Multimodal Engagement Prediction in the Wild

Yichen Kang, Yanchun Zhang, Jun Wu
The Hong Kong University of Science and Technology
{yckang, yzhangut, jwueu}@connect.ust.hk

Abstract

Predicting engagement levels of a video remains a complex yet significant challenge. This paper addresses the EmotiW2024 Engagement Prediction Challenge, aiming to classify engagement levels in group videos featuring diverse scenarios into one of four predefined categories. To tackle this, we developed a novel multimodal fusion framework that integrates pose tracking, facial landmarks, facial features, and video understanding, ensembled with an early-fusion Transformer Fusion model. Our approach achieved a test accuracy of 74.74%, representing an 8.24% improvement over the baseline.

1. Introduction

The rise of online education and software-as-a-service (SaaS) platforms has revolutionized the way people access knowledge and interact digitally. However, accurately assessing user engagement remains a significant hurdle. Engagement, encompassing behavioral, cognitive, and emotional dimensions, is essential for improving user experience across various domains, including user experience research, human-robot interaction, gaming, digital marketing, and healthcare.

Recent advancements in deep learning have enabled more precise analysis of user engagement [10]. Deep learning techniques are being widely adopted across various fields to understand and predict user behavior. These methods allow for more accurate and nuanced insights into user engagement, enabling more personalized and effective interactions.

However, one significant trend in this area is analyzing engagement "in the wild" [17], where datasets reflect real-world complexities, remains under-explored. These real-world scenarios embody a complexity that captures the intricacies of everyday human behavior, which is far more practical than controlled environment datasets, where conditions are predictable.

This paper outlines our participation in the EmotiW2024 Engagement Prediction Challenge, presenting a robust framework that advances the state-of-the-art in engagement prediction. We aim to classify group videos featuring diverse scenarios of subjects' engagement while watching MOOCs, into one of four engagement levels.

Traditional 3D-CNN [9] and Video-ViT [3] methods have proven insufficient for video classification task, especially when dealing with realistic "in the wild" factors. In the recent years, a multi-pipeline approach has emerged [8, 13], specifically designed to leverage multiple modalities to analyze different features all at once.

Our approach includes preprocessing multiple independent modalities, including: (i) **Pose:** Pose keypoints extraction (ii) **Landmarks:** Facial landmarkings extraction (iii) **Facial Features:** Facial feature representations extraction (iv) **Video Understanding:** Video visual understanding and reasoning. The outputs from each modalities are combined and ensemble to predict the final engagement level.

The **baseline** for comparison is the highest accuracy of EngageNet Fusion model by Manisha *et al*. [24]. The highest reported validation accuracy is 68.49%.

Our key contributions include: (i) A novel Transformer-Fusion model optimized for multimodal integration, combining pose, facial landmarks, facial features, and video understanding. (ii) Develop a Video-LLaVa [15] based Video LLM zero-shot video understanding modality.

2. Related Works

Recent years have witnessed a shift in engagement prediction research from unimodal to multimodal analysis, and from controlled laboratory environments to real-world scenarios. This section discusses relevant research progress from multiple perspectives.

2.1. Engagement Assessment in Online Learning

The study of engagement prediction has evolved significantly with advances in computer vision and deep learning. Early work by [21] introduced the concept of "with-me-ness" as a gaze-based measure for student attention in MOOCs, establishing foundational metrics for engagement assessment. [6] further explored the dynamics of affective

states during complex learning, highlighting the importance of emotional engagement in learning processes.

2.2. Facial Analysis Techniques

Recent developments in facial analysis have contributed significantly to engagement prediction. [14] proposed reliable crowdsourcing methods for expression recognition in unconstrained environments, while [30] developed advanced face alignment techniques across large poses, both crucial for accurate engagement assessment in real-world scenarios. [12] provided a comprehensive review of face detection techniques, establishing the technological foundation for engagement analysis through facial features.OpenFace 2.0 [Baltrusaitis et al., 2018] [4]introduced a comprehensive facial behavior analysis toolkit that integrates facial landmark detection, head pose estimation, and action unit recognition using Convolutional Experts Constrained Local Model (CE-CLM), achieving state-of-the-art accuracy while maintaining real-time performance without specialized hardware.

2.3. Video Understanding and Transformer Architectures

In the context of video understanding, significant progress has been made through transformer-based architectures. [2] introduced ViViT, a video vision transformer that effectively captures temporal dependencies in video data. This was followed by [25] who proposed VideoMAE, demonstrating that masked autoencoders are efficient learners for self-supervised video pre-training. The recent Video-LLaVA framework by [16] further advanced the field by learning unified visual representations through alignment before projection.

2.4. Multimodal Learning Analytics

The integration of multimodal data for learning analytics has been explored by [20], who investigated how multimodal data can provide insights into learning processes. [27] examined the relationship between interaction and learning engagement in online learning, emphasizing the role of self-efficacy and academic emotions as mediating factors.

2.5. Recent Advances in Video Recognition

Most recently, [29] proposed a second-order transformer network specifically designed for video recognition, demonstrating improved performance in capturing complex temporal relationships in video data. These advances in multimodal analysis and transformer architectures have paved the way for more sophisticated approaches to engagement prediction in naturalistic settings.

3. Datasets

3.1. EngageNet Dataset

The primary dataset used in this work is EngageNet [24], from the EmotiW2024 Challenge. It consists of 10-second user videos labeled as Highly Engaged, Engaged, Barely Engaged, and Not Engaged, from 127 participants engagement level while watching MOOCs. There is a total of 7983 train, 1071 validation and 2257 test videos. Figure 1 shows sample video clips from each engagement levels.

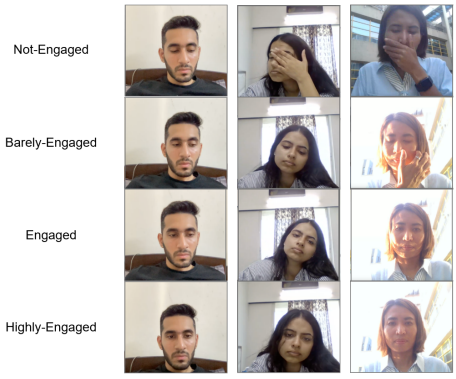

Figure 1. **Sample Dataset Clips** Video clips from each engagement levels

3.2. Data Augmentation

In the training process, insufficient data may lead to overfitting or suboptimal video classification results. In our dataset, there is a notable imbalance across four distinct categories, with the class containing the most data (highly engaged) significantly outperforming the other three in terms of accuracy. To address this imbalance and ensure a fairer distribution across categories, we employ data augmentation techniques to equalize the number of samples per class.

Data augmentation involves applying a series of transformations to the original images to generate new data samples. The primary goal of this approach is to expand the dataset, which is particularly effective in situations where data is scarce. Previous research [11, 22] has demonstrated the effectiveness of data augmentation in improving model performance. By introducing various transformations to the original dataset, models are able to learn more diverse features, thereby enhancing their generalization ability and improving performance on unseen data.

In our research, to better recognize and extract facial landmarks from video data, which are crucial for assessing participant engagement levels, we apply a combination of flipping and color jitter as data augmentation techniques [19]. Flipping involves mirroring the original image horizontally to create new data samples. Horizontal flipping

is particularly effective in our dataset, as it preserves the semantic meaning of facial features when flipped along the horizontal axis [7]. Color jitter modifies the color properties of video frame by changing parameters such as brightness, contrast, saturation, and hue randomly. By applying these transformations, it ensures that the model focuses on unchanging facial features rather than relying on specific lighting conditions or color information [28], enabling the model to learn color-invariant features and improving its generalization ability.

4. Approach

Figure 2 demonstrates that our architecture employs a fully connected, holistic fusion approach. After the video is input, it first goes through a four-layer data processing stage. The first layer is Pose Tracking, which performs head pose recognition via OpenFace 2.0 [4]. The second layer is Facial Landmarks, which tracks facial keypoints in the video via MediaPipe [18]. The third layer is Face Features Extraction, using Masked Autoencoder for facial video Representation LearnINg (MARLIN) [5] to capture expression changes. The fourth layer is Video Understanding, which extract visual features in text-form using the open-source visual language model Video-LLaVa [15]. After obtaining these feature matrices, we perform Data Merge on these data for processing and analysis. The final engagement class prediction is output by the Transformer Fusion model.

