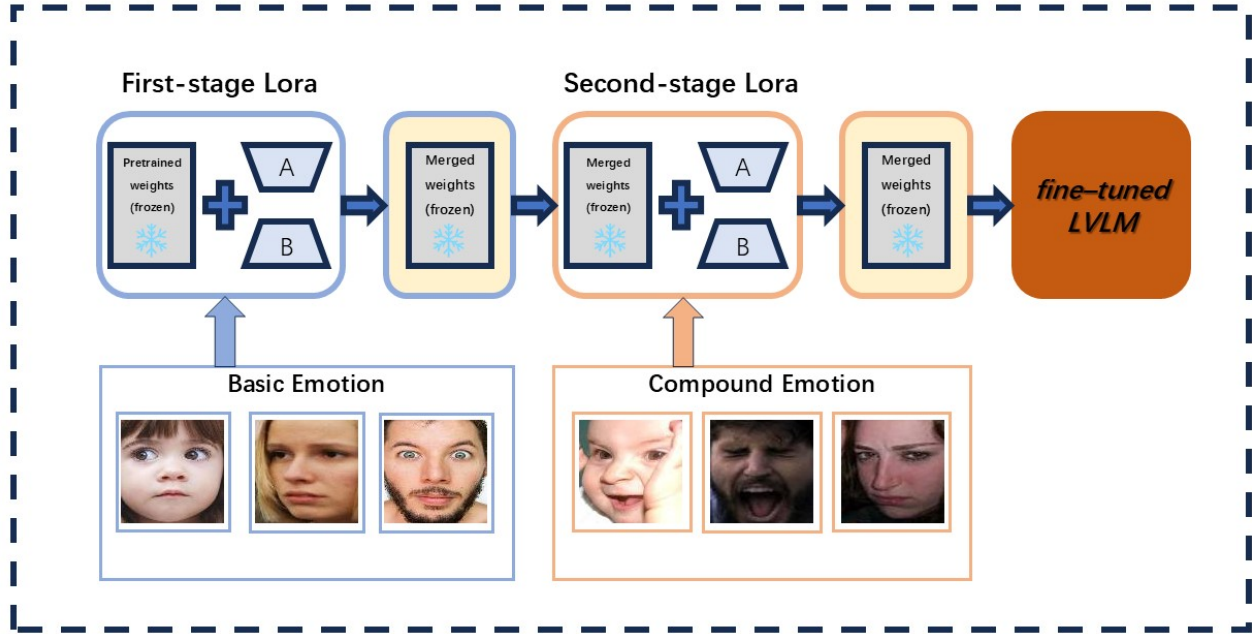


Figure 1. Stage-wise LoRA Fine-tuning Framework for Compound Expression Recognition

Another popular method is adapter-based fine-tuning, which introduces small adapter layers into the pre-trained model. These adapters are trained independently while keeping the original model parameters fixed [7]. This not only reduces computational overhead but also facilitates task switching, as different adapters can be trained for different tasks on the same base model.

In the realm of low-rank adaptation, the LoRA (Low-Rank Adaptation) method has gained significant attention. Unlike full fine-tuning, LoRA updates only a small number of low-rank matrices, drastically reducing memory usage and training time [9]. LoRA has demonstrated strong performance in various fields, including natural language processing and vision-language tasks. By approximating weight updates through low-rank matrices, LoRA enables efficient fine-tuning of LVLMs for specific tasks like CER, making it particularly suitable for resource-constrained settings.

### 3. Method

Our approach for Compound Expression Recognition (CER) leverages the multimodal capabilities of Large Vision-Language Models (LVLMs) through a stage-wise LoRA (Low-Rank Adaptation) fine-tuning strategy and carefully designed context prompts, as shown in Figure 1 and 2. This methodology adapts LVLMs to the complex

task of CER, enabling efficient and accurate recognition of compound facial expressions.

