Evaluating Vision Language Models in Detecting Learning Engagement

Jayant Teotia, Xulang Zhang, Rui Mao and Erik Cambria

Nanyang Technological University, Singapore
jayant002@e.ntu.edu.sg; {xulang.zhang, rui.mao, cambria}@ntu.edu.sg

***Abstract*—With the advancement of both computer vision and natural language processing, there is growing interest in incorporating Vision Language Models (VLMs) into the classroom to empower students and educators. Despite the VLMs' sophisticated abilities in context-aware emotion recognition, their effectiveness in detecting classroom-specific emotions, e.g., engagement, distraction, and absent-mindedness, remains underexplored. As such, this paper aims to investigate the capabilities of two state-of-the-art VLMs in this domain through an empirical study, focusing on two research questions: 1) Is learning engagement detection more challenging for VLMs compared to conventional emotion detection? 2) What are the key difficulties faced by VLMs in processing learning engagement detection tasks? To address these questions, we perform a series of evaluation experiments by utilizing a classroom behavior detection dataset and an emotion recognition dataset. We conclude that VLMs that perform well on basic emotion recognition struggle with in-context engagement detection, due to the nuanced and context-dependent nature of the task. Specifically, experiments show that VLMs have difficulty distinguishing engaged and distracted classroom behavior, e.g., reading versus bowing the head. It suggests that VLMs still have significant room for improvement in engagement analysis. This issue can potentially be addressed by incorporating more classroom-specific training data or commonsense reasoning frameworks.**

***Index Terms*—vision language models, emotion recognition, learning engagement detection, multimodal learning.**

I. Introduction

With the monumental development of Large Language Models (LLM), Vision Language Models (VLMs), which extend the LLMs by integrating visual inputs, have gained popularity in a variety of application fields. One of such scenarios is pedagogy [1]. There are a number of works that utilize computer vision models in the classroom setting to detect students' behaviour [2], [3]. With their promising visual understanding abilities, VLMs can be implemented to automatically monitor the mental and cognitive states of students in the classroom, providing support for students' engagement and well-being.

However, despite the recent advancement in multimodal emotion recognition [4]–[7], current VLMs might not be adept at recognizing emotion states relevant to the classroom setting, e.g., engagement, distraction, absent-mindedness, etc. This is because such emotional states are more contextual, requiring the machine to understand not only facial expressions, but also body language and classroom-related activities.

To the best of our knowledge, existing works have not extensively studied whether the prevalent VLMs are competent at detecting classroom-related emotion states, which we define as the learning engagement detection task in this paper. Therefore, the objective of the paper is to conduct an empirical study to answer the following two research questions:

1) Compared to conventional visual emotion detection tasks, are learning engagement detection tasks more difficult for VLMs?
2) What are the most challenging factors for VLMs to process learning engagement detection tasks?

To this end, we conduct a set of evaluation experiments, utilizing a classroom behavior detection dataset [8] and an emotion recognition dataset [9], [10]. We repurpose these datasets into binary classification tasks, i.e., engaged and non-engaged, positive and negative, to evaluate the abilities of two publicly available VLMs to detect engagement and conventional emotions. We further conduct multi-class classification evaluations on the original dataset labels to analyze what VLMs find challenging to identify (Fig. 1). Through the experiments, we find that learning engagement detection tasks pose greater challenges for VLMs as they require interpreting the interaction between individuals and their surrounding environment, unlike emotion detection, which primarily focuses on the individual. The limited availability of relevant examples for analyzing learning engagement in real world, unlabeled corpora likely contribute to the inconsistent performance of VLMs across different engagement label classes. This is because, unlike emotions, engagement analysis depends largely on commonsense reasoning rather than visual and textual cues [11]. However, humans often omit explicit mention of commonsense knowledge when it appears too obvious. The contribution of this work can be summarized as two-fold:

1) We provide a systematic empirical evaluation of two state-of-the-art VLMs in the context of detecting learning engagement, addressing a gap in the existing literature regarding their effectiveness in classroom-specific emotion recognition tasks.
2) We identify and discuss the specific challenges that VLMs face when processing learning engagement detection tasks compared to conventional emotion detection, and suggest potential avenues for improving VLM performance in learning engagement analysis.

