Language-guided Multi-modal Emotional Mimicry Intensity Estimation

Feng Qiu[1,*], Wei Zhang[1,*], Chen Liu[1,2], Lincheng Li[1,†], Heming Du[2], Tianchen Guo[2], Xin Yu[2]
[1] Netease Fuxi AI Lab
[2] The University of Queensland
{qiufeng, zhangwei05, lilincheng}@corp.netease.com, chen.liu7@uqconnect.edu.au,
{Heming.du, tianchen.guo, xin.yu}@uq.edu.au

Abstract

Emotional Mimicry Intensity (EMI) estimation aims to identify the intensity of mimicry exhibited by individuals in response to observed emotions. The challenge in EMI estimation lies in discerning nuanced facial expression cues on mimicry behaviors based on the seed video and the text instructions. In this paper, we propose a multi-modal EMI estimation framework by leveraging visual, auditory, and textual modalities to capture a comprehensive emotional profile. We first extract representations for each modality separately and then fuse the modality-specific representations via a Temporal Segment Network, optimizing for temporal coherence and emotional context. Furthermore, we find that participants demonstrate notable proficiency in mimicking text instructions, yet exhibit less effectiveness in replicating facial expressions and vocal tones. In light of this, we design a contrastive learning mechanism to refine the extracted feature based on textual guidance. By doing so, features derived from similar text instructions are closely aligned, enhancing the estimation of emotional mimicry intensity by leveraging the dominant textual modality. Experiments conducted on the Hume-Vidmimic2 dataset illustrate the effectiveness of our framework in EMI estimation. Our framework is recognized as the leading solution in the Emotional Mimicry Intensity (EMI) Estimation Challenge at the 6th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW). More information for the Competition can be found in: 6th ABAW.

1. Introduction

Emotional Mimicry Intensity (EMI) refers to the degree to which individuals mimic the emotional expressions, voices, or gestures of others during social interactions [35, 38, 39, 41, 42]. With the development of artificial intelligence technology, how to utilize AI systems to accurately identify and respond to human emotional states has drawn a widespread concern [22, 66]. This task is crucial for improving Human-Computer Interaction (HCI), enabling computers and robots to respond more naturally and making interactions more engaging and effective [20, 57, 67].

To investigate the analysis of emotional behavior in real-world environments, the 6th Affective Behavior Analysis in-the-wild (ABAW) Competition [33–42, 42, 43, 82] establishes a track for the Emotional Mimicry Intensity Estimation Challenge. This challenge focuses on analyzing and assessing the emotional intensity that participants exhibit when they mimic or respond to emotions displayed in a "seed" video. It employs the multi-modal Hume-Vidmimic2 [43] dataset, which consists of 15,000 videos, totaling more than 25 hours of content. Participants in these videos imitate the emotional expression seen in the seed videos and then evaluate the intensity of these emotions across several dimensions, such as "Admiration", "Amusement", "Determination", "Empathic Pain", "Excitement", and "Joy".

In this paper, we propose an effective Emotional Mimicry Intensity estimation framework by fully integrating the emotional features from multi modalities *i.e.,* visual, audio, and text. In this competition, we mainly focus on the following two aspects: (1) how to obtain the advantageous modal representations and (2) the impact of various modalities on the accuracy of emotional mimicry. We demonstrate that these are the two key factors for achieving a more robust emotional intensity assessment approach.

To achieve the first objective, we employ three large-scale foundation models as our feature extractors, *i.e.,* the Masked Auto Encoding (MAE) [25, 84], the Wav2Vec2 [3], and the ChatGLM3 [17, 83]. Note that, since the MAE is pre-trained on the large-scale general dataset, which prioritizes broad feature representation, we fintune the model on AffectNet [53]. In this fashion, the visual feature extractor is more suitable for the emotion analysis task. Wav2Vec2 [3] and ChatGLM3 [17, 83] are trained on large and diverse

[*]Equal contribution
[†]Corresponding author

datasets, and we directly leverage their pre-trained models to extract audio and text features, respectively. Then we devise a Temporal Segment Network to fuse the specific multi-modal representations. Specifically, we employ the GRUs (BiGRUs) to model the temporal information and integrate emotional cues from various modalities.

Moreover, we observe that the impact of various modalities varies in the seed video on mimicry accuracy. Compared with audio and visual cues, text instructions can provide the mimicry with more explicit guidance. Inspired by this, we consider the text feature as the primary feature and leverage contrastive learning [17, 83] to narrow the gap between the three modalities. This design enhances the correlation between different modalities, allowing our model to better integrate and utilize various information, thereby improving the emotional intensity estimation accuracy.

