

Supplementary Materials: FineCLIPER: Multi-modal Fine-grained CLIP for Dynamic Facial Expression Recognition with Adapter

Anonymous Authors

1 Introduction

The content of our supplementary material is organized as follows:

- 1) In Sec. 2, we present the information of demo, code, and variants illustration for FineCLIPER;
- 2) In Sec. 3, we analyze two competitive baseline models;
- 3) In Sec. 4, we further analyze the proposed adaptive weighting strategy;
- 4) In Sec. 5, we present detailed information regarding fine-grained text descriptions of facial action movements.

2 Additional Information

Demo. We presented a simple demo in the compressed file to facilitate a better understanding of our FineCLIPER.

Code. The main anonymous code is available at <https://anonymous.4open.science/r/FineCLIPER>. The full version of the code, model weights, and data will be publicly available upon the paper notification.

FineCLIPER Variants. 1) FineCLIPER: The variant utilizes only low-semantic level video frames along with PN (Postive-Negative) descriptors for label augmentation; 2) FineCLIPER*: Building upon FineCLIPER, this variant incorporates middle-semantic level face parsing and landmarks; 3) FineCLIPER[†]: Extending FineCLIPER, this variant directly integrates high-semantic level fine-grained descriptions of facial action changes; 4) FineCLIPER^{*†}: Expanding on FineCLIPER, this variant includes both middle-semantic level and high-semantic level information.

3 Competitive Baseline Analysis

S2D [1] achieved notable results with minimal tunable parameters on the ViT-B/16 backbone. However, it is noteworthy that it first undergoes pre-training on the Static Facial Expression Recognition (SFER) dataset, specifically AffectNet-7 [5] (consisting of 283,901 training samples) for 100 epochs, before fine-tuning on the DFER dataset. This pre-training step significantly contributes to its performance;

A³lign-DFER [6], as the latest CLIP-based DFER model, predominantly relies on the CLIP-ViT-L/14 backbone to further empower DFER from an alignment perspective. The training process is delineated into three stages spanning a total of 100 epochs. Regrettably, pertinent information regarding tunable parameters was not found within its paper.

In contrast, our FineCLIPER model employs the CLIP-ViT-B/16 backbone and undergoes training solely on the DFER dataset for 30 epochs, achieving state-of-the-art performance with 13-20M tunable parameters in both supervised and zero-shot settings.

4 Ablation of Adaptive Weighting

Due to the inherent correlation between the expanded multimodal data, *i.e.*, face parsing, landmarks, and fine-grained text, with videos, this work endeavors to explore the feasibility of further modeling

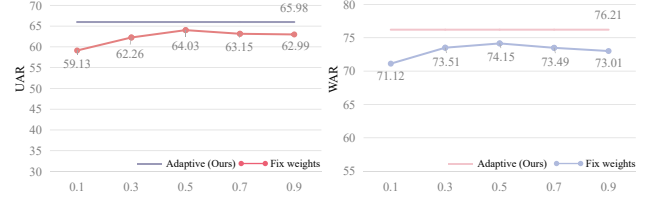


Figure 1: Comparison between our adaptive weighting strategy and fixed weights on the DFEW dataset, where the x-axis represents the weights of video features.

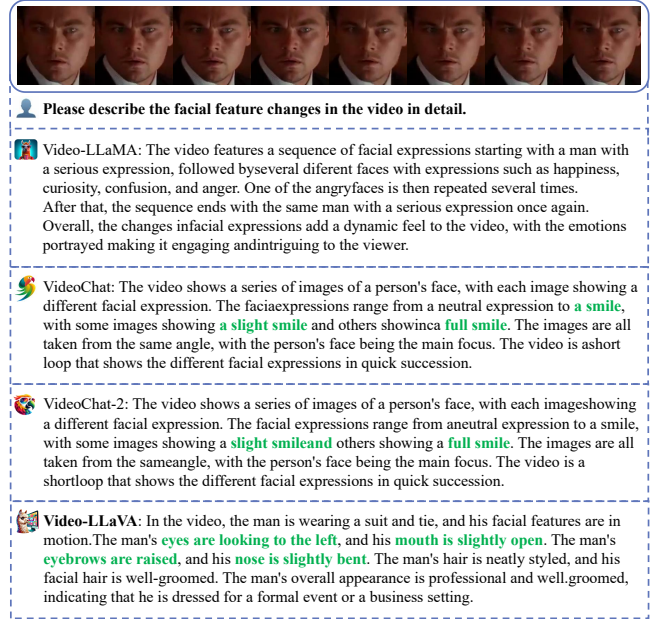


Figure 2: Comparison of MLLM-generated captions for video. Facial expressions related are highlighted in green.