2375-9259/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDMW65004.2024.00069

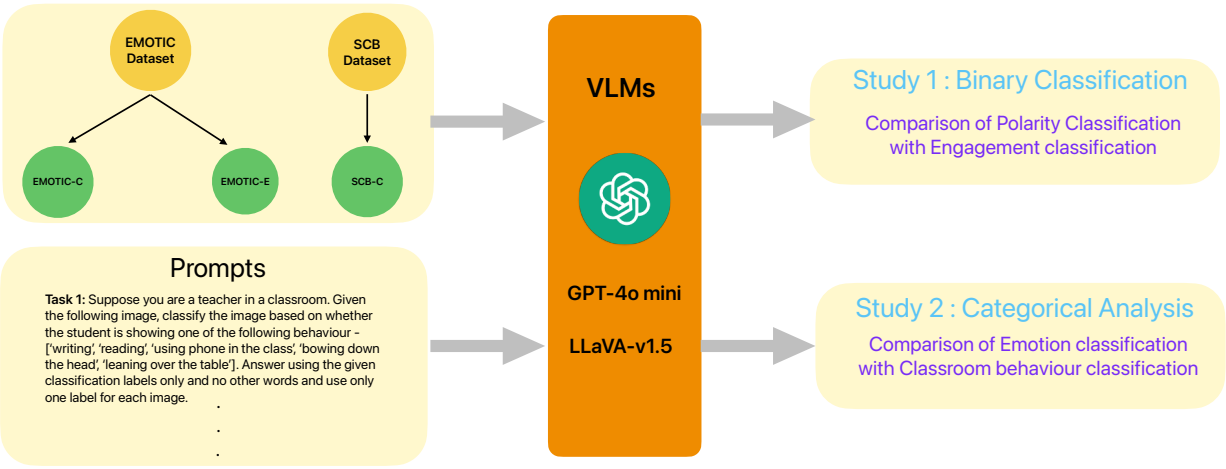

Fig. 1. Workflow of our evaluations.

II. Related Work

A. Emotion Recognition

Emotion recognition encompasses a variety of task settings, each with its own set of challenges [12]. For text-based emotion recognition, the research community is concerned with sophisticated problems such as bias [13], [14], commonsense [15], cognition [16]–[18], ambiguity [19], [20], crosslingual and multilingual [21], few-shot instances [22], and multitask settings [23], discourse relations [24], domain-adaptation [25], etc. Speech-based emotion recognition deals with intonation, stress, speaking speed, and accent [26]–[28]. In image-based emotion recognition, some of the challenges include contextual understanding [4], [7], [29], and cultural differences [30].

In multimodal emotion recognition, there are various combinations of modalities, including text-image, text-audio, video-audio, text-video, etc [31], [32]. Research in this field mainly focuses on the fusion of features from different modalities, e.g., tensor fusion [33], [34], hierarchical fusion [35], [36], bi-modal fusion [37], [38], attention-based fusion [39]–[41], etc. As VLMs are making impressive progress, the research focus of text-image emotion recognition has shifted to more complex scenarios, emotion recognition in context [9], [42], [43] being one of the challenging tasks.

B. Vision Language Model

VLMs can be divided into three categories based on training paradigm, namely, contrastive-based, masking-based [44], and pretrained-backbone-based. The most commonly used contrastive-based method is Contrastive Language–Image Pre-training (CLIP) [45], by which the model learns to map vision and language in a shared vector space. Due to its robustness, this approach has been extended by a variety of studies, e.g., BLIP [46], SigLIP [47], and Llip [48].