Experiments conducted on the official validation dataset demonstrate the effectiveness of our method designs. Moreover, our team (*i.e.,* **NetEase Fuxi AI Lab**) attain first place in the EMI track, further proving the generalization capability of our method. Overall, our contributions are two-fold:

- We leverage the large foundation models to generate the multi-modal representations and integrate them via a Temporal Segment Network. This enriches the emotional features at the spatial and temporal dimensions.
- We regard the text feature as a guiding force and align the multimodal features with it. This enhances the generalization ability and estimation performance of our model.

2. Related Work

2.1. Emotional Mimicry Intensity Estimation

Emotional mimicry, defined as the automatic replication of another's non-verbal expressions, has long been recognized as a fundamental component in the communication of affective states [4, 21]. Lipps [45] and Rogers [58] posit that mimicry facilitates empathic communication and offers insights into an individual's internal state, a concept further embraced by various therapeutic practices [65]. Facial mimicry, often described as a reflex-like, automatic process [24, 27, 45, 76–81], involves an observer's facial expressions mirroring those observed, contributing to emotional contagion—a phenomenon where an individual's affective state aligns with that of another. Despite the close relationship between mimicry and emotional contagion, distinctions are made, with mimicry pertaining solely to expressive components and contagion encompassing affective states [24].

Empirical evidence [26] support the prevalence of mimicry across various behaviors and age groups, highlighting its role in congruent emotional displays. However, instances of counter-mimicry, as found in competitive versus collaborative settings [26, 44] suggest that mimicry's automaticity may be influenced by context and task type. Moreover, the relationship between mimicry and emotion recognition remains complex. While mimicry is hypothesized to facilitate emotional understanding through feedback mechanisms [27, 45], recent findings challenge this assumption, indicating no significant link between mimicry and enhanced emotion recognition [5, 23, 26].

The context-dependency of mimicry, particularly in response to less prototypical and more natural expressions [23, 26, 44], as well as its modulation by personal attitudes [6, 26, 52], suggests a nuanced understanding of mimicry, beyond reflex-like responses to extreme stimuli. This insight raises questions about the everyday applicability of mimicry and its role in emotion recognition, especially when employing prototypical, intense expressions as stimuli.

Prior research on quantifying emotional mimicry has relied on facial muscle activity measured through electromyography (EMG) or the Facial Action Coding System (FACS) applied to facial movements [18]. While these methods offer precise measurement, they are either invasive (EMG) or require extensive manual analysis (FACS). To address these limitations, recent work has explored utilizing computer vision and statistical techniques for automatic estimation of facial expressions, postures, and emotions from video recordings [13, 31, 63, 64, 70]. This video-based approach offers a non-invasive, automatable, and scalable solution for real-world applications like human-agent interaction, albeit with the current drawback of potentially lower precision compared to physiological signal-based measurements. In this work, by leveraging multi-modal data, including visual, audio, and textual inputs, we aim to enrich the quality of expression features derived, thereby enhancing the robustness and applicability of emotion recognition systems [48, 55, 84–87] in complex, uncontrolled environments.

2.2. Multi-modal Feature Extraction

Multi-modal Feature Extraction plays a pivotal role in enhancing the performance of emotion recognition systems [30, 32, 46, 47, 54, 62, 87] by leveraging diverse sources of information. In the realm of audio features, the adoption of models such as Wav2Vec 2.0 [3], HuBERT [28], and WavLM [10] exemplifies the trend towards utilizing self-supervised learning techniques to capture rich speech representations from large-scale unlabeled audio data [7, 8, 72]. Specifically, the Wav2Vec 2.0 [3] framework, with its predictive audio encoder and quantization module, has been instrumental in learning nuanced speech representations that are highly beneficial for emotion-related tasks.

The evolution of text features has been markedly accelerated with the advent of Large Language Models (LLMs),