faces using multi-modal data. In our proposed adaptive weighting algorithm, we determine the fusion of features and the weighting of the loss function adaptively by computing the similarity between multi-modal features and label features. To further validate the superiority of this strategy, we fix the weight of the video feature, w_v , ranging from 0.1 to 0.9, while evenly distributing the remaining weight among the other three modal features, *i.e.*, $1 - w_v$. As illustrated in Fig. 1, our proposed adaptive weighting strategy exhibits greater stability and effectiveness compared to fixed weights.

Certainly, the fusion of features can be further optimized. The simplicity of our fusion strategy in this study serves to further substantiate the feasibility and potential of leveraging multi-modal

Instruction:
You are a facial expression analyst. Identify emotions by analyzing the detailed movements of various facial features (mouth, chin, eyes, etc.) and implied actions (breathing rate, gaze direction) across a sequence. Provide a detailed description of the movements of various facial features and implied actions concisely.

Example:
[Input Video]: Initial neutrality transitions to anger; mouth corners downturn, chin protrudes, eyes narrow, gaze intensifies, and breathing appears heavier.
Answer: Downward mouth, protruding chin, narrowed eyes, intense gaze, and heavy breathing collectively signify anger.

Requirements:

1. Format: [Detailed feature and action analysis].
2. Focus on detailed progression of multiple facial parts and implied actions to emotions.
3. Role: analyze nuanced facial and action cues.
4. Use the human-annotated video.
5. Don't include emotional words.
6. Max 100 tokens.

For the given video, human-annotation: {Label}

Figure 3: Text Prompt Demonstration

data for DFER. In future works, we intend to delve deeper into the potential of feature fusion.

5 Fine-grained Text Generation

MLLM Selection. Considering resource consumption, we initially evaluated several open-source Multi-modal Large Language Models (MLLMs) capable of processing videos, namely Video-LLaMA [8], VideoChat [2], VideoChat-2 [3], and Video-LLaVA [4]. To expedite the assessment of their ability to comprehend facial videos, we employed a simple prompt at this stage, *i.e.*, "Please describe the facial feature changes in the video in detail". As depicted in Fig. 2, we highlight in green the descriptions of facial features outputted by the four MLLMs. It is evident that Video-LLaVA, compared to the other three, more accurately captures facial feature information. Consequently, we adopt Video-LLaVA as our fine-grained text generation model. Subsequent sections will elaborate on the detailed text prompt and refinement process for fine-grained text generation.

Prompt Design. With the advancement of large language models, the significance of prompt engineering has become increasingly apparent. Well-crafted prompts can significantly enhance a model's ability to generate responses tailored to specific tasks. As illustrated in Fig. 3, in addition to explicit instructions, corresponding examples are provided for the model to reference and learn from. Furthermore, to further standardize the model's responses, six requirements are delineated. Finally, considering that the model may describe actions based on its analysis of facial expressions, ground truth labels for each video are also provided for the model's reference.

Text Refinement. Text refinement plays a pivotal role in our proposed FineCLIPER framework. Specifically, we identify two categories of low-quality text: 1) Directly expressing emotions. For example, stating "The man in the video wears a sad expression..." This can lead to data leakage during the training process. 2) Indirectly implying emotions. For example, stating "The man's mouth is slightly ajar, showing his teeth, and his eyes are narrowed, suggesting a feeling of joy or amusement." Despite not explicitly containing label information, such descriptions still pose a risk of potential data leakage.

To this end, we introduce a two-stage heuristic process for text refinement in this study, as outlined by [7], which comprises text cleaning and counterfactual verification as illustrated in the main content. Specifically, manual inspection involves the participation of numerous master's and undergraduate students with backgrounds in psychology or computer science.

