

Zero-Shot Multimodal Compound Expression Recognition Approach using Off-the-Shelf Large Visual-Language Models

Elena Ryumina

St. Petersburg Federal Research Center
of the Russian Academy of Sciences
St. Petersburg, Russia

ryumina.e@iiias.spb.su

Alexandr Axyonov

St. Petersburg Federal Research Center
of the Russian Academy of Sciences
St. Petersburg, Russia

axyonov.a@iiias.spb.su

Mikhail Dolgushin

St. Petersburg Federal Research Center
of the Russian Academy of Sciences
St. Petersburg, Russia

dolgushin.m@iiias.spb.su

Maxim Markitantov

St. Petersburg Federal Research Center
of the Russian Academy of Sciences
St. Petersburg, Russia

markitantov.m@iiias.spb.su

Dmitry Ryumin

St. Petersburg Federal Research Center
of the Russian Academy of Sciences
St. Petersburg, Russia

ryumin.d@iiias.spb.su

Alexey Karpov

ITMO University
St. Petersburg, Russia

karpov@iiias.spb.su

Abstract

Compound Expression Recognition (CER), a subfield of affective computing, aims to detect complex emotional states formed by combinations of basic emotions. In this work, we present a novel zero-shot multimodal approach for CER that combines six heterogeneous modalities into a single pipeline: static and dynamic facial expressions, scene and label matching, scene context, audio, and text. Unlike previous approaches relying on task-specific training data, our approach uses zero-shot components, including Contrastive Language-Image Pretraining (CLIP)-based label matching and Qwen-VL for semantic scene understanding. We further introduce a Multi-Head Probability Fusion (MHPF) module that dynamically weights modality-specific predictions, followed by basic-to-compound emotion conversion that uses Pair-wise Probability Aggregation (PPA) or Pair-wise Feature Similarity Aggregation (PFSA) methods to produce interpretable compound emotion outputs. Evaluated under multi-corpus training, the proposed approach achieves macro-F1 scores of 46.95% on AffWild2, 49.02% on Acted Facial Expressions in The Wild (AFEW), and 34.85% on C-EXPR-DB via zero-shot testing, comparable to supervised approaches trained on target data. Thus, our approach effectively captures Compound Expressions (CE) without domain adaptation.

The source code is publicly available at https://github.com/SMIL-SPCRAS/ICCVW_25.

1. Introduction

Compound Expression Recognition (CER), a subfield of affective computing, represents an emerging challenge in intelligent human-computer interaction and multimodal interface design. Unlike traditional emotion recognition systems that focus on discrete, basic emotional states – such as Fear, Happiness, Sadness, Anger, Surprise, and Disgust – CER aims at detecting complex, compound emotional expressions that naturally occur in human behavior. These Compound Expressions (CE), including combinations such as Happily Surprised, Angrily Surprised, or Sadly Fearful, reflect more realistic and nuanced affective states that occur in everyday interactions.

Although the need to develop approaches for CER has long been recognized, significant interest has emerged in this area since the 6th Affective Behavior Analysis in-the-Wild (ABAW) Competition in 2024 [6, 11, 15, 16]. A wide range of approaches have been proposed, differing in experimental setups. Some researchers have emphasized task-

specific fine-tuning [20, 21], while others have focused on zero-shot learning [26, 30]. The former works have demonstrated high accuracy; however, they are limited to a single target task. In contrast, the latter works have achieved lower accuracy, but address multiple tasks simultaneously, recognizing both basic emotions and CEs.

In this research, we introduce a novel multimodal zero-shot approach specifically designed for CER. It integrates six different modalities – static and dynamic facial expressions, scene and label matching, scene context, audio signals, and textual descriptions, and probability-level fusion – into a unified system. The main contribution of our approach is a specially selected prompt for an off-shelf Large Visual Language Model (LVLM) describing the scene and human behavior. We also propose a Multi-Head Probability Fusion (MHPF) model that adaptively balances contributions from individual modality-specific classifiers. Finally, we present two methods for basic-to-compound emotion conversion, namely Pair-wise Probability Aggregation (PPA) and Pair-wise Feature Similarity Aggregation (PFSA).

In general, our main contributions are as follows:

- We introduce a novel approach based on six modalities for CER.
- We present an efficient MHPF model for dynamically weighting the importance of modalities.
- We propose a novel PFSA method for basic-to-compound emotion conversion.

2. Related Work

We compare our approach with the State-of-the-Art (SOTA) approaches presented in the 6th, 7th, 8th and 9th ABAW Competitions [16, 17, 19]. Several CER approaches have been developed for this task, including our own [28], proposed by the RAS team, the winners of the 9th ABAW challenge.

Savchenko [29, 30] focused on lightweight EmotiNet models to recognize basic facial emotions. The CEs labels were obtained by aggregating the predictions of basic emotions and applying temporal smoothing using block-based and Gaussian filters. In contrast, Richet et al. [26] relied on the text modality and proposed an approach that converted visual and audio signals into textual descriptions. These descriptions, as well as the extracted text transcriptions, served as prompts for Large Language Models (LLMs).