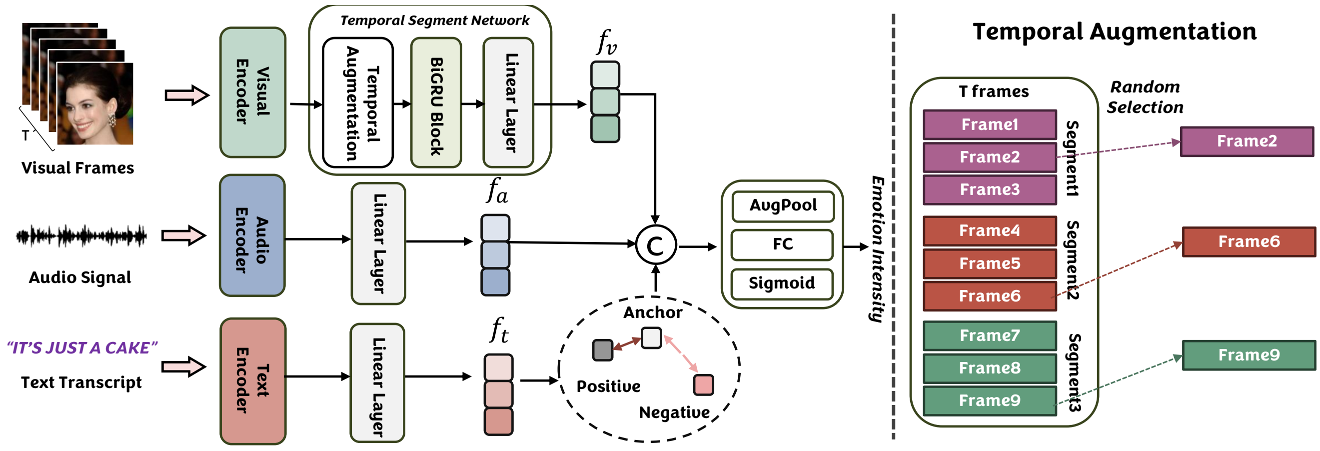

Figure 1. The overview of our proposed framework. First, we extract unimodal features from images, audio, and text separately. Then, we introduce a temporal augmentation module to sample image feature sequences, enhancing their temporal generalizability. We consider the text modality as the leading modality and introduce contrastive learning to align the final multimodal features with text features. Finally, we use a late fusion strategy to obtain the multimodal features and estimate the intensity of emotions.

such as LLaMA [69], GLM [16], and GPT [1]. These models, by virtue of their vast parameter scales and extensive pre-training on diverse corpora, have significantly outperformed traditional models in extracting text features that are more effective for emotion recognition [74]. The use of LLMs in combination with instruction tuning techniques has further pushed the boundaries, enabling these models to adapt more effectively to the task-specific nuances of emotion recognition from textual data.

Video feature extraction [9, 15, 19, 25] has also seen notable advancements with the integration of models like Vision Transformer (ViT) [14] and Facial Action Unit (FAU) detectors [29]. The use of ViT, particularly those pre-trained methods such as Masked Autoencoder (MAE) [25] and DINO [9], underscores the shift towards self-supervised learning paradigms in the video domain [68]. In this paper, our designed framework allows for the extraction of facial features that are more aligned with emotional expressions, thereby enhancing the overall efficacy of multi-modal emotion recognition systems.

3. Method

This section presents our multimodal framework designed to estimate the emotional mimicry intensities of individuals in videos. Our method consists of three branches, each extracting unimodal features from images, audio, and text, respectively. To enhance the generalizability of the model, we introduced a temporal augmentation module to sample the feature sequences. Then, we use a late fusion strategy to integrate multimodal features for estimating the intensities of six mimicry emotions. Additionally, we position text as the dominant modality and introduce contrastive learning to constrain the final output results.

3.1. Unimodal feature extraction

We define the input face images for the visual branch as $\mathcal{X}_v \in \{I_1, I_2, .., I_t, ...I_T\}$, where T is the number of total frames and I_t is the t-th image in an image sequence.

3.1.1 Visual feature

For visual features, we utilize the vision transformer (ViT) [14] model as the visual encoder to extract spatial feature $f_v \in \mathbb{R}^{T \times d_v}$ from each frame, where d_v is the feature dimension of the output in ViT encoder. To obtain more robust visual features, our ViT encoder is trained in two steps. First, in a self-supervised manner, we employ the masked auto encoding (MAE) method to train the model on an image reconstruction task in an unlabeled large-scale face dataset, which includes AffectNet [53], CASIA-WebFace [73], CelebA [49] and IMDB-WIKI [59]. Specifically, we train a model comprising a ViT encoder and a decoder. The input of the model is a facial image with a large portion (75%) of patches masked, and it is required to reconstruct raw pixel values and output the complete original image. MAE is capable of learning a network with excellent generalization ability. After training, we retain only the ViT encoder as our visual encoder.

In the second step, we add two fully connected layers after the ViT encoder and then finetune the model on an expression classification task. The purpose of this step is to enhance the model's ability to understand specific downstream tasks, specifically improving the network's capability to analyze emotional behaviors. This enables us to extract more effective visual features for our final task. More specifically, we finetune the ViT encoder on the AffectNet [53] dataset. Our model achieves the top-1 accuracy of

69.77% and F1 score of 0.3515 on the test set of AffectNet. After completing the training, we freeze the parameters of the ViT encoder, using it as our visual encoder to extract facial expression features from images. We denoted our ViT as EmoViT.