#### 3.1. Stage-wise LoRA Fine-Tuning

##### 3.1.1. Low-Rank Adaptation (LoRA)

LoRA [9] is a parameter-efficient fine-tuning technique designed for large pre-trained models like LVLMs. In a standard neural network layer, the weight matrix  $W$  is high-dimensional, and updating all its parameters during fine-tuning can be computationally expensive and prone to overfitting. LoRA addresses this by introducing two low-rank matrices,  $A$  and  $B$ , to approximate the weight update  $\Delta W$  as:

$$\Delta W \approx BA \quad (1)$$

Here,  $A$  reduces the input dimensionality from  $d$  to  $r$  (where  $r \ll d$ ), and  $B$  maps it back to the original dimension. The output  $y$  of the LoRA-modified layer is computed as:

$$y = W_0x + (BA)x \quad (2)$$

where  $W_0$  is the original weight matrix. The number of trainable parameters is reduced from  $d^2$  to  $2dr$ , significantly lowering computational costs. This reduction is formalized as:

$$N_{original} = d \times d = d^2 \quad (3)$$

$$N_{LoRA} = d \times r + r \times d = 2dr \quad (4)$$

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Objective:</b><br>Your task is to analyze the facial expression of the person in the provided image and classify it into one of the seven predefined categories:<br>Fearfully Surprised: A mix of fear and surprise, characterized by widened eyes, raised eyebrows, and a slightly open mouth.<br>Happily Surprised: A blend of happiness and surprise, featuring a bright smile, raised eyebrows, and wide-open eyes.<br>Sadly Surprised: A combination of sadness and surprise, with downturned lips, raised eyebrows, and a look of shock.<br>Disgustedly Surprised: A mix of disgust and surprise, with a wrinkled nose, raised eyebrows, and a slightly open mouth.<br>Angrily Surprised: A blend of anger and surprise, featuring furrowed brows, wide eyes, and a tense mouth.<br>Sadly Fearful: A combination of sadness and fear, with a worried expression, downturned lips, and wide eyes.<br>Sadly Angry: A mix of sadness and anger, characterized by a furrowed brow, downturned lips, and a tense jawline. |                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Instructions:</b><br><b>Image Analysis:</b><br>Carefully examine the image to identify any visible facial features such as the eyes, eyebrows, mouth, and overall facial tension. Note the position and movement of facial muscles.<br><b>Classification:</b><br>Based on your analysis, select the most appropriate category from the seven provided.<br>If the image does not contain any person, clearly state that no one is present.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | <b>Output Format:</b><br>If no one appears in the image, output:<br>Analysis: [Provide a detailed analysis of the image, noting the absence of any person.]<br>Conclusion: There is no one in the image.<br>If a person is present, output:<br>Analysis: [Provide a detailed analysis of the facial expression, describing the features that led to your conclusion.]<br>Conclusion: The facial expression of the person in the image is '[Selected Category]' . |
| <b>Example Outputs:</b><br>No Person in Image:<br>Analysis: The image depicts a serene landscape with trees and a river. Conclusion: There is no one in the image.<br>Person Present:<br>Analysis: The person in the image has wide-open eyes, raised eyebrows, and a bright smile, indicating a mix of happiness and surprise.<br>Conclusion: The facial expression of the person in the image is 'Happily Surprised'.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | <b>Additional Notes:</b><br>Ensure your analysis is thorough and considers all visible facial features.<br>If the facial expression does not clearly fit into one of the seven categories, provide the closest match and justify your choice.<br>Be mindful of any potential ambiguities in the image and address them in your analysis.                                                                                                                         |

Figure 2. Context Prompt Example on Compound Expressions

This efficiency makes LoRA particularly suitable for fine-tuning large-scale LVLMs in resource-constrained settings.

### 3.1.2. First-stage Fine-Tuning on Basic Emotions

In the first stage, we fine-tune the LVLM using the Basic subset of FER datasets. We initialize LoRA layers within the pre-trained LVLM and freeze the majority of its original parameters, updating only the low-rank matrices  $A$  and  $B$ . This stage focuses on learning fundamental patterns of basic emotions, such as happiness, sadness, and anger, without overfitting to the idiosyncrasies of the pre-trained model. By capturing these foundational features, the model establishes a robust basis for recognizing more complex compound expressions in the subsequent stage.

### 3.1.3. Second-stage Fine-Tuning on Compound Emotions

In the second stage, we fine-tune the model using the Compound subset of FER datasets. We continue to update only the LoRA layers while keeping the original LVLM parameters frozen. This stage focuses on learning the intricate combinations of emotions present in compound expressions, such as "happily surprised" or "sadly angry." The two-stage approach enables the model to incrementally adapt to the increasing complexity of CER, building on its knowledge of basic emotions to accurately recognize compound expressions.