Yu et al. [35] and Wang et al. [33] proposed visual approaches based on static models trained specifically for CER. For model training, both approaches used Real-world Affective Faces Database (RAF-DB) annotated with CEs. In contrast, Wang et al. [33] employed a LVLM using the pre-trained Claude3¹ model to generate additional annota-

¹<https://www.anthropic.com/claude>

tions for test samples in the C-EXPR-DB [7, 18, 19, 36]. Lu et al. [22] also introduced a LVLM approach for CER using a two-stage LoRA adaptation of Qwen-VL. In the first stage, basic emotion representations were learned; in the second stage, CE features were refined using prompt engineering to focus on relevant facial regions and emotional concepts.

Liu, X. et al. [21], Liu, C. et al. [20] and Qiu et al. [24] proposed visual approaches for CER. Liu, X. et al. [21] combined ResNet50, Vision Transformer (ViT), and Facial Action Units (FAU) features with transformer-based temporal modeling and Fully Connected Layers (FCLs) fusion, while Liu, C. et al. [20] introduced a curriculum learning approach based on CutMix and Mixup augmentation methods of CE samples and a Masked Autoencoder for progressive CE training. Qiu et al. [24] collected their own corpus with data annotated for CE and developed an audio-visual ensemble approach based on convolutional features from both spectrograms and facial regions. These features were concatenated and processed using transformer layers, while the final predictions of CEs were obtained by majority voting.

The SOTA approaches focus mainly on the visual modality, LLM and LVLM. Using LVLM, we comprehensively analyze the performance of individual models and multimodal models for CER.

3. Methodology

The pipeline of the approach proposed is presented in Figure 1. Initially, the video files are divided into 4-second segments with a 2-second overlap, as previous studies show. This window is sufficient to extract multimodal emotional patterns. Our approach combines six different modalities, which are detailed in the following.

3.1. Face Models

For each segment, the frame rate is downsampled to 20 frames per 4-second window, ensuring uniform temporal resolution and reducing computational complexity. Facial detection is performed using a two-stage method: an initial YOLO v11² model, pre-trained on WIDER FACE, and fine-tuned for real-world conditions, detects faces under challenging conditions, while a secondary refinement step based on MediaPipe FaceMesh [23] validates face geometry and filters out false-positive regions.

Two models extract static facial features. EmoAffectNet [27], a ResNet-50-based model fine-tuned on AffectNet, classifies six basic emotions (anger, disgust, fear, happiness, sadness, surprise) and neutral states. A Contrastive Language-Image Pretraining (CLIP) [25] model, based on ViT-B/32 with 12 layers and 12 attention heads, is designed