3.1.2 Audio feature

We utilize a speech model, Wav2Vec2 [3] to extract audio features $f_a \in \mathbb{R}^{T \times d_a}$ from the raw wavefrom of the speech signal, where d_a is the feature dimension of the output. Wav2Vec2 conceals parts of the speech input within the latent space and addresses a contrastive task that is defined based on a quantization of the latent representations, which are learned simultaneously.

3.1.3 Text feature

To extract text features, we first need to transcribe the text from the audio. In this work, we use Whipser [56] to convert the speech into text. Whisper is a state-of-the-art automatic speech recognition (ASR) system developed through training on approximately 680,000 hours of supervised multilingual and multitask data sourced from the internet. The extensive and varied nature of this dataset enhances its adaptability to various accents, and resilience against background accents, background noise, and technical language. Additionally, this system supports transcription in numerous languages and offers capabilities for translating these languages into English.

After that, we incorporate a large language model (LLM) to extract text features. Large language models stand out for their capacity to perform general-purpose language generation and to tackle various natural language processing tasks, such as classification, text generation and emotion analysis. LLMs develop these capabilities by learning statistical correlations from text documents during a computationally intensive self-supervised and semi-supervised training process. Specifically, we use ChatGLM3 [17, 83] as the text encoder to extract features $f_t \in \mathbb{R}^{T \times d_t}$ from words, where d_t is the feature dimension of the output. To ensure the accuracy of text extraction, we used the Fuxi Youling Crowdsourcing Platform and Fuxi Agent-Oriented Programming (AOP) System for text verification.

3.2. Temporal augmentation

Given the considerable variation in the number of frames T across videos, we implement a segment-based sampling approach akin to the one utilized in Temporal Segment Networks (TSN) [71]. This strategy allows our model to capture the temporal characteristics of the entire video, independent of its duration. Moreover, this approach also serves as a form of temporal augmentation. Performing random sampling within each video segment effectively broadens the model's capacity for temporal generalization.

Specifically, for a sequence of features $\mathcal{F}$, we divide it into K segments $\{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_k\}$ with the same frame number. And then we sample one frame from each segment randomly to form a new sequence of feature $\hat{\mathcal{F}} \in \{f_1, f_2, .., f_K\}$ with K frames.

Following the temporal augmentation module, or each of the three modalities, we employ a separate Bidirectional Recurrent Unit (BiGRU) block to aggregate contextual information, thereby extracting temporal features from sequences. Our BiGRU block consists of two BiGRU layers following a layer normalization [2] and a linear layer. To further augment the feature representation capacity, we additionally use a linear layer to produce the final 256-dimensional unimodal features $\hat{f}_v$, $\hat{f}_a$ and $\hat{f}_t$ for image, audio, and text, respectively.

3.3. Late fusion

When expressing and understanding emotions, signals from different modalities often complement each other. To conduct a more comprehensive analysis of emotions and avoid overfitting to any specific modality during training, we employed a late-fusion approach to integrate features from multiple modalities.

Specifically, we simply use an average pooling layer of three features $\hat{f}_v$, $\hat{f}_a$ and $\hat{f}_t$ extracted by visual, audio, and text branches, respectively. To enhance the generalization ability of the model, two fully connected layers following a dropout layer are adopted to estimate the emotional intensities. Because the value range of labels is between 0 and 1, we add a sigmoid activation function to normalize the predicted results to (0,1). The process can be formulated as:

$$\hat{y} = Sigmoid(FC(AvgPool(\hat{f}_v, \hat{f}_a, \hat{f}_t))) \quad (1)$$

where $\hat{y}$ denotes the predicted intensity.

3.4. Text-based contrastive learning

Our experiments reveal that the dominant modality for emotional mimicry intensity estimation is the textual modality (refer to Table 1). This is attributed to the dataset's collection method, which involves imitating a seed video. Participants are generally able to mimic the dialogue accurately, but their imitation of facial expressions and tone of voice is less precise. Therefore, based on contrastive learning, we introduced a triplet loss based on textual features. This means constraining the relative distances of the output results to align with the relative distances of the input text features.