Masking, on the other hand, takes inspiration from masked language modeling [49] and learns image representations by reconstructing masked image patches given some unmasked text, and vice versa for text representations [50]. With the advancement of LLMs, however, it is also possible to forgo such expensive training, and learn a mapping between a pretrained image encoder and an open-source LLM instead. Such an approach is much less computationally demanding than training from scratch. VLMs of this type includes Frozen [51], MiniGPT [52], GPT-4 [53], Gemini [54], and LLaVA [55]. VLMs have been widely used in the field of affective computing [56]–[59]. Specifically, many existing works leverage VLM for facial emotion recognition [60], [61], demonstrating impressive generalization and zero-shot abilities. However, the application of VLMs for in-context emotion recognition is relatively under-explored. Zhang et al. [4] propose EmotionCLIP, which extends CLIPby extracting emotion representations from verbal and nonverbal communication. Xenos et al. [7] leverage the generative capabilities of LLaVA [55] for context-aware emotion recognition by employing a two-stage approach to make use of its commonsense reasoning knowledge, achieving impressive zero-shot performance. However, to the best of our knowledge, existing studies have not explored VLMs' ability to recognize classroom-specific emotions, i.e., learning engagement detection.

III. Methodology

In this section, we will introduce how we prompt the VLMs for our evaluations on three datasets (introduced in Section IV-A). The prompts are constructed by following common practices in the field of prompt engineering for VLMs [62]. For the binary classification experiments (shown in Table II), we use the following prompts to evaluate the VLMs' performance on the different tasks, i.e., learning engagement detection, sentiment polarity classification, and general engagement detection.

Task: Learning engagement detection
Prompt: `Suppose you are a teacher. Given the following image, identify if the student is concentrating on the studies or not. Answer in 'yes' or 'no' only, 'yes' if the student is concentrating.`

Task: Sentiment polarity classification
Prompt: `Suppose you are an emotional analyst. Given the following image, identify its only label from the label set - ['positive', 'negative'].Answer by only giving the label classification and nothing else.`

Task: General engagement detection
Prompt: `Suppose you are an emotional analyst. Given the following image, identify its only label from the label set - ['disconnection', 'engagement'].Answer by only giving the label classification and nothing else.`

For the multiclass classification experiments (shown in Section V-B), we use the following prompts to evaluate the VLMs' performance on the different tasks, i.e., classroom behaviour classification and emotion recognition.

Task: Learning behavior recognition
Prompt: `Suppose you are a teacher in a classroom. Given the following image, classify the image based on whether the student is showing one of the following behaviour - ['writing', 'reading', 'using phone in the class', 'bowing down the head', 'leaning over the table']. Answer using the given classification labels only and no other words and use only one label for each image.`

Task: Emotion recognition
Prompt: `Suppose you are an emotional analyst. Given the following image, identify its only emotion label from the label set ['affection', 'pleasure', 'anger', 'sadness', 'fear']. Answer by only giving the label classification and nothing else.`

IV. Experiment

To evaluate VLMs ability to handle conventional and classroom-related emotion states, we first compare the performance of two VLMs, i.e., LLaVA and GPT, on the binary classification tasks of sentiment polarity detection, general engagement detection, and learning engagement detection. Furthermore, we evaluate the two VLMs' performance on basic emotion recognition and learning behavior recognition to analyze what are the challenging factors in learning engagement detection.

A. Datasets

As shown in Fig. 2, we conducted our experiments by repurposing two benchmark datasets into the following three datasets for different evaluation scenarios. Details of the datasets can be found in Table IV-A.

1) The SCB-C dataset originates from the validation set of SCB-dataset3 [8], containing images with 5 categories of classroom behaviors, including *reading*, *writing*, *using phone*, *bowing the head*, and *leaning over the table*. For the learning engagement detection task, we combine the labels *reading* and *writing* into *engaged*, and *using phone*, *bowing the head*, and *leaning over the table* as *non-engaged*.