²<https://github.com/akanametov/yolo-face>

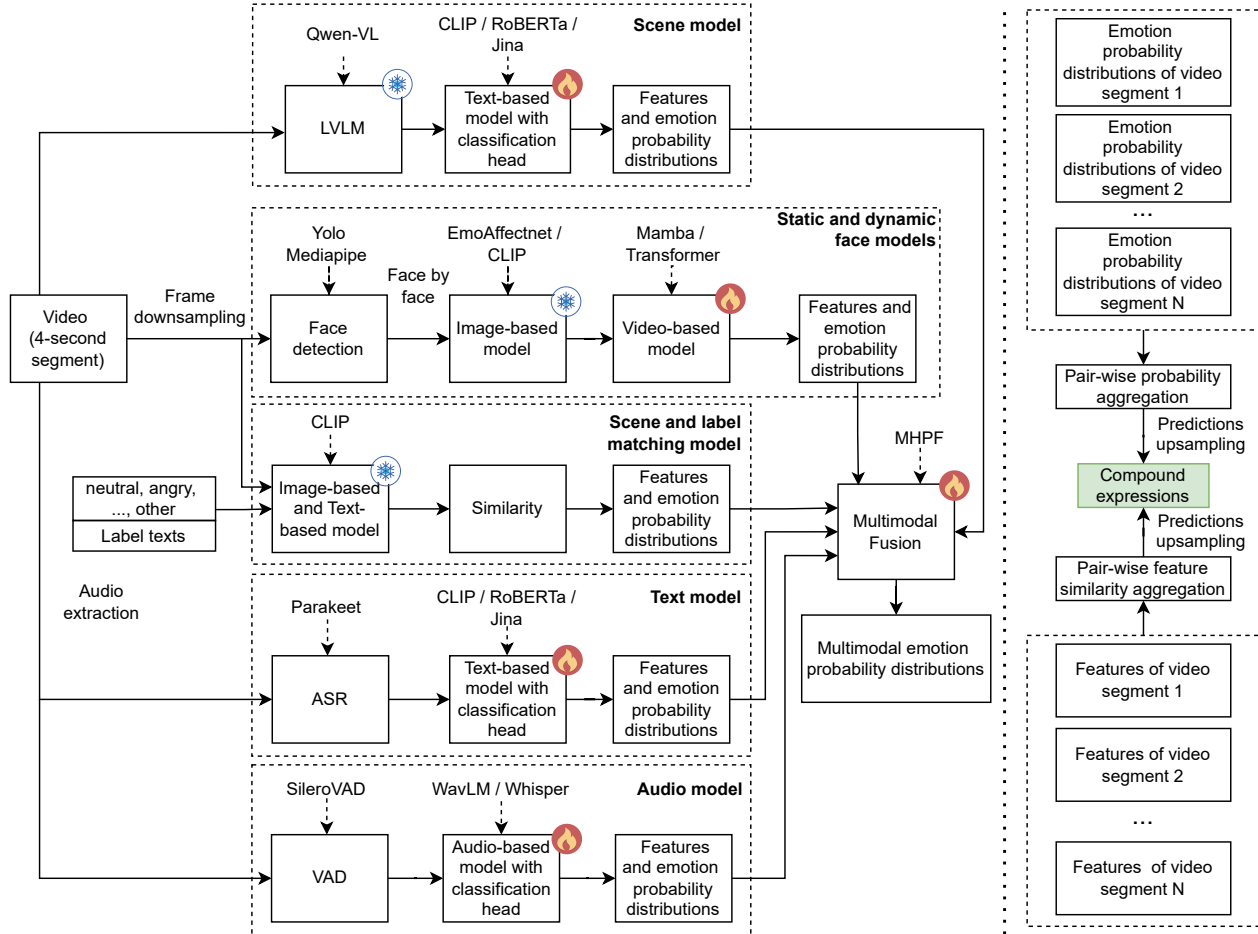


Figure 1. Pipeline of the proposed multimodal CER approach. MHPF refers to Multi-Head Probability Fusion.

for cross-modal matching between images and text. It extracts both textual and visual features simultaneously.

Temporal modeling is carried out using the Mamba [2] and Transformer [32] models. The Mamba uses a selective state-space mechanism that efficiently models long sequences with linear complexity. The Transformer captures global dependencies using self-attention, but at a higher computational cost.

3.2. Scene and Label Matching Model

We use CLIP to generate bimodal features for video frames and eight emotion labels. We compute the cosine similarity between frame and label features to quantify their matching and then apply softmax normalization for probability predictions. By matching visual and textual features, the model supports zero-shot emotion classification and can be extended to other multimodal tasks.

3.3. Scene Models

To include contextual semantics, a modality based on textual scene descriptions is introduced. For each 4-second video segment, a textual description is generated using the Qwen-VL 2.5 32B LVLML³, which processes visual data and produces unlimited natural language output. We experimented with several prompts; however, the most effective one encouraged a free-form emotional interpretation of non-verbal behavior and the surrounding environment, focusing on facial expressions, gestures, and posture.

The generated descriptions, aligned with the CE labels via corpus annotations, are encoded using three transformer-based models: CLIP [25], Emotion English DistilRoBERTa Base (RoBERTa)⁴, and Jina Embeddings V3 (Jina) [31]. A linear classification head was added to each encoder.

³<https://huggingface.co/Qwen/Qwen-VL-Chat-Int4>

⁴<https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/>

3.4. Audio Models

The audio is extracted from each video segment and processed using Silero⁵ Voice Activity Detection (VAD). Only segments with speech are used for training to avoid the influence of non-speech segments. We propose three models for speech emotion recognition.

The first model employs a pre-trained WavLM model [1] as a feature extractor, fine-tuning only the upper encoder and projection layers. To capture local temporal dependencies in audio, a Residual Bidirectional Long Short-Term Memory (BiLSTM) Block is used. Attention Pooling then aggregates the temporal information into a fixed-size feature vector. Finally, a classification head with several FCLs predicts eight emotion labels.

Unlike the first model, which produces deterministic results, the second provides a distribution-based prediction with explicit confidence estimation. This model is based on the WavLM model [1] and consists of a Temporal Relation block that applies multi-head attention and aggregates temporal information with an average pooling. These features are concatenated, processed through FCLs and passed to the Emotion Uncertainty head, which predicts class-wise means and log-variances.

3.5. Text Models

For text analysis, transcriptions are extracted from 4-second audio segments using the Parakeet Token-and-Duration Transducer 0.6B V2 (Parakeet) model⁶. The original attention mechanism is replaced with a local attention block using a 128×128 relative position window to improve computational efficiency on longer samples at the cost of transcription accuracy.

Several transformer-based models are used to extract textual features: CLIP [25], RoBERTa and Jina [31]. Due to the short length of most segments, only the first 1024 tokens of Jina features are used. For other models, the maximum supported token length is used. The resulting features are then aggregated and passed through the final classification head, predicting eight emotion labels.