To be specific, we first use a Global Average Pooling layer to the word features extracted by ChatGLM3 along the temporal dimension, resulting in a one-dimensional text feature. During the training process, we randomly sample three examples from a mini-batch to form a triplet. Based on the relative distances of their text features, we label them as anchor, positive, and negative, respectively. Within a triplet, the distance from the anchor to the positive should be smaller than the distance from the anchor to the negative. Then, we utilize the triplet loss to constrain the final prediction results, which is calculated as follows:

$$\hat{f} = AvgPool(\hat{f}_v, \hat{f}_a, \hat{f}_t) \tag{2}$$

$$\begin{aligned}\mathcal{L}_{triplet} = &\max\left(0, \left\|\hat{f}_{anc} - \hat{f}_{pos}\right\|^2 - \left\|\hat{f}_{anc} - \hat{f}_{neg}\right\|^2 + \gamma\right) \\ &+ \max\left(0, \left\|\hat{f}_{anc} - \hat{f}_{pos}\right\|^2 - \left\|\hat{f}_{pos} - \hat{f}_{neg}\right\|^2 + \gamma\right)\end{aligned} \tag{3}$$

where $\hat{f}$ is the multimodal features fused by three unimodal features, $\hat{f}_{anc}$ and $\hat{f}_{pos}$ are interchangeable with each other, γ is a enforced margin and set to 0.1.

4. Experiment

4.1. Dataset

For the Emotional Mimicry Intensity Estimation Challenge, we utilize the multimodal Hume-Vidmimic2 [11] dataset, which aims to address the problem of acquiring data related to human affective behavior. In this dataset, subjects are required to mimic the individuals in the seed videos. Subsequently, the seed videos need to be annotated with the intensity of seven specified emotions, e.g. Admiration, Amusement, Determination, Empathic Pain, Excitement, and Joy. In total, Hume-Vidmimic2 collects more than 15,000 videos, with a total duration exceeding 25 hours.

4.2. Experimental Setting

We first extract frames from all videos in the Hume-Vidmimic2 database by OpenCV. Then we utilize RetinaFace [12] for face detection and subsequently crop the facial images from the original pictures. Besides that, we employ a speech recognition model, Whisper [56], to transcribe the spoken words in the videos, facilitating the extraction of text features for subsequent analysis.

Before fed into the network, all facial images are uniformly resized to a dimension of 224 $\times$ 224. The experimental codebase is developed in the PyTorch framework, with the training and validations executed on NVIDIA A30 GPUs. For optimization, we use AdamW[51] as our optimizer and set the size of the mini-batch to 16. When training the multimodal network, we set different learning rates for the various modules. Specifically, for the visual feature extraction module, we set the learning rate at 1e-6; for the text and audio feature extraction modules, we use a learning rate of 1e-4. As for the final multimodal feature fusion module, we apply a learning rate of 1e-6. The learning rate is dynamically adjusted according to the Cosine Annealing [50] strategy, featuring a minimum learning rate of 1e-8 and restart epochs every 5 cycles. To ensure optimal training duration and efficiency, an early-stopping mechanism is enforced, activating after a patience interval of 10 epochs. To further ensure the stability of the training phase, the Exponential Moving Average (EMA) strategy is adopted, characterized by a decay rate of 0.999.

Table 1. Comparison of the results of emotion mimicry intensity estimation models trained on different features.

Visual	Audio	Text	ρ
ViT			0.0873
EmoViT			0.1685
ViT +EmoViT			0.1490
	Wav2Vec2		0.2576
	HuBERT		0.1472
	Wav2Vec2 +HuBERT		0.2028
		ChatGLM3_28th	0.4665
		ChatGLM3_21st	0.4846
		ChatGLM3_14th	0.4842
		ChatGLM3_28th_lora	0.4592
		ChatGLM3_21st +ChatGLM3_28th	0.4879
EmoViT	HuBERT	ChatGLM3_21st +ChatGLM3_28th	0.4931

Table 2. Comparison of the results of emotion mimicry intensity estimation models with different settings.

Temporal Augment	EMA	Triplet Loss	Fusion	ρ
×	×	×	Average	0.4931
×	×	×	Concatnate	0.4879
×	×	×	Multimodal Transformer	0.4645
✓	×	×	Average	0.5124
✓	✓	×	Average	0.5247
✓	✓	✓	Average	0.5851

4.3. Metrics

For the Emotional Mimicry Intensity Estimation Challenge, we evaluate the performance by averaging Pearson's corre-

Table 3. The Pearson's correlations of models that are trained and tested on different folds (including the original training/validation set of Hume-Vidmimic2).

