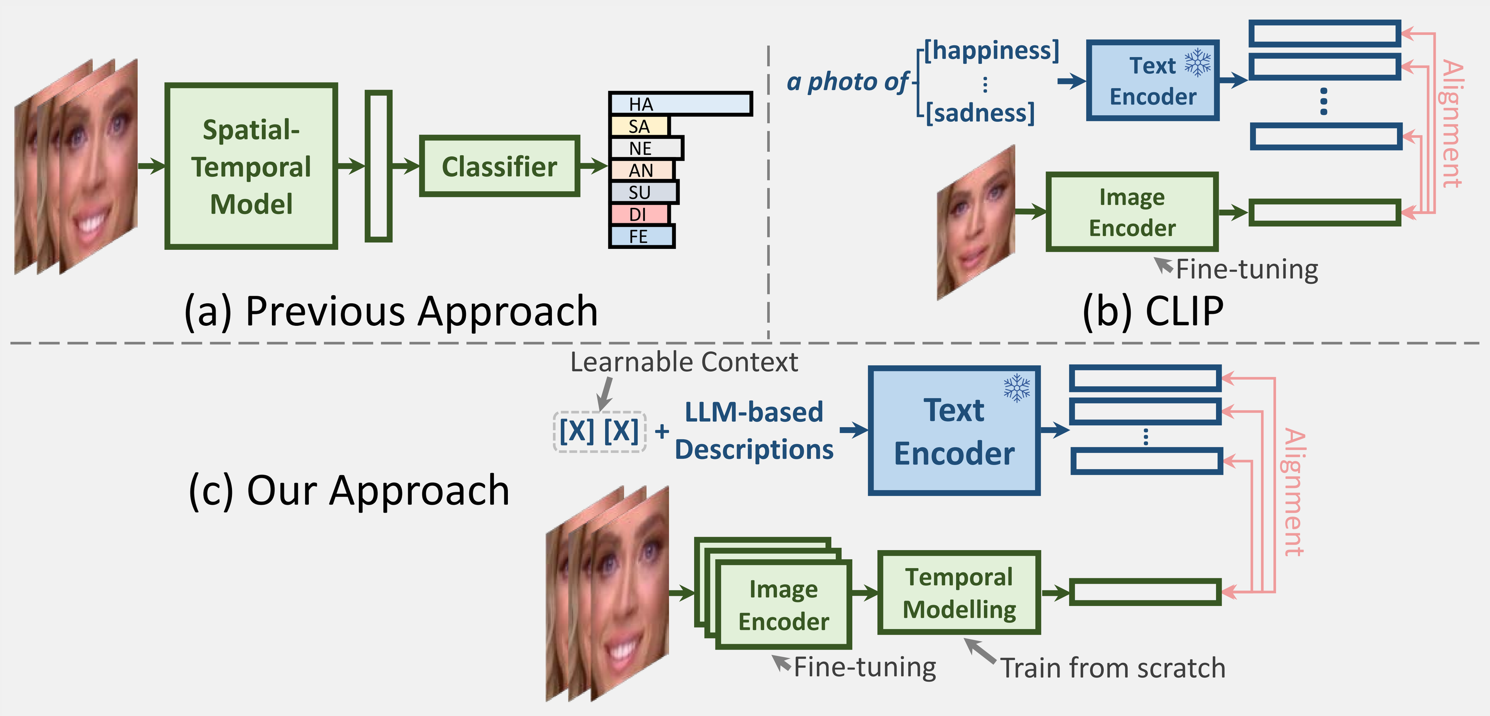
# Prompting Visual-Language Models for Dynamic Facial Expression Recognition


Zengqun Zhao Ioannis Patras

Queen Mary University of London


BMVC 2023

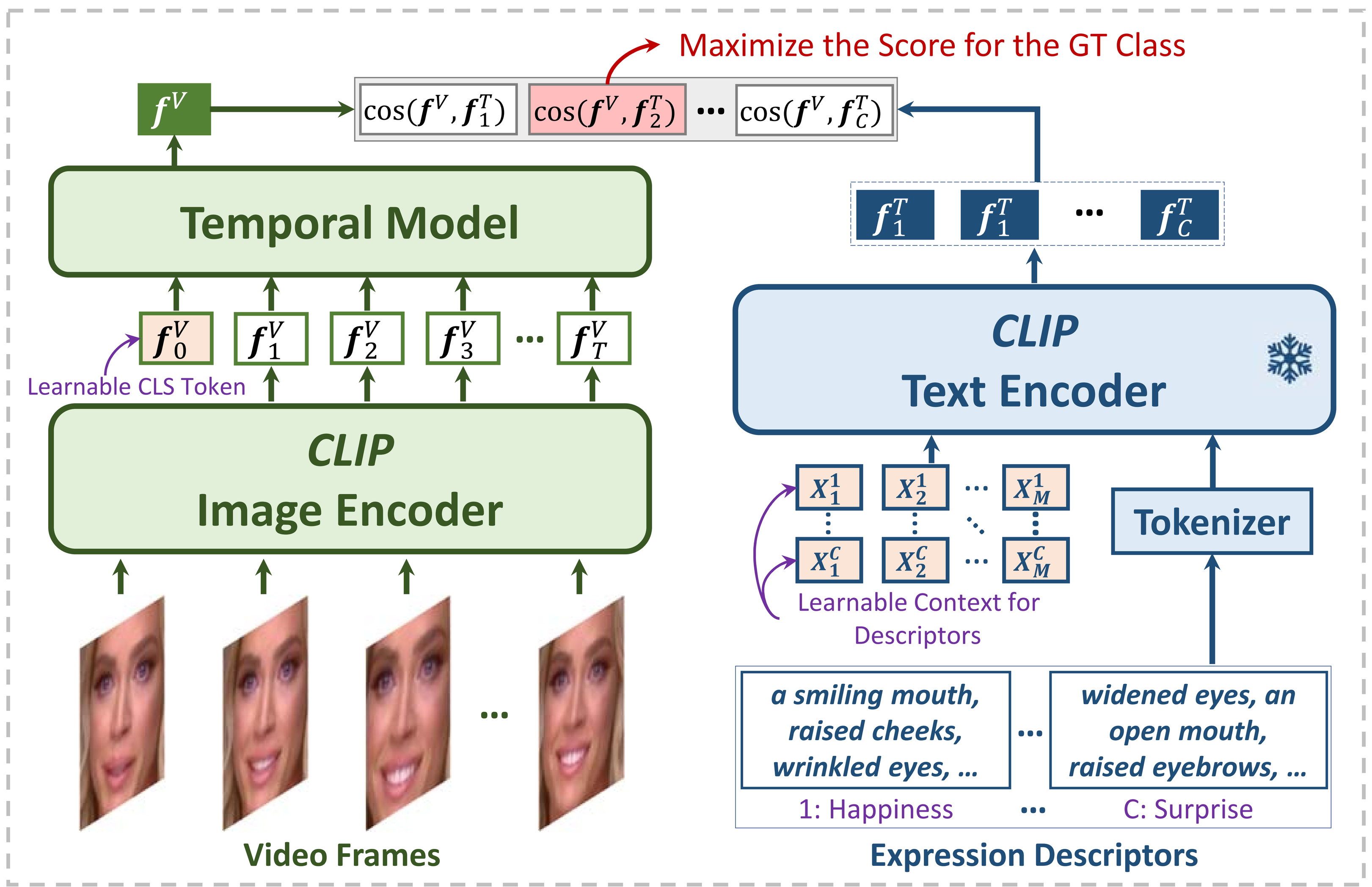
## Overview




We propose a novel visual-language model called DFER-CLIP, based on the CLIP model and designed for in-the-wild Dynamic Facial Expression Recognition (DFER). The DFER-CLIP:

- **Temporal feature learning**: learns spatial as well as temporal facial expression features by re-fining a CLIP image-encoder and training a temporal model.
- **Text description**: trains in a supervised manner with text descriptions, capturing facial behaviour, instead of class name.
- **Learnable context**: with a learnable prompt for descriptors of each class to learn relevant context information for each expression during training.

## Method




- **Visual part**: the frame-level features are first learnt by a shared CLIP visual encoder. Then all of the frame-level features along with an additional learnable class token will feed into the temporal model, in which the learnable position embedding is added to encode the temporal position.
- **Textual part**: we utilize descriptions related to facial behaviour instead of class names for the text encoder. Furthermore, we adopt the learnable prompt as a context for descriptors of each class, which does not require experts to design context words and allows the model to learn relevant context information for each expression during training.

## LLM-based Descriptions Building

- The CLIP text encoder learns semantic information from natural language text, we propose taking the facial action description as the input for the text encoder.
- We prompt a large language model such as ChatGPT to automatically generate descriptions based on contextual information, instead of manually designing.
- We prompt the language model with the input:
  *Q: What are useful visual features for the facial expression of {class name}?*
  *A: Some useful visual features for facial expressions of {class name} include: ...*

## Comparison with SOTA

| Methods | DFEW | | FERV39k | | MAFW | |
|---|---|---|---|---|---|---|
| | UAR | WAR | UAR | WAR | UAR | WAR |
| Former-DFER [1] [MM'21] | 53.69 | 65.70 | 37.20 | 46.85 | 31.16 | 43.27 |
| DPCNet [2] [MM'22] | 57.11 | 66.32 | - | - | - | - |
| T-ESFL [3] [MM'22] | - | - | - | - | 33.28 | 48.18 |
| EST [4] [PR'23] | 53.94 | 65.85 | - | - | - | - |
| IAL [5] [AAAI'23] | 55.71 | 69.24 | 35.82 | 48.54 | - | - |
| CLIPER [6] [arXiv'23] | 57.56 | 70.84 | 41.23 | 51.34 | - | - |
| M3DFEL [7] [CVPR'23] | 56.10 | 69.25 | 35.94 | 47.67 | - | - |
| AEN [8] [CVPRW'23] | 56.66 | 69.37 | 38.18 | 47.88 | - | - |
| DFER-CLIP (Ours) | **59.61** | **71.25** | **41.27** | **51.65** | **39.89** | **52.55** |

## Ablation Analysis

Table 1. Evaluation of the learnable context prompt numbers & the temporal model depths.

| Number of the Context Prompts | Depth of the Temporal Model | DFEW | | FERV39k | | MAFW | |
|---|---|---|---|---|---|---|---|
| | | UAR | WAR | UAR | WAR | UAR | WAR |
| 4 | ✗ | 56.91 | **69.01** | 40.26 | **50.96** | 38.03 | 50.62 |
| 8 | ✗ | **57.39** | 69.00 | **40.64** | 50.92 | **38.51** | **50.91** |
| 16 | ✗ | 57.32 | 68.96 | 40.22 | 50.64 | 37.98 | 50.40 |
| 8 | 1 | **59.61** | **71.25** | **41.27** | **51.65** | **39.89** | **52.55** |
| 8 | 2 | 58.87 | 70.92 | 40.41 | 51.08 | 39.13 | 52.10 |
| 8 | 3 | 58.64 | 70.80 | 40.35 | 50.98 | 38.90 | 51.86 |

- By adopting the temporal model, the UAR performance can be improved by 2.22%, and 1.38%, and WAR performance can be improved by 2.25%, 0.73%, and 1.64% on DFER, FERV39k, and MAFW datasets, respectively.

Table 2. Evaluation of different training strategies. TM stands for the temporal model.

| Strategies | | DFEW | | FERV39k | | MAFW | |
|---|---|---|---|---|---|---|---|
| | | UAR | WAR | UAR | WAR | UAR | WAR |
| Classifier-based | Linear Probe | 45.46 | 57.40 | 32.47 | 43.72 | 30.74 | 42.95 |
| | Fully Fine-Tuning (w/o TM) | 55.70 | 68.41 | 39.64 | 50.77 | 37.53 | 50.48 |
| | Fully Fine-Tuning (w/ TM) | 58.28 | 70.27 | 40.55 | 51.22 | 38.39 | 50.92 |
| Text-based (Classifier-free) | Zero-shot CLIP | 23.34 | 20.07 | 20.99 | 17.09 | 18.42 | 19.16 |
| | Zero-shot FaRL | 23.14 | 31.54 | 21.67 | 25.65 | 14.18 | 11.78 |
| | CoOp | 44.98 | 56.68 | 31.72 | 42.55 | 30.79 | 42.77 |
| | Co-CoOp | 46.80 | 57.52 | 32.91 | 44.25 | 30.81 | 43.23 |
| | DFER-CLIP (w/o TM) (Ours) | 57.39 | 69.00 | 40.64 | 50.92 | 38.51 | 50.91 |
| | DFER-CLIP (w TM) (Ours) | **59.61** | **71.25** | **41.27** | **51.65** | **39.89** | **52.55** |

- Our method outperforms Fully Fine-tuning in UAR by 3.91%, 1.63%, and 2.36%, and in WAR by 2.84%, 0.88%, and 2.07% on DFER, FERV39k, and MAFW datasets, respectively. Even without the temporal model, our method is better than all the classifier-based methods.

Table 3. Evaluation of different prompts.

| Prompts | | DFEW | | FERV39k | | MAFW | |
|---|---|---|---|---|---|---|---|
| | | UAR | WAR | UAR | WAR | UAR | WAR |
| w/o TM | a photo of [Class] | 56.21 | 68.44 | 39.44 | 49.94 | 37.91 | 50.87 |
| | an expression of [Class] | 56.16 | 68.73 | 39.28 | 50.41 | 37.71 | 51.08 |
| | [Learnable Prompt] [Class] | 57.37 | 68.86 | 40.42 | 50.50 | 38.01 | 50.81 |
| | [Learnable Prompt] [Descriptors] | 57.39 | 69.00 | 40.64 | 50.92 | 38.51 | 50.91 |
| w/ TM | [Learnable Prompt] [Class] | 58.28 | 70.29 | 40.60 | 51.18 | 39.64 | 51.21 |
| | [Learnable Prompt] [Descriptors] | **59.61** | **71.25** | **41.27** | **51.65** | **39.89** | **52.55** |

- Our method outperforms manually designed prompts on both DFEW and FERV39k datasets. Furthermore, our method outperforms the prompt of the class name with the learnable context approach, which indicates the effectiveness of using descriptions.

## Acknowledgments


Zengqun Zhao is funded by Queen Mary Principal's PhD Studentships, and this work is supported by the EU H2020 AI4Media No. 951911 project.


## References


[1] Z. Zhao *et al.*, "Former-dfer: Dynamic facial expression recognition transformer," in *MM*, 2021.
[2] Y. Wang *et al.*, "Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos," in *MM*, 2022.
[3] Y. Liu *et al.*, "Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild," in *MM*, 2022.
[4] Y. Liu *et al.*, "Expression snippet transformer for robust video-based facial expression recognition," *Pattern Recognition*, 2023.
[5] H. Li *et al.*, "Intensity-aware loss for dynamic facial expression recognition in the wild," in *AAAI*, 2023.
[6] H. Li *et al.*, "Cliper: A unified vision-language framework for in-the-wild facial expression recognition," *arXiv*, 2023.
[7] H. Wang *et al.*, "Rethinking the learning paradigm for dynamic facial expression recognition," in *CVPR*, 2023.
[8] B. Lee *et al.*, "Frame level emotion guided dynamic facial expression recognition with emotion grouping," in *CVPRW*, 2023.


Code is available at `https://github.com/zengqunzhao/DFER-CLIP`


Feel free to contact us `{zengqun.zhao,i.patras}@qmul.ac.uk`