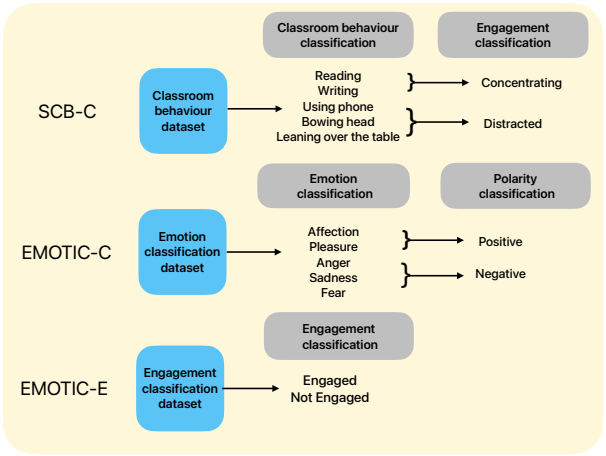

Fig. 2. Illustration of the repurposed datasets.

TABLE I
NUMBER OF EVALUATION SAMPLES IN EACH DATASET USED IN THIS PAPER. BLUE DENOTES POSITIVE/ENGAGED; RED DENOTES NEGATIVE/NON-ENGAGED.

	SCB-C	EMOTIC-C	EMOTIC-E
# reading	782	-	-
# writing	329	-	-
# using phone	631	-	-
# bowing head	94	-	-
# leaning over table	99	-	-
# affection	-	253	-
# pleasure	-	982	-
# anger	-	76	-
# sadness	-	168	-
# fear	-	84	-
# engagement	-	-	738
# disconnection	-	-	272
# total	1,935	1,563	1,010

2) The EMOTIC-C dataset originates from the testing set of the EMOTIC dataset [9], which contains images with people in real environments, annotated with their apparent emotions. To make the comparison fair, we select samples from 5 out of 26 the emotion categories, i.e., *pleasure*, *affection*, *sadness*, *fear*, and *anger*, to form the EMOTIC-C dataset for multiclass classification, which contains 1,563 images in total. For the sentiment polarity classification task, we regard *pleasure* and *affection* as *positive*, and *sadness*, *fear*, and *anger* as *negative*.
3) The EMOTIC-E dataset also originates from the testing set of the EMOTIC dataset. We select samples from 2 of the emotion categories, engagement and disconnection, to form the EMOTIC-E dataset, which contains 1,010 images in total.

Note that the original datasets, i.e., SCB-dataset3 and EMOTIC, have multiple targets in a single image. For the purpose of fair evaluation in our experiments, we utilize bounding boxes given by the datasets and convert all included samples into single-target images for classification.

B. VLM Setup

In this paper, we conduct our evaluations on two VLMs, namely, LLaVA-v1.5[1] and GPT-4o mini[2]. For each experiment, the images are resized to 128x128. The inference for GPT was done using the online GPT-4o-mini API, while LLaVA-v1.5 was implemented offline on Nvidia A100 GPU.

V. Results

A. Binary Classification

We first compare the VLMs' binary classification performance on sentiment polarity (EMOTIC-C), general engagement (EMOTIC-E), and learning engagement (SCB-C). As seen in Table II, both VLMs are well capable of understanding the sentiment polarities from visual data, achieving higher accuracy than the other two engagement evaluation tasks. Specifically, the GPT model yields 99.48% accuracy, showing that GPT-4o mini's visual-linguistic understanding is well-developed in common sentimental analysis scenarios. On the other hand, we can also observe that the performance of both VLMs drops sharply on the engagement analysis tasks. Accuracy gaps larger than 15% can be observed when comparing the two models on sentiment and engagement analysis tasks, respectively. It indicates that VLMs have difficulties in understanding these images in an engagement context. Considering the nearly perfect performance of the GPT model (99.48% on EMOTIC-C), such a big drop suggests that engagement analysis may involve the understanding of more abstract, context-dependent, and nuanced human behaviors. Engagement analysis extends beyond merely interpreting affective signals from individuals, as seen in sentiment analysis. It requires an understanding of the dynamic interaction between individuals and their surroundings, e.g., if someone is focused on their phone, it may indicate high engagement with the device. However, in contexts such as a classroom or meeting, the same behavior could suggest disengagement, as the individual is expected to focus on the speaker or presentation.