## 3.2. Context Prompt Design

### 3.2.1. Design Principles

Context prompts are designed to provide the LVLM with task-relevant information, guiding its attention to key facial features and their emotional significance [6]. Our prompt architecture follows two core principles:

- **Task Alignment:** Establishes direct mapping between visual features and emotional states through:
  - Seven precisely defined compound emotion categories
  - Visual-semantic anchors (e.g., "wide-open eyes" for surprise)
  - Operational definitions for ambiguous features
- **Process Guidance:** Enforces structured reasoning through:
  - Step-by-step analysis instructions
  - Dual-phase workflow (image analysis → classification)
  - Edge case handling protocols

### 3.2.2. Structural Components

The prompt comprises five interconnected modules:

- **Task Specification Module:**
  - Defines 7 emotion categories with distinguishing features

- Provides visual prototypes (e.g., Sadly Angry:
  - \* Furrowed brow (anger indicator)
  - \* Downturned lips (sadness marker)
  - \* Tense jawline (combined affect)

- **Analytical Framework:**

- Feature Extraction:
  - \* Spatial analysis: eyes, eyebrows, mouth, facial tension
  - \* Temporal analysis: muscle movement dynamics
- Decision Protocol:
  - \* Evidence-based category selection
  - \* Empty image detection routine

- **Output Schema:**

- Structured templates ensure consistency:
  - \* Negative case: Analysis: [...] Conclusion: No person
  - \* Positive case: Analysis: [...] Conclusion: [Category]

- **Exemplar Bank:**

- Paired demonstrations:
  - \* Negative example: Landscape image analysis
  - \* Positive example: Happily Surprised classification

- **Quality Assurance:**

- Completeness verifiers: Feature checks
- Uncertainty handlers: Closest-match justification
- Ambiguity resolvers: Conflict acknowledgement protocols

This structured approach enhances the model’s ability to accurately recognize compound expressions and provide interpretable results.

## 4. Experiments

### 4.1. Datasets

#### 4.1.1. RAF-DB

We utilize the Basic and Compound subsets from the RAF-DB (Real-world Affective Faces Database) [26]. The Basic subset within RAF-DB contains a substantial number of facial images labeled with basic emotions, including happiness, sadness, anger, surprise, fear, and disgust. These serve as a fundamental resource for the model to learn basic facial expression patterns. The Compound subset, on the other hand, is composed of images with more complex, compound emotional expressions. Each image in this subset is precisely annotated with the specific combination of emotions present, equipping the model to learn the subtleties of

compound expressions.

#### 4.1.2. Aff-Wild2

Aff-Wild2 is a large-scale in-the-wild dataset designed for affective computing tasks [13], including emotion recognition and valence-arousal estimation. It contains over 1,000 videos with more than 2.5 million frames, annotated for 7 basic emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral) and 6 compound emotions. The dataset is highly diverse, covering a wide range of ages, ethnicities, and lighting conditions, making it suitable for training robust models for real-world applications. We use Aff-Wild2 in conjunction with RAF-DB for training, leveraging its large-scale and diverse annotations to enhance the model’s generalization ability.

#### 4.1.3. C-EXPR-DB

C-EXPR-DB stands as the largest and most diverse in-the-wild audiovisual database to date [12]. It encompasses 400 videos, totaling around 200,000 frames, meticulously annotated for 12 compound expressions and various affective states. Additionally, C-EXPR-DB includes annotations for continuous valence-arousal dimensions [-1, 1], speech detection, facial landmarks and bounding boxes, 17 action units (facial muscle activation), and facial attributes. In the Compound Expression (CE) Recognition Challenge, a total of 56 unlabeled videos were selected, covering 7 types of compound expressions. The extracted video tags consist of seven compound expressions: Fearfully Surprised, Happily Surprised, Sadly Surprised, Disgustedly Surprised, Angrily Surprised, Sadly Fearful, and Sadly Angry. C-EXPR-DB serves as the final test set for evaluating the model’s performance in a real-world competition setting.

## 4.2. Implement Details

### 4.2.1. Model Selection

We selected Qwen-vl [1] as the base model for our experiments. This model has demonstrated strong performance in various multimodal tasks, making it a suitable choice for compound expression recognition. We initialized the model with pre-trained weights on large-scale vision language datasets, which provides the model with a good starting point for learning facial expression-related features.

### 4.2.2. LoRA Configuration

In the stage-wise LoRA fine-tuning process, we configured the LoRA layers. For the first-stage fine-tuning on the Basic subset, we set the rank of the low-rank matrices to 16. This value was chosen based on preliminary experiments to balance the reduction of trainable parameters and the model’s representational ability. The learning rate was set to 0.0001. During the second-stage fine-tuning on the Compound subset, the rank of the low-rank matrices was adjusted to 8, and the learning rate was set to 0.0001. These adjustments were

made to better adapt the model to the increasing complexity of compound expressions.