Generated Instances. As demonstrated in Fig. 4, despite the careful design of the prompt in Fig. 3, emphasizing "Don't include emotional words," the generated text still contains several direct or indirect emotional expressions, as highlighted in **red**. Subsequently, through the implementation of the two-stage text refinement process, the refined text predominantly encompasses facial features and implied actions, as highlighted in **bold**. Such refined fine-grained text significantly enhances and strengthens facial modeling from a high semantic level. All the fine-grained descriptions, along with the face parsing and landmarks data will be released after the paper notification.

References

- [1] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. 2023. From Static to Dynamic: Adapting Landmark-Aware Image Models for Facial Expression Recognition in Videos. *arXiv:2312.05447 [cs.CV]*
- [2] Kunchang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. VideoChat: Chat-Centric Video Understanding. *arXiv preprint arXiv:2305.06355 (2023)*.
- [3] Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. *arXiv:2311.17005 [cs.CV]*
- [4] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv:2311.10122 [cs.CV]*
- [5] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* 10, 1 (Jan. 2019), 18–31. <https://doi.org/10.1109/taffc.2017.2740923>
- [6] Zeng Tao, Yan Wang, Junxiong Lin, Haoran Wang, Xinji Mai, Jiawen Yu, Xuan Tong, Ziheng Zhou, Shaoqi Yan, Qing Zhao, Liyuan Han, and Wenqiang Zhang. 2024. A³lign-DFER: Pioneering Comprehensive Dynamic Affective Alignment for Dynamic Facial Expression Recognition with CLIP. *arXiv:2403.04294 [cs.CV]*
- [7] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2023. When Urban Region Profiling Meets Large Language Models. *arXiv:2310.18340 [cs.CL]*
- [8] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858 (2023)*. <https://arxiv.org/abs/2306.02858>

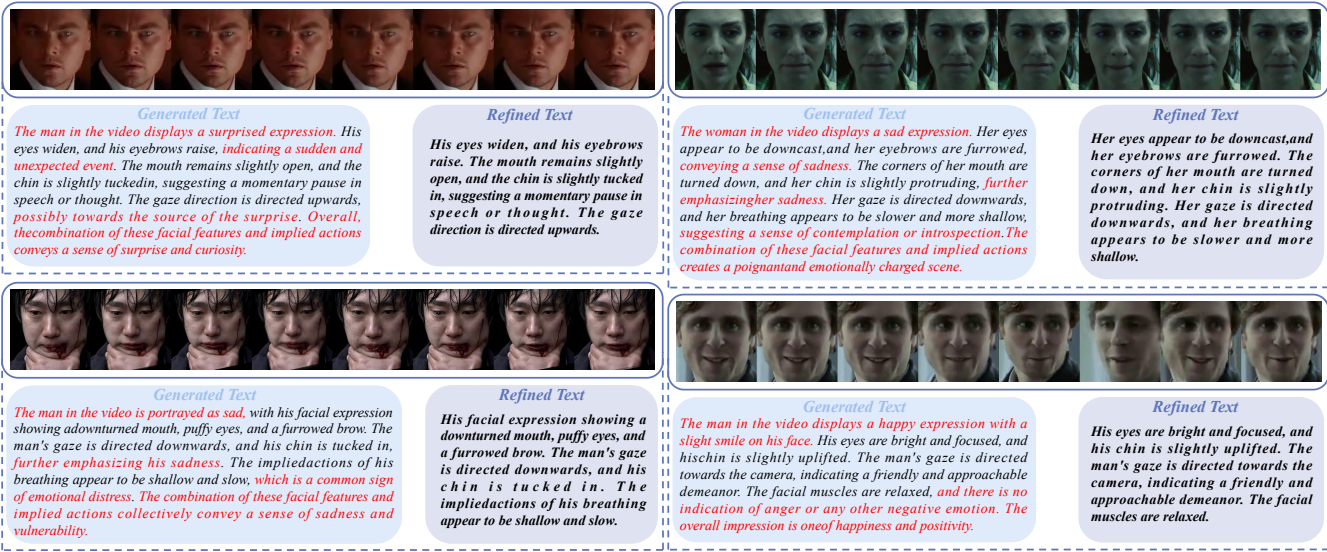


Figure 4: Examples of the generated text and the refined text.