Therefore, engagement analysis necessitates a more advanced level of contextual reasoning, one that can assess the appropriateness of an individual's behavior relative to the specific environment, social expectations, and situational dynamics. This added complexity often proves challenging for models designed primarily for simpler tasks like sentiment detection. Furthermore, engagement is inherently multifaceted, shaped by factors that may not be directly observable, such as cognitive states, personal interests, or underlying motivations. For example, a neutral expression in a professional context could signify intense focus or, alternatively, boredom, depending on the situation and the individual's engagement with the task [63]. This highlights the need for models to integrate not only visual data but also contextual information, temporal patterns, and external factors such as task relevance or social interactions, in order to provide a more accurate assessment of engagement [64].

[1]https://llava-vl.github.io

[2]https://platform.openai.com/docs/models/gpt-4o-mini

TABLE II
Binary evaluation results, evaluated by Accuracy.

Model	LLaVA	GPT
SCB-C	67.23%	72.65%
EMOTIC-C	82.79%	99.48%
EMOTIC-E	40.59%	71.38%

B. Categorical Analysis

TABLE III
Behavior classification results on SCB-C, evaluated by Accuracy.

Label	LLaVA	GPT
reading	11.13%	18.49%
writing	95.44%	90.58%
using phone	39.30%	57.89%
bowing the head	0.00%	28.72%
leaning over the table	0.00%	56.56%
average	29.17%	50.45%

TABLE IV
Emotion classification results on EMOTIC-C, evaluated by Accuracy.

Label	LLaVA	GPT
affection	22.53%	47.43%
pleasure	37.47%	62.73%
anger	23.68%	40.79%
sadness	93.45%	51.19%
fear	17.86%	20.23%
average	39.00%	44.48%

By investigating the results on specific label classes, VLM performance varies widely in their ability to understand images with different types of labels. First, as seen in Table III, both LLaVA and GPT yield the highest accuracy in "writing" on the SCB-C dataset, exceeding the performance on other classes significantly. However, LLaVA utterly fails to correctly identify "bowing the head" and "leaning over the table". On the other hand, "reading" and "bowing the head" classes are challenging for GPT-4o mini. Such results imply that learning these labels from real-world unlabeled data by pre-training is still challenging. This is likely because the actions of "bowing the head", "leaning over the table", and "reading" are highly dependent on their environmental and situational context, which may be difficult for a VLM to accurately interpret. For example, "bowing the head" could take place while reading, but without explicit contextual indicators, such as holding and focusing on a book, the model may struggle to correctly classify the action.

Compared to the behavior classification results on SCB-C in Table III, the performance gaps between the highest and lowest accuracy labels in the EMOTIC-C dataset (emotion classification) appear to be narrower (see Table IV). This is likely due to the greater number of learning instances for emotion detection during the pre-training phase compared to those available for in-class behavior classification.

VLMs have more opportunities to learn the connection between textual descriptions of emotions and their corresponding visual representations from the pre-training corpora, e.g., a person says "I am feeling sad" with a sad face in a video, making their performance on traditional emotion classification tasks relatively robust. In contrast, text describing in-class behaviors is less common in real-world training data, contributing to the wider performance gaps in behavior classification. The learning of rare cases is a significant challenge for pre-trained models [65].