3.6. Modality Fusion

We consider the modality fusion at the probability-level only. To combine modalities, we propose the MHPF model. This model uses H independent attention heads to assign adaptive, class-specific weights to each of the M input probability distributions $\mathbf{p}_1, \dots, \mathbf{p}_M \in \mathbb{R}^C$, where C is the number of emotion classes. Each head h learns a weight matrix $W_h \in \mathbb{R}^{M \times C}$, which is normalized using a softmax operation across the input dimension and used to calculate

a weighted sum of the input probabilities:

$$\mathbf{o}_h = \sum_{m=1}^M \text{softmax}(W_h, \text{dim} = 0)[m] \odot \mathbf{p}_m,$$

where \odot denotes element-wise multiplication. These head outputs are then linearly combined using learned scalar weights $\alpha \in \mathbb{R}^H$, normalized via softmax:

$$\hat{\mathbf{y}} = \sum_{h=1}^H \text{softmax}(\alpha)_h \cdot \mathbf{o}_h$$

The resulting prediction $\hat{\mathbf{y}} \in \mathbb{R}^C$ represents a flexible and interpretable fusion of class-wise probabilities, where each head captures a distinct weighting strategy. This promotes robust aggregation across varying modalities and distributions.

3.7. Basic-to-Compound Emotion Conversion

In order to convert the original eight emotion labels to a set of seven CEs, we propose two methods: PPA and PFSA.

The PPA method is based on summarizing the predicted probabilities of the corresponding basic emotions to estimate the probabilities of CEs, assuming that CEs can be represented as additive combinations of their corresponding basic emotions.

The PFSA method is based on cosine similarity in the latent space of the model. Emotion-specific features are computed from the validation subsets by averaging feature vectors of correctly classified samples in each basic emotion. Let S_k denote the set of feature vectors correctly classified as basic emotion k , then the prototype vector \mathbf{p}_k is computed as:

$$\mathbf{p}_k = \frac{1}{|S_k|} \sum_{\mathbf{f}_i \in S_k} \frac{\mathbf{f}_i}{\|\mathbf{f}_i\|_2}$$

These prototypes serve as class centroids for basic emotions. The CE prototypes are then constructed by averaging the features of their corresponding basic emotions, followed by the L2 normalization:

$$\mathbf{c}_{(i,j)} = \frac{1}{2} \left(\frac{\mathbf{p}_i}{\|\mathbf{p}_i\|_2} + \frac{\mathbf{p}_j}{\|\mathbf{p}_j\|_2} \right)$$

Each feature vector \mathbf{f} from the test subset of C-EXPR-DB is compared to all CE prototypes using cosine similarity. The resulting scores s_k are passed through a temperature-scaled softmax to calculate the final CEs probabilities \hat{y}_k :

$$s_k = \frac{\mathbf{f}}{\|\mathbf{f}\|_2} \cdot \mathbf{c}_k, \quad \hat{y}_k = \frac{e^{s_k/T}}{\sum_j e^{s_j/T}}$$

This strategy produces a smooth probability distribution over compound emotions with temperature T controlling the confidence-sharpness trade-off.

⁵<https://github.com/snakers4/silero-vad>

⁶<https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2>

To produce stable and consistent frame-wise CER, segment-level predictions are duplicated across the corresponding frames. Overlapping predictions are averaged per frame to maintain temporal coherence in the final CER.

4. Experiments

During training of all video-based models, hyperparameters were varied, including hidden state dimensions {128, 256, 512}, classification feature vectors {128, 256, 512}, number of layers {1 – 4}, Mamba-specific parameters {d_state: 4 – 16; kernel_size: 1 – 7}, Transformer-specific parameters {num_heads: 2 – 16}. The model was trained for 100 epochs with a fixed batch size of 64 and a dropout rate of 0.15. The Adam optimizer with a learning rate of $1e - 5$ was used. Early stopping was applied with a patience of 25 epochs if validation performance plateaued, promoting stable convergence.

Scene-based models were trained for 75 epochs with a batch size of 16 for RoBERTa and 48 for both Jina and CLIP. The maximum token lengths were fixed at 192 for RoBERTa and Jina, and 77 for CLIP. All models were optimized using Adam with a learning rate of $1e - 5$. Early stopping was applied with a patience of 10 epochs.

Audio-based models were trained for 75 epochs using the AdamW optimizer with a learning rate of $1e - 5$ and a CosineAnnealingLR scheduler. The batch size was 32. Early stopping was applied with a patience of 10 epochs. The first model was trained using the Focal loss [34], while the last model was trained using Negative Log-Likelihood (NLL) loss with additional Kullback-Leibler divergence regularization and log-variance penalization.

Text-based models were trained for 100 epochs with batch sizes of 16, 32, and 64 for Jina, RoBERTa, and CLIP, respectively, with the final 4, 3, and 2 layers unfrozen. Training was carried out using Adam with a learning rate of $1e - 4$. Early stopping was applied with a patience of 30 epochs.

4.1. Research Corpora

In our study, we firstly trained our models for emotion recognition using two corpora: Acted Facial Expressions in The Wild (AFEW) [3] and AffWild2 [8–10, 12, 13]. The AFEW dataset contains audio-visual data annotated for six basic emotions and a neutral state, while AffWild2 includes an additional 8th class labeled “other”.