	Admiration	**Amusement**	**Determination**	**Empathic Pain**	**Excitement**	**Joy**	**Average**
Official	0.7155	0.6159	0.6303	0.3488	0.6174	0.5793	0.5851
fold-1	0.6305	0.6355	0.6242	0.6070	0.6399	0.6322	0.6282
fold-2	0.6365	0.6419	0.6319	0.6124	0.6450	0.6397	0.6346
fold-3	0.6399	0.6450	0.6353	0.6150	0.6490	0.6451	0.6382
fold-4	0.6509	0.6526	0.6436	0.6266	0.6554	0.6560	0.6475
fold-5	0.6442	0.6488	0.6397	0.6197	0.6509	0.6491	0.6421

lations (ρ) across the 6 emotion dimensions, defined as:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

$$P_{EMI} = \frac{\sum_{c=1}^{6} \rho_c}{6}. \tag{5}$$

where n is the number of data points, x_i and y_i are the individual sample points indexed with i, $\bar{x}$ and $\bar{y}$ are the means of the samples X and Y, respectively. The coefficient ρ ranges from -1 to 1. A value of 1 implies a perfect positive linear relationship, -1 implies a perfect negative linear relationship, and 0 implies no linear relationship between the variables. And c represents the category ID.

4.4. Results

4.4.1 Comparison of different features.

First, we compared the results of different unimodal features based on the official validation set of Hume-Vidmimic2, which are shown in Table 1. For the visual modality, it can be observed that the EmoViT performance, which we pre-trained on a large-scale face dataset, significantly surpasses the official ViT features, achieving a ρ of 0.1685. Subsequently, we attempted to fuse the ViT and EmoViT features, but this resulted in a decrease in performance. For the audio modality, the official Wav2Vec2 features significantly outperform the HuBERT features, achieving a ρ of 0.2576. However, the fusion of the two also resulted in a decrease in performance.

In addition to audio and visual modalities, we also utilized text features for estimating emotion intensity. Specifically, we employed ChatGLM3 [16, 83] to extract text features, which is a generation of pre-trained dialogue models jointly released by Zhipu AI and Tsinghua KEG. We attempted to use different layers of hidden states in ChatGLM3 and found that the 21st layer produced the best results. We also tried fusing features from the 21st and 28th layers, which led to a slight improvement, with the ρ index reaching 0.4879. Furthermore, inspired by previous

Table 4. Final competition results of the Emotional Mimicry Intensity Estimation Challenge. The ρ is evaluated on the test set of the Hume-Vidmimic2 test set.

Rank	Teams	ρ
#1	**Netease Fuxi AI Lab**	**0.7185**
#2	HCAI-VIS [61]	0.5536
#3	USTC-IAT-United [75]	0.3594
#4	HSEmotion [60]	0.3316
-	baseline [43]	0.48

work [74], we attempt to first finetune ChatGLM3 on this task and then extract features for training. However, we find that performance slightly decreased.

Finally, we select the EmoViT feature, HuBERT feature, and the 21st and 28th layers of ChatGLM3 as unimodal features for training a multimodal network, achieving a ρ of 0.4931. We find that in this challenge, the text modality's features are dominant, outperforming the other two modalities significantly. Even with the fusion of multimodal features, performance was only slightly better than that of the single text modality. We believe this is related to the data collection method of Hume-Vidmimic2, which requires subjects to imitate a seed video. Often, subjects cannot accurately replicate the facial expressions and tone of voice from the seed video, but they can generally reproduce the spoken words quite well. Therefore, in this task, the importance of the text modality far exceeds the other two modalities.

4.4.2 Comparison of different settings.

As can be seen in Table 2, we also conduct extensive experiments with different settings to further investigate the effectiveness of our used components, including Temporal Augment, EMA, Triplet Loss and different fusion strategy.

4.4.3 Validation results

To further enhance the generalization ability and test the models' performance, we use 5-fold cross-validation to train multiple models and then ensemble them. We combine the training and validation set of the Hume-Vidmimic2 dataset. Then we split them into 5 folds randomly train the model on 4 folds of them and take the rest on as the validation set. The results can be found in Table 3.

4.4.4 Competition results

For Emotional Mimicry Intensity (EMI) Estimation Challenge, we need to predict the intensity of 6 predefined emotions of the videos from the test set of Hume-Vidmimic2. The results can be seen in Table 4. Our method achieves a average ρ of 0.7185 and wins the first place in the EMI Estimation Challenge.

5. Conclusion

In this work, we propose a multi-modal framework for emotional mimicry intensity estimation. We explore multiple effective features for different modalities and incorporate temporal augment module to improve the model's generalization ability. Additionally, we find that text features are the most important for this task across different modalities. Therefore, we introduced contrastive learning to refine the extracted multimodal features. Our method shows superior performance and win the first place in the Emotional Mimicry Intensity (EMI) Estimation Challenge of ABAW6.