#### 4.2.3. Context Prompt Generation

The context prompts were generated using a rule-based approach. We defined a set of templates related to facial expression analysis. For example, for an image, the prompt might start with "Analyze the following facial image to identify compound emotions. Look for signs of key facial features related to basic emotions and determine how they combine." The generation process was automated, and the prompts were carefully designed to provide relevant information to the model without over-specifying the answer.

#### 4.2.4. Training and Validation

The training was carried out using 8 RTX3090 GPUs to expedite the process. The batch size was set to 1 for both the first-stage and second-stage fine-tuning. The number of epochs for the first-stage fine-tuning was 20, and for the second-stage fine-tuning was 10.

We train our model on a combination of the Aff-Wild2 and RAF-DB datasets, leveraging the large-scale and diverse annotations from Aff-Wild2 to enhance the model's generalization ability. The model is then validated on the RAF-DB dataset to ensure its performance on both basic and compound emotion recognition tasks. Finally, the model is evaluated on the C-EXPR-DB dataset, which serves as the final test set for the Compound Expression Recognition Challenge. This approach ensures that the model is robust and capable of handling real-world variability in facial expression recognition.

## 5. Results

### 5.1. RAF-DB Basic Emotion Recognition

To adapt to the real-world image environment of the C-EXPR-DB dataset, we conducted experiments on both **aligned images** and **original images** for RAF-DB Basic Emotion Recognition. The results are presented in Table 1.

Table 1. Accuracy for RAF-DB Basic Emotion Recognition

| Emotion        | Aligned (%)  | Original (%) |
|----------------|--------------|--------------|
| Sadness        | 88.12        | 86.89        |
| Surprise       | 94.45        | 92.58        |
| Happiness      | 96.50        | 94.39        |
| Disgust        | 68.25        | 64.88        |
| Anger          | 83.30        | 80.78        |
| Fear           | 73.60        | 69.22        |
| Neutral        | 92.20        | 90.65        |
| <b>Overall</b> | <b>89.78</b> | <b>87.84</b> |

As shown in Table 1, the model achieves higher accuracy on aligned images compared to original images across

all emotion categories. For example, the accuracy for **Happiness** increases from 94.39% on original images to 96.50% on aligned images. Similarly, the overall accuracy improves from 87.84% to 89.78%. This indicates that aligned images, with their standardized facial features, provide a more consistent and reliable input for emotion recognition. However, the performance on original images remains strong, demonstrating the model's robustness in real-world scenarios.

### 5.2. RAF-DB Compound Emotion Recognition

For RAF-DB Compound Emotion Recognition, we also conducted experiments on both **aligned images** and **original images**. The results are presented in Table 2.

Table 2. Accuracy for RAF-DB Compound Emotion Recognition

| Emotion               | Aligned (%)  | Original (%) |
|-----------------------|--------------|--------------|
| Happily Surprised     | 90.25        | 88.52        |
| Sadly Disgusted       | 75.30        | 71.70        |
| Happily Disgusted     | 80.45        | 78.09        |
| Fearfully Angry       | 77.80        | 73.64        |
| Angrily Disgusted     | 94.10        | 92.18        |
| Angrily Surprised     | 50.20        | 41.58        |
| Sadly Surprised       | 52.40        | 43.33        |
| Fearfully Surprised   | 80.60        | 78.10        |
| Disgustedly Surprised | 55.80        | 45.71        |
| Sadly Fearful         | 60.30        | 55.45        |
| Sadly Angry           | 55.20        | 46.36        |
| <b>Overall</b>        | <b>78.50</b> | <b>74.39</b> |

As shown in Table 2, the model achieves higher accuracy on aligned images compared to original images for all compound emotion categories. For example, the accuracy for **Angrily Disgusted** increases from 92.18% on original images to 94.10% on aligned images. Similarly, the overall accuracy improves from 74.39% to 78.50%. This suggests that aligned images provide a more consistent representation of compound emotions, which are inherently more complex and challenging to recognize. However, the performance on original images remains competitive, highlighting the model's ability to handle real-world variability.

### 5.3. C-EXPR-DB Compound Emotion Recognition

For C-EXPR-DB Compound Emotion Recognition, we evaluated our model's performance. The results are presented in Table 3.