C. Summary

Compared to conventional visual emotion detection tasks, learning engagement detection tasks are more difficult for VLMs. This is primarily because engagement analysis relies on interpreting the interaction between individuals and their surrounding environment, whereas emotion detection is largely focused on the individual alone. Consequently, for effective engagement detection, VLMs must capture a wider array of contextual signals that extend beyond the person's emotional state.

The most challenging factor for VLMs to process learning engagement detection tasks arises from the scarcity of relevant examples in real-world, unlabeled corpora. Unlike emotions, which are often expressed through both visual and textual cues, engagement analysis relies heavily on commonsense understanding. Since humans tend to omit commonsense information when it is perceived as obvious, e.g., "I am bowing the head" or "he is leaning over the table", it is rare to find parallel examples indicating engagement in unlabeled data. This lack of explicit training instances may contribute to the underperformance of VLMs in detecting learning engagement.

VI. Conclusion

In this work, we evaluated two publicly available VLMs for their ability to handle the learning engagement detection tasks. We repurposed an in-context emotion recognition dataset and a classroom behavior dataset to perform polarity classification and engagement classification, drawing the conclusion that the engagement analysis is more difficult than conventional emotion detection tasks for VLMs. Furthermore, we conduct multiclass classification using the emotion and behavior labels of these datasets respectively. From the experiment results, we find that the performance of VLMs is more robust on emotion detection tasks than the performance on detecting in-classroom behaviors, indicating a research gap for classroom engagement detection.

Future research could take multiple directions to enhance VLM performance in engagement detection tasks. One promising approach is to introduce more diverse and contextually rich datasets specifically designed for engagement analysis. Expanding the training data with instances that represent a wider variety of behaviors and environmental interactions may help reduce the current performance gap.

Additionally, incorporating commonsense reasoning frameworks and external knowledge sources could improve VLMs' ability to infer engagement by compensating for the absence of explicit contextual details in real-world data. Future studies should also focus on refining model architectures and developing specialized systems capable of addressing the complex relationship between individuals and their environments across various learning contexts.

References

[1] H. Singh and S. J. Miah, "Smart education literature: A theoretical analysis," *Education and Information Technologies*, vol. 25, no. 4, pp. 3299–3328, 2020.

[2] H. Chen, G. Zhou, and H. Jiang, "Student behavior detection in the classroom based on improved yolov8," *Sensors*, vol. 23, no. 20, p. 8385, 2023.

[3] D. Zhang, C. Chen, F. Tan, B. Qian, W. Li, X. He, and S. Lei, "Multi-view and multi-scale behavior recognition algorithm based on attention mechanism," *Frontiers in Neurorobotics*, vol. 17, p. 1276208, 2023.

[4] S. Zhang, Y. Pan, and J. Z. Wang, "Learning emotion representations from verbal and nonverbal communication," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 993–19 004.

[5] N. Wagner, F. Mätzler, S. R. Vossberg, H. Schneider, S. Pavlitska, and J. M. Zöllner, "Cage: Circumplex affect guided expression inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4683–4692.

[6] M. S. U. Khan, M. F. Naeem, F. Tombari, L. Van Gool, D. Stricker, and M. Z. Afzal, "Focusclip: Multimodal subject-level guidance for zero-shot transfer in human-centric tasks," *arXiv preprint arXiv:2403.06904*, 2024.

[7] A. Xenos, N. M. Foteinopoulou, I. Ntinou, I. Patras, and G. Tzimiropoulos, "Vllms provide better context for emotion understanding through common sense reasoning," *arXiv preprint arXiv:2404.07078*, 2024.

[8] F. Yang and T. Wang, "Scb-dataset3: A benchmark for detecting student classroom behavior," *arXiv preprint arXiv:2310.02522*, 2023.

[9] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotic: Emotions in context dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 61–69.

[10] Y. Susanto, A. Livingstone, B. C. Ng, and E. Cambria, "The Hourglass Model revisited," *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.

[11] E. Cambria, "Understanding natural language understanding," *Springer, ISBN 978-3-031-73973-6*, 2024.