6. Acknowledgments

The experiments and the data management and storage are supported by NetEase Fuxi Youling platform, based on Fuxi Agent-Oriented Programming (AOP) system that is carefully designed to facilitate task modeling. This work is also supported by the National Key R&D Program of China (No. 2022YFF09022303).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 1, 2, 4

[4] Janet B Bavelas, Alex Black, Charles R Lemery, and Jennifer Mullett. " i show how you feel": Motor mimicry as a communicative act. *Journal of personality and social psychology*, 50(2):322, 1986. 2

[5] Sylvie Blairy, Pedro Herrera, and Ursula Hess. Mimicry and the judgment of emotional facial expressions. *Journal of Nonverbal behavior*, 23:5–41, 1999. 2

[6] P Bourgeois and U Hess. Emotional reactions to political leaders facial displays: a replication. In *Psychophysiology*, pages S36–S36. CAMBRIDGE UNIV PRESS 40 WEST 20TH STREET, NEW YORK, NY 10011-4211 USA, 1999. 2

[7] Jean-Pierre Briot and François Pachet. Deep learning for music generation: challenges and directions. *Neural Computing and Applications*, 32(4):981–993, 2020. 2

[8] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. Deep learning techniques for music generation–a survey. *arXiv preprint arXiv:1709.01620*, 2017. 2

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[10] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 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. 2

[11] Lukas Christ, Shahin Amiriparian, Alice Baird, Alexander Kathan, Niklas Müller, Steffen Klug, Chris Gagne, Panagiotis Tzirakis, Lukas Stappen, Eva-Maria Meßner, et al. The muse 2023 multimodal sentiment analysis challenge: Mimicked emotions, cross-cultural humour, and personalisation. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, pages 1–10, 2023. 5

[12] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 5

[13] Muhterem Dindar, Sanna Järvelä, Sara Ahola, Xiaohua Huang, and Guoying Zhao. Leaders and followers identified by emotional mimicry during collaborative learning: A facial expression recognition study on emotional valence. *IEEE Transactions on Affective Computing*, 13(3):1390–1400, 2020. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[15] Heming Du, Xin Yu, and Liang Zheng. Learning object relation graph and tentative policy for visual navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 19–34. Springer, 2020. 3

[16] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021. 3, 6
[17] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022. 1, 2, 4
[18] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 2
[19] Xiaoyu Feng, Heming Du, Hehe Fan, Yueqi Duan, and Yongpan Liu. Seformer: Structure embedding transformer for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 632–640, 2023. 3
[20] Chiara Filippini, David Perpetuini, Daniela Cardone, Antonio Maria Chiarelli, and Arcangelo Merla. Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: A review. *Applied Sciences*, 10(8):2924, 2020. 1
[21] Sigmund Freud. *Group psychology and the analysis of the ego*. WW Norton & Company, 1975. 2
[22] Riccardo Gervasi, Federico Barravecchia, Luca Mastrogiacomo, and Fiorenzo Franceschini. Applications of affective computing in human-robot interaction: State-of-art and challenges for manufacturing. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 237(6-7):815–832, 2023. 1
[23] Brooks B Gump and James A Kulik. Stress, affiliation, and emotional contagion. *Journal of personality and social psychology*, 72(2):305, 1997. 2
[24] Elaine Hatfield, John T Cacioppo, and Richard L Rapson. Emotional contagion. *Current directions in psychological science*, 2(3):96–100, 1993. 2
[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 3
[26] Ursula Hess and Sylvie Blairy. Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International journal of psychophysiology*, 40(2):129–141, 2001. 2
[27] Martin L Hoffman. Interaction of affect and cognition in empathy. *Emotion, cognition, and behavior*, pages 103–131, 1984. 2
[28] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 2
[29] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7680–7689, 2021. 3
[30] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint arXiv:2105.03588*, 2021. 2
[31] MD Wahiduzzaman Khan, Hongwei Sheng, Hu Zhang, Heming Du, Sen Wang, Minas Coroneo, Farshid Hajati, Sahar Shariflou, Michael Kalloniatis, Jack Phu, et al. Rvd: a handheld device-based fundus video dataset for retinal vessel segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 2
[32] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018. 2
[33] 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
[34] 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.
[35] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1
[36] 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.
[37] 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.
[38] 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
[39] 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. 1
[40] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020.
[41] 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
[42] 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
[43] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Chunchang Shao, and Guanyu Hu. The 6th

affective behavior analysis in-the-wild (abaw) competition. *arXiv preprint arXiv:2402.19344*, 2024. 1, 6

[44] John T Lanzetta and Basil G Englis. Expectations of cooperation and competition and their effects on observers' vicarious emotional responses. *Journal of personality and social psychology*, 56(4):543, 1989. 2

[45] Theodor Lipps. Das wissen von fremden ichen. *Psychologische untersuchungen*, 4:694, 1905. 2