Both corpora have a pre-defined training (773 videos or 855 4-second segments for AFEW and 248 videos or 15158 4-second segments for AffWild2 [4, 5]) and validation (383 videos or 411 4-second segments for AFEW and 70 videos or 6450 4-second segments for AffWild2 [4, 6, 14]) subsets. Although both corpora capture facial expressions in natural settings, AFEW specifically includes acted emotional expressions. Both corpora are relatively close in domain to

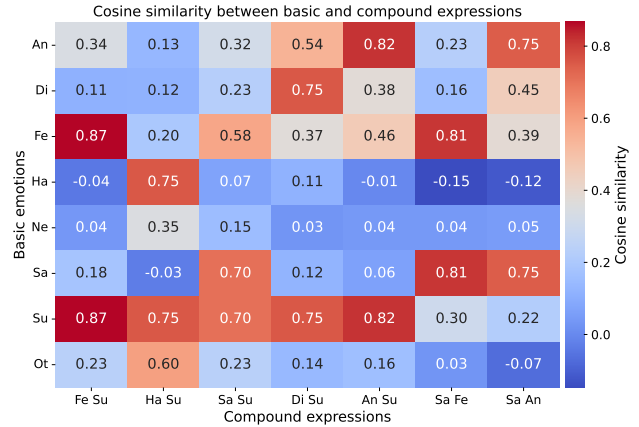


Figure 2. Heatmap visualization of cosine similarity between basic emotions and CEs. Ne, An, Di, Fe, Ha, Sa, Su, Ot refer to Neutral, Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Other

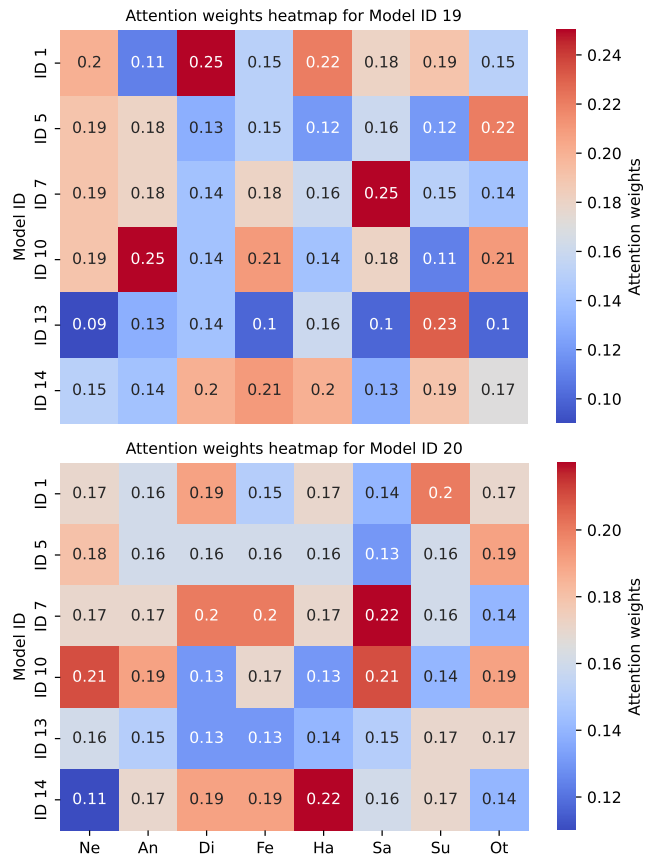


Figure 3. Heatmaps visualization of attention weights for MHPF with one (top) and three heads (bottom).

the target corpus for CER, named C-EXPR-DB [6], which has only 56 videos or 449 4-second segments available for evaluation.

ID	Model	Test corpus							C-EXPR-DB (7cl), macro-F1	
		AffWild2 (8cl)			AFEW (8cl)			PPA	PFSA	
		macro-F1	UAR	Average	macro-F1	UAR	Average			
1	Face & EmoAffectnet	26.08	42.08	34.08	45.90	45.69	45.80	23.09	23.59	
2	Face & EmoAffectnet & Transformer	38.66	42.83	40.74	37.52	36.49	37.01	13.21	15.52	
3	Face & EmoAffectnet & Mamba	35.96	40.61	38.29	38.26	37.70	37.98	19.48	20.54	
4	Face & CLIP & Transformer	41.53	43.40	42.47	34.23	34.20	34.22	14.80	12.30	
5	Face & CLIP & Mamba	40.49	45.47	42.98	37.60	37.31	37.46	16.79	19.04	
6	Scene & LVLM & Jina	34.60	43.33	38.97	31.48	31.70	31.59	22.70	28.83	
7	Scene & LVLM & RoBERTa	36.31	44.68	40.49	32.30	32.04	32.17	34.62	34.85	
8	Scene & LVLM & CLIP	29.01	39.29	34.15	24.38	25.38	24.88	22.37	20.61	
9	Audio & WavLM & U-aware	33.45	33.88	33.67	26.44	26.33	26.39	12.19	17.44	
10	Audio & WavLM & ReBiLSTM	26.38	32.78	29.58	23.09	23.26	23.18	22.90	23.27	
11	Text & Jina	32.54	41.67	37.11	14.36	13.90	14.13	21.68	27.46	
12	Text & RoBERTa	27.21	35.71	31.46	16.24	15.40	15.82	22.07	25.08	
13	Text & CLIP	35.83	45.83	40.83	17.15	17.40	17.25	14.97	12.86	
14	Scene and label matching & CLIP & Cosine similarity	10.84	13.09	11.97	11.79	30.41	24.10	12.18	0.58	
The top three best-performing modality combinations for MHPF with different heads										
15	Models ID 1 & 5 & 6 & 10 & 14 & MHPF (one head)	44.92	49.58	47.25	47.01	46.89	46.95	28.81	–	
16	Models ID 1 & 5 & 6 & 10 & 14 & MHPF (six heads)	45.21	50.54	47.88	46.41	46.47	46.44	31.92	–	
17	Models ID 1 & 5 & 6 & 10 & 13 & MHPF (one head)	46.95	51.08	49.02	45.47	46.00	45.74	32.29	–	
18	Models ID 1 & 5 & 6 & 10 & 13 & MHPF (two heads)	46.33	50.62	48.47	44.30	45.11	44.71	29.06	–	
19	Models ID 1 & 5 & 6 & 10 & 13 & 14 & MHPF (one head)	46.86	50.93	48.89	44.78	45.60	45.19	32.61	–	
20	Models ID 1 & 5 & 6 & 10 & 13 & 14 & MHPF (three heads)	46.50	51.04	48.77	47.14	47.53	47.34	30.70	–	

Table 1. Experimental results. PF is Probability Fusion. MHPF is Multi-Head Probability Fusion. PPA is Pair-wise Probability Aggregation. PFSA is Pair-wise Feature Similarity Aggregation. The gray-highlighted performance measures represent our official submissions to the CER challenge in the 9th ABAW Competition.

Table 2. Performance comparison (macro-F1, %) of SOTA methods. FT is a fine-tuning setup. ZS is a zero-shot setup.

Method	ABAW	Setup	Modality	C-EXPR-DB
Wang et al. [33]	6th	FT	V	18.45
Yu et al. [35]	6th	FT	V	22.40
Liu et al. [21]	7th	FT	V	22.81
Qiu et al. [24]	6th	FT	AV	55.26
Lu et al. [22]	8th	FT	V	57.23
Liu et al. [20]	7th	FT	V	60.63
Richet et al. [26]	7th	ZS	AVT	25.91
Savchenko [30]	7th	ZS	V	32.43
Ours	9th	ZS	V	34.85

4.2. Experimental Results

The experimental results demonstrate (see Table 1) varying performance across different model configurations on three corpora: AffWild2 (8 classes), AFEW (8 classes), and C-EXPR-DB (7 classes).

All dynamic facial models (Model IDs 2–5) demonstrate a higher average performance on the AffWild2 and AFEW corpora, but show a significantly lower macro-F1 on the C-EXPR-DB corpus. This suggests that these models become more confident in predicting basic emotions, while more complex facial emotions are captured less effectively. In contrast, the static facial model (Model ID 1), which was not fine-tuned on any of the corpora used in this study, achieves moderate performance across all the corpora, indi-

ating strong generalization capabilities on unseen data.

Scene-based models (Model IDs 6–8) show moderate performance on both corpora (AffWild2 and AFEW) and the target C-EXPR-DB corpus, with the exception of Model ID 7, which outperforms all models on C-EXPR-DB (macro-F1 of 34.85%). This implies that scene-descriptive prompts, combined with behavioral context, enable cross-corpus generalization, and the inherent emotional knowledge of the RoBERTa model increases its advantage.

Despite the complexity of corpora designed for visual analysis, audio-based (Model IDs 9–10) and text-based models (Model IDs 11–13) achieve lower performance on AffWild2 and AFEW, but outperform video-based models (Model IDs 1–8) on C-EXPR-DB. This emphasizes the importance of integrating acoustic and linguistic features alongside visual features. The Scene and Label Matching model (Model ID 14) shows low performance across all corpora, highlighting the need for task-specific fine-tuning of the CLIP model.

The PFSA method shows better performance than the PPA method. This suggests that probability predictions alone are insufficient to capture the underlying structure of the data. Instead, the creation of CE prototypes in the feature space provides a more coherent representation of basic-to-compound emotion relationships. Figure 2 presents a heatmap of the cosine similarity between basic and CEs for Model ID 7. Four out of seven CEs correctly matched their basic components. However, three CEs show confusion: Happily Surprised is misclassified as Other (Ot), Sadly Sur-

prised as Fear (Fe), and Disgustedly Surprised as Anger (An). This may reflect semantic proximity in the scene and behavioral descriptions. The heatmap also highlights the opposites of some expressions. For example, the cosine similarity between Happiness (Ha) and such CEs as Fearfully Surprised, Angrily Surprised, Sadly Fearful, and Sadly Angry has a negative value. This suggests that the features are not directly comparable and that the model is able to distinguish between them.

For multimodal fusion, we combined the best unimodal models (in terms of average performance on AffWild2 and AFEW), resulting in 56 configurations involving 2 to 6 modalities across six models. Table 1 shows the top three combinations for MHPF with different numbers of heads. While Model ID 20 (three heads) achieved a higher average performance on AffWild2 and AFEW, Model ID 19 (one head) showed the best result on C-EXPR-DB (macro-F1 of 32.61%). Although multimodal fusion proves to be effective for CER, it achieves lower results than Model ID 7.

Figure 3 shows heatmaps of the attention weights for two models (IDs 19 and 20), illustrating how attention is distributed between different modalities when predicting various emotions. The weights vary significantly between models, indicating different decision strategies. For example, Model ID 19 relies more on the static facial modality and the scene and label matching modality (Model IDs 1 and 14) when predicting Disgust (Di), while Model ID 20 places additional emphasis on the scene modality (Model ID 7).

In addition, heatmaps show which modalities each model relies on when predicting CEs. For example, when predicting Sadly Surprised, Model ID 19 assigns greater weight to the scene and textual modalities (Model IDs 7 and 13), while Model ID 20 focuses on the static facial, audio, and scene modalities (Model IDs 1, 7 and 10). In general, Model ID 19 distributes attention across modalities less uniformly, assigning higher weights to select modalities, which likely contributes to its better performance in CER. This suggests that its simple attention mechanism (with only one head) is more context-sensitive and better suited to the complex nature of CEs.

Therefore, while multimodal fusion yields improvements on emotional corpora with basic emotions, the scene model is the most effective on C-EXPR-DB [19] for CER. This shows that the correct prompting of LVLM can achieve reliable results for both basic emotion recognition and CER. Comparison with the SOTA results (see Table 2) shows that the proposed approach outperforms all approaches not trained on the target task, and is only surpassed by [20, 22, 24], which are not zero-shot solutions.

5. Conclusions

In this paper, we proposed a novel multimodal approach for CER that combines six modalities: static and dynamic

facial expressions, scene and label matching, scene context, audio, and text, as well as probability-level fusion. Each modality is processed using zero-shot or general-purpose pipelines based on the CLIP, Qwen-VL, WavLM, and RoBERTa models, and temporal dynamics is modeled by the Transformer and Mamba models. MHPF dynamically combines emotion probability distributions, followed by a basic-to-compound emotion conversion based on Pair-wise Probability Aggregation (PPA) or Pair-wise Feature Similarity Aggregation (PFSA). Our results show that PFSA significantly outperforms PPA in CER, demonstrating the potential of constructing CE prototypes. In addition, while simple probability weighting (MHPF with one head) outperforms complex one (MHPF with three heads) in CER, it is less reliable for basic emotion recognition.

Evaluated in the multi-corpus training setup, the method proposed achieves macro-F1 scores of 46.95% on AffWild2 and 49.02% on AFEW, as well as 34.85% on C-EXPR-DB without any fine-tuning on the target task. Multimodal fusion showed the best results on AffWild2, AFEW, while on C-EXPR-DB the best results were achieved by LVLM, highlighting the potential of prompt-based strategies and scene descriptions in CER. These results are competitive with those of supervised approaches, demonstrating the effectiveness of our zero-shot pipeline.

6. Acknowledgements

This work was supported by the Ministry of Economic Development of the Russian Federation (IGK 000000C313925P4C0002), agreement No139-15-2025-010.

References

- [1] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. 4
- [2] Tri Dao and Albert Gu. Transformers are ssms: generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, pages 10041–10071, 2024. 3
- [3] Abhinav Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *International Conference on Multimodal Interaction (ICMI)*, pages 546–550, 2019. 5
- [4] Dimitrios Kollias. ABAW: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *CVPR*, pages 2328–2336, 2022. 5
- [5] Dimitrios Kollias. ABAW: Learning from synthetic data &

- multi-task learning challenges. In *ECCV*, pages 157–172, 2023. [5](#)
- [6] Dimitrios Kollias. Multi-label compound expression recognition: C-EXPR database & network. In *CVPR*, pages 5589–5598, 2023. [1](#), [5](#)
- [7] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and arface. *arXiv preprint arXiv:1910.04855*, pages 1–15, 2019. [2](#)
- [8] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, pages 1–20, 2021. [5](#)
- [9] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second ABAW2 competition. In *CVPR*, pages 3652–3660, 2021.
- [10] 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*, pages 1–11, 2019. [5](#)
- [11] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, et al. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *IJCV*, pages 1–23, 2019. [1](#)
- [12] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *International Conference on Automatic Face and Gesture Recognition (FG)*, pages 794–800, 2020. [5](#)
- [13] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, pages 1–15, 2021. [5](#)
- [14] 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 *CVPR*, pages 5888–5897, 2023. [5](#)
- [15] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for multi-task learning of classification tasks: a large-scale study on faces & beyond. In *AAAI Conference on Artificial Intelligence*, pages 2813–2821, 2024. [1](#)
- [16] 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 *CVPR*, pages 4587–4598, 2024. [1](#), [2](#)
- [17] 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*, pages 1–14, 2024. [2](#)
- [18] Dimitrios Kollias, Damith C Senadeera, Jianian Zheng, Kaushal KK Yadav, Greg Slabaugh, Muhammad Awais, and Xiaoyun Yang. Dvd: A comprehensive dataset for advancing violence detection in real-world scenarios. *arXiv preprint arXiv:2506.05372*, pages 1–10, 2025. [2](#)
- [19] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Eric Granger, Marco Pedersoli, Simon 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. In *CVPR*, pages 5572–5583, 2025. [2](#), [7](#)
- [20] Chen Liu, Feng Qiu, Wei Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Compound expression recognition via curriculum learning. In *European Conference on Computer Vision*, pages 282–293. Springer, 2025. [2](#), [6](#), [7](#)
- [21] Xuxiong Liu, Kang Shen, Jun Yao, Boyan Wang, Yu Wang, Yujie Guan, Xin Liu, Gengchen Li, Liuwei An, Zishun Cui, Minrui Liu, Xiao Sun, and Weijie Feng. Abaw7 challenge: A facial affect recognition approach based on transformer encoder and multilayer perceptron. In *European Conference on Computer Vision*, pages 267–281. Springer, 2025. [2](#), [6](#)
- [22] Xilong Lu, Jun Yu, Yunxiang Zhang, Lingsi Zhu, Yang Zheng, Yongqi Wang, and Qiang Ling. Robust stage-wise lvlm adaptation: Multi-phase prompt lora fine-tuning for compound expression recognition. In *CVPRW*, pages 5770–5777, 2025. [2](#), [6](#), [7](#)
- [23] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, pages 1–9, 2019. [2](#)
- [24] Feng Qiu, Heming Du, Wei Zhang, Chen Liu, Lincheng Li, Tianchen Guo, and Xin Yu. Learning transferable compound expressions from masked autoencoder pretraining. In *CVPRW*, pages 4733–4741, 2024. [2](#), [6](#), [7](#)
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. [2](#), [3](#), [4](#)
- [26] Nicolas Richet, Soufiane Belharbi, Haseeb Aslam, Meike Emilie Schadt, Manuela González-González, Gustave Cortal, Alessandro Lameiras Koerich, Marco Pedersoli, Alain Finkel, Simon Bacon, and Eric Granger. Textualized and feature-based models for compound multimodal emotion recognition in the wild. In *European Conference on Computer Vision*, pages 60–78. Springer, 2025. [2](#), [6](#)
- [27] Elena Ryumina, Denis Dresvyanskiy, and Alexey Karpov. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, 514: 435–450, 2022. [2](#)
- [28] Elena Ryumina, Maxim Markitantov, Alexandr Axyonov, Dmitry Ryumin, Mikhail Dolgushin, and Alexey Karpov. Team ras in 9th abaw competition: Multimodal compound expression recognition approach. *arXiv preprint arXiv:2507.02205*, pages 1–7, 2025. [2](#)
- [29] Andrey V Savchenko. Leveraging pre-trained multi-task deep models for trustworthy facial analysis in affective be-

- haviour analysis in-the-wild. In *CVPRW*, pages 4703–4712, 2024. [2](#)
- [30] Andrey V Savchenko. Smoothing predictions of multi-task emotinet models for compound facial expression recognition. In *European Conference on Computer Vision*, pages 257–266. Springer, 2025. [2](#), [6](#)
- [31] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*, pages 1–20, 2024. [3](#), [4](#)
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, pages 1–11, 2017. [3](#)
- [33] Jiahe Wang, Jiale Huang, Bingzhao Cai, Yifan Cao, Xin Yun, and Shangfei Wang. Zero-shot compound expression recognition with visual language model at the 6th abaw challenge. *arXiv preprint arXiv:2403.11450*, pages 1–4, 2024. [2](#), [6](#)
- [34] Zhenyuan Wang, Xuemei Xie, Jianxiu Yang, and Guangming Shi. Soft focal loss: Evaluating sample quality for dense object detection. *Neurocomputing*, 480:271–280, 2022. [5](#)
- [35] Jun Yu, Jichao Zhu, Wangyuan Zhu, Zhongpeng Cai, Gongpeng Zhao, Zhihong Wei, Guochen Xie, Zerui Zhang, Qingsong Liu, and Jiaen Liang. Multi model ensemble for compound expression recognition. In *CVPRW*, pages 4873–4879, 2024. [2](#), [6](#)
- [36] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, et al. Aff-wild: Valence and arousal ‘in-the-wild’ challenge. In *CVPRW*, pages 1980–1987, 2017. [2](#)