Table 3. Results for C-EXPR-DB Compound Emotion Recognition

| Team                   | Rank     | F1-score      |
|------------------------|----------|---------------|
| <b>USTC-IAT-United</b> | <b>1</b> | <b>0.5723</b> |
| CAS-MAIS               | 2        | 0.1162        |

As shown in Table 3, our model achieved an F1-score of 57.23% on the C-EXPR-DB dataset for compound emotion recognition and ranked first in the 8th ABAW Compound Expression Recognition Challenge. These results further demonstrate the effectiveness of our proposed approach in handling the complex and diverse compound emotions present in the C-EXPR-DB dataset. Although the dataset poses unique challenges such as a wider variety of real-world scenarios and more subtle emotional cues, our model still shows strong performance, indicating its potential for practical applications in real-life emotion analysis tasks.

## 6. Conclusion

The results demonstrate that aligned images consistently outperform original images in both basic and compound emotion recognition tasks. This is expected, as aligned images provide standardized facial features, reducing variability and improving model performance. These findings are derived from our experimental results on the RAF-DB dataset, where facial alignment preprocessing significantly enhanced recognition accuracy by normalizing facial geometry variations. However, the strong performance on original images underscores the model's robustness and generalizability to real-world scenarios, as evidenced by our top-ranking results with an F1 score of 0.5723 in the 8th ABAW Compound Expression Recognition Challenge using the C-EXPR-DB dataset. This competition achievement validates our model's effectiveness in handling unconstrained facial expressions without alignment preprocessing. Future work could focus on further bridging the gap between aligned and original image performance, potentially through domain adaptation techniques or more sophisticated data augmentation strategies. This is an important direction highlighted by our competition experience with real-world expression recognition challenges.

## Acknowledgements

This work was supported by the Natural Science Foundation of China (62276242), National Aviation Science Foundation (2022Z071078001), Dreams Foundation of Jianghuai Advance Technology Center (2023-ZM01Z001).

## References

- [1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 5
- [2] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(1):25–36, 1999. 2
- [3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. and... Dhariwal, and D. Amodei. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1
- [4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019. 2
- [5] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999. 2
- [6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, 2024. Association for Computational Linguistics. 4
- [7] N. Houlsby, A. Giurgiu, S. Jastrzebski, Ł. Morrone, D. De Laroussilhe, A. and... Gesmundo, and F. Zenke. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*, 2019. 3
- [8] A. G. Howard and S. Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018. 2
- [9] E. Hu, Y. Shen, P. Wallis, Z. Allen Zhu, Y. Li, S. Wang, and Z. Ma. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3
- [10] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 1
- [11] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023. 1
- [12] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023. 5
- [13] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1, 5
- [14] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 1
- [15] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 1
- [16] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800, .
- [17] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Irene Kotsia, UK Cogitat, Eric Granger, Marco Pedersoli, Simon Bacon, Alice Baird, Chunchang Shao, et al. Advancements

- in affective and behavior analysis: The 8th abaw workshop and competition. . [1](#)
- [18] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. [1](#)
- [19] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019.
- [20] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. [1](#)
- [21] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2023. [1](#)
- [22] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for multi-task learning of classification tasks: a large-scale study on faces & beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2813–2821, 2024. [1](#)
- [23] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Alice Baird, Chris Gagne, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4587–4598, 2024. [1](#), [2](#)
- [24] Dimitrios Kollias, Stefanos Zafeiriou, Irene Kotsia, Abhinav Dhall, Shreya Ghosh, Chunchang Shao, and Guanyu Hu. 7th abaw competition: Multi-task learning and compound expression recognition. *arXiv preprint arXiv:2407.03835*, 2024. [2](#)
- [25] Dimitrios Kollias, Panagiotis Tzirakis, Alan S. Cowen, Stefanos Zafeiriou, Irene Kotsia, Eric Granger, Marco Pedersoli, Simon L. Bacon, Alice Baird, Chris Gagne, Chunchang Shao, Guanyu Hu, Soufiane Belharbi, and Muhammad Haseeb Aslam. Advancements in Affective and Behavior Analysis: The 8th ABAW Workshop and Competition. 2025. [1](#)
- [26] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. [5](#)
- [27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. [2](#)
- [28] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000. [1](#), [2](#)
- [29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. [1](#)
- [30] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal ‘in-the-wild’ challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. [1](#)
- [31] X. Zhang and Q. Ji. Spatio-temporal modeling of facial expressions using local binary patterns and multi-kernel learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(6):1543 – 1554, 2010. [2](#)
- [32] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. [2](#)