[12] E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, "Seven pillars for the future of artificial intelligence," *IEEE Intelligent Systems*, vol. 38, no. 6, pp. 62–69, 2023.

[13] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1743–1753, 2023.

[14] X. Zhang, R. Mao, and Cambria, "SenticVec: Toward robust and human-centric neurosymbolic sentiment analysis," in *Proceedings of ACL*, 2024, pp. 4851–4863.

[15] E. Cambria, X. Zhang, R. Mao, M. Chen, and K. Kwok, "SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing," in *International Conference on Human-Computer Interaction (HCII)*, Washington DC, USA, 2024.

[16] R. Mao, K. Du, Y. Ma, L. Zhu, and E. Cambria, "Discovering the cognition behind language: Financial metaphor analysis with MetaPro," in *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2023, pp. 1211–1216.

[17] R. Mao, T. Zhang, Q. Liu, A. Hussain, and E. Cambria, "Unveiling diplomatic narratives: Analyzing United Nations Security Council debates through metaphorical cognition," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, vol. 46, 2024, pp. 1709–1716.

[18] R. Mao, Q. Lin, Q. Liu, G. Mengaldo, and E. Cambria, "Understanding public perception towards weather disasters through the lens of metaphor," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, (IJCAI-24)*. International Joint Conferences on Artificial Intelligence Organization, 2024, pp. 7394–7402.

[19] X. Zhang, R. Mao, K. He, and E. Cambria, "Neuro-symbolic sentiment analysis with dynamic word sense disambiguation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 8772–8783.

[20] L. Zhu, R. Mao, E. Cambria, and B. J. Jansen, "Neurosymbolic AI for personalized sentiment analysis," in *Proceedings of International Conference on Human-Computer Interaction (HCII)*, Washington DC, USA, 2024.

[21] X. Zhang, R. Mao, and E. Cambria, "Multilingual emotion recognition: Discovering the variations of lexical semantics between languages," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024.

[22] K. He, R. Mao, T. Gong, C. Li, and E. Cambria, "Meta-based self-training and re-weighting for aspect-based sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1731–1742, 2023.

[23] R. Mao and X. Li, "Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 534–13 542.

[24] W. Li, L. Zhu, R. Mao, and E. Cambria, "SKIER: A symbolic knowledge integrated model for conversational emotion recognition," *Proceedings of the AAAI Conference on Arificial Intelligence*, vol. 37, no. 11, pp. 13 121–13 129, 2023.

[25] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–42, 2024.

[26] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6907–6911.

[27] C. Min, H. Lin, X. Li, H. Zhao, J. Lu, L. Yang, and B. Xu, "Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective," *Information Fusion*, vol. 96, pp. 214–223, 2023.

[28] M. Agarla, S. Bianco, L. Celona, P. Napoletano, A. Petrovsky, F. Piccoli, R. Schettini, and I. Shanin, "Semi-supervised cross-lingual speech emotion recognition," *Expert Systems with Applications*, vol. 237, p. 121368, 2024.

[29] Q. Xu, Y. Wei, S. Yuan, J. Wu, L. Wang, and C. Wu, "Learning emotional prompt features with multiple views for visual emotion analysis," *Information Fusion*, vol. 108, p. 102366, 2024.

[30] M. Sun and A. S. Lau, "Exploring cultural differences in expressive suppression and emotion recognition," *Journal of cross-cultural psychology*, vol. 49, no. 4, pp. 664–672, 2018.

[31] C. Fan, J. Lin, R. Mao, and E. Cambria, "Fusing pairwise modalities for emotion recognition in conversations," *Information Fusion*, vol. 106, p. 102306, 2024.

[32] E. Cambria, N. Howard, J. Hsu, and A. Hussain, "Sentic blending: Scalable multimodal fusion for continuous interpretation of semantics and sentics," in *IEEE SSCI*, 2013, pp. 108–117.