[46] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *arXiv preprint arXiv:2308.10175*, 2023. 2

[47] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7590–7598, 2023. 2

[48] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–4. IEEE, 2019. 2

[49] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 3

[50] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[52] Gregory J McHugo, John T Lanzetta, and Lauren K Bush. The effect of attitudes on emotional reactions to expressive displays of political leaders. *Journal of Nonverbal Behavior*, 15(1):19–41, 1991. 2

[53] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. 1, 3

[54] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *arXiv preprint arXiv:2305.18891*, 2023. 2

[55] Feng Qiu, Bowen Ma, Wei Zhang, and Yu Ding. Multimodal emotion reaction intensity estimation with temporal augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5776–5783, 2023. 2

[56] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 4, 5

[57] Minglun Ren, Nengying Chen, and Hui Qiu. Human-machine collaborative decision-making: An evolutionary roadmap based on cognitive intelligence. *International Journal of Social Robotics*, 15(7):1101–1114, 2023. 1

[58] Carl R Rogers. The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology*, 21(2):95, 1957. 2

[59] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018. 3

[60] Elena Ryumina, Maxim Markitantov, Dmitry Ryumin, Heysem Kaya, and Alexey Karpov. Audio-visual compound expression recognition method based on late modality fusion and rule-based decision. *arXiv preprint arXiv:2403.12687*, 2024. 6

[61] Andrey V Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction. *arXiv preprint arXiv:2403.11590*, 2024. 6

[62] Nicu Sebe, Ira Cohen, and Thomas S Huang. Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision*, pages 387–409. World Scientific, 2005. 2

[63] Xin Shen, Shaozu Yuan, Hongwei Sheng, Heming Du, and Xin Yu. Auslan-daily: Australian sign language translation for daily communication and news. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[64] Hongwei Sheng, Xin Yu, Feiyu Wang, MD Wahiduzzaman Khan, Hexuan Weng, Sahar Shariflou, and S Mojtaba Golzan. Autonomous stabilization of retinal videos for streamlining assessment of spontaneous venous pulsations. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023. 2

[65] Elaine V Siegel. Psychoanalytic dance therapy: The bridge between psyche and soma. *American Journal of Dance Therapy*, 17(2):115–128, 1995. 2

[66] Boštjan Šumak, Saša Brdnik, and Maja Pušnik. Sensors and artificial intelligence methods and algorithms for human–computer intelligent interaction: A systematic mapping study. *Sensors*, 22(1):20, 2021. 1

[67] Martina Szabóová, Martin Sarnovskỳ, Viera Maslej Krešňáková, and Kristína Machová. Emotion analysis in human–robot interaction. *Electronics*, 9(11): 1761, 2020. 1

[68] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 3

[69] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[70] Giovanna Varni, Isabelle Hupont, Chloe Clavel, and Mohamed Chetouani. Computational study of primitive emotional contagion in dyadic interactions. *IEEE Transactions on Affective Computing*, 11(2):258–271, 2017. 2

[71] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment

networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 4

[72] Qingzheng Xu. Analyzing music effects on brain waves according to functional measures. 2

[73] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 3

[74] Guofeng Yi, Yuguang Yang, Yu Pan, Yuhang Cao, Jixun Yao, Xiang Lv, Cunhang Fan, Zhao Lv, Jianhua Tao, Shan Liang, et al. Exploring the power of cross-contextual large language model in mimic emotion prediction. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, pages 19–26, 2023. 3, 6

[75] Jun Yu, Jichao Zhu, and Wangyuan Zhu. Compound expression recognition via multi model ensemble. *arXiv preprint arXiv:2403.12572*, 2024. 6

[76] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *European conference on computer vision*, pages 318–333. Springer, 2016. 2

[77] Xin Yu and Fatih Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[78] Xin Yu and Fatih Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3760–3768, 2017.

[79] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–233, 2018.

[80] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 908–917, 2018.

[81] Xin Yu, Fatemeh Shiri, Bernard Ghanem, and Fatih Porikli. Can we see more? joint frontalization and hallucination of unaligned tiny faces. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2148–2164, 2019. 2

[82] 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

[83] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 1, 2, 4, 6

[84] Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2428–2437, 2022. 1, 2

[85] Wei Zhang, Lincheng Li, Yu Ding, Wei Chen, Zhigang Deng, and Xin Yu. Detecting facial action units from global-local fine-grained expressions. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[86] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th abaw competition. *arXiv preprint arXiv:2303.10849*, 2023.

[87] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tiancheng Guo, and Xin Yu. Affective behaviour analysis via integrating multi-modal knowledge. *arXiv preprint arXiv:2403.10825*, 2024. 2