[33] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017, pp. 1114–1125.

[34] E. J. Barezi, P. Momeni, and P. Fung, "Modality-based factorization for multimodal fusion," *arXiv preprint arXiv:1811.12624*, 2018.

[35] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 481–492.

[36] J. Tang, Z. Ma, K. Gan, J. Zhang, and Z. Yin, "Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment," *Information Fusion*, vol. 103, p. 102129, 2024.

[37] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proceedings of the 2021 international conference on multimodal interaction*, 2021, pp. 6–15.

[38] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, and Y. Huang, "Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowledge-Based Systems*, vol. 235, p. 107676, 2022.

[39] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017, pp. 1033–1038.

[40] C. Xi, G. Lu, and J. Yan, "Multimodal sentiment analysis based on multi-head attention mechanism," in *Proceedings of the 4th international conference on machine learning and soft computing*, 2020, pp. 34–39.

[41] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional lstm," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13 059–13 076, 2021.

[42] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 143–10 152.

[43] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "Arbee: Towards automated recognition of bodily expression of emotion in the wild," *International journal of computer vision*, vol. 128, pp. 1–25, 2020.

[44] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman *et al.*, "An introduction to vision-language modeling," *arXiv preprint arXiv:2405.17247*, 2024.

[45] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[46] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[47] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.

[48] S. Lavoie, P. Kirichenko, M. Ibrahim, M. Assran, A. G. Wildon, A. Courville, and N. Ballas, "Modeling caption diversity in contrastive vision-language pretraining," *arXiv preprint arXiv:2405.00740*, 2024.

[49] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[50] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto, "Masked vision and language modeling for multi-modal representation learning," *arXiv preprint arXiv:2208.02131*, 2022.

[51] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 200–212, 2021.

[52] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[53] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[54] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[55] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[56] J. Mu, F. Nie, W. Wang, J. Xu, J. Zhang, and H. Liu, "Mocolnet: A momentum contrastive learning network for multimodal aspect-level sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[57] F. Zhao, C. Li, Z. Wu, Y. Ouyang, J. Zhang, and X. Dai, "M2df: Multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9057–9070.

[58] S. Yan, N. Dong, L. Zhang, and J. Tang, "Clip-driven fine-grained text-image person re-identification," *IEEE Transactions on Image Processing*, 2023.

[59] L. Xiao, X. Wu, J. Xu, W. Li, C. Jin, and L. He, "Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis," *Information Fusion*, p. 102304, 2024.

[60] N. M. Foteinopoulou and I. Patras, "Emoclip: A vision-language method for zero-shot video facial expression recognition," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2024, pp. 1–10.

[61] X. Luwei, R. Mao, X. Zhang, L. He, and E. Cambria, "Vanessa: Visual connotation and aesthetic attributes understanding network for multimodal aspect-based sentiment analysis," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, 2024, pp. 4851–4863.

[62] J. Gu, Z. Han, S. Chen, A. Beirami, B. He, G. Zhang, R. Liao, Y. Qin, V. Tresp, and P. Torr, "A systematic survey of prompt engineering on vision-language foundation models," *arXiv preprint arXiv:2307.12980*, 2023.

[63] A. Valdivia, V. Luzón, E. Cambria, and F. Herrera, "Consensus vote models for detecting and filtering neutrality in sentiment analysis," *Information Fusion*, vol. 44, pp. 126–135, 2018.

[64] R. Mao, M. Ge, S. Han, W. Li, K. He, L. Zhu, and E. Cambria, "A survey on pragmatic processing techniques," *Information Fusion*, vol. 114, p. 102712, 2025.

[65] R. Mao, K. He, C. B. Ong, Q. Liu, and E. Cambria, "MetaPro 2.0: Computational metaphor processing on the effectiveness of anomalous language modeling," in *Findings of the Association for Computational Linguistics: ACL*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 9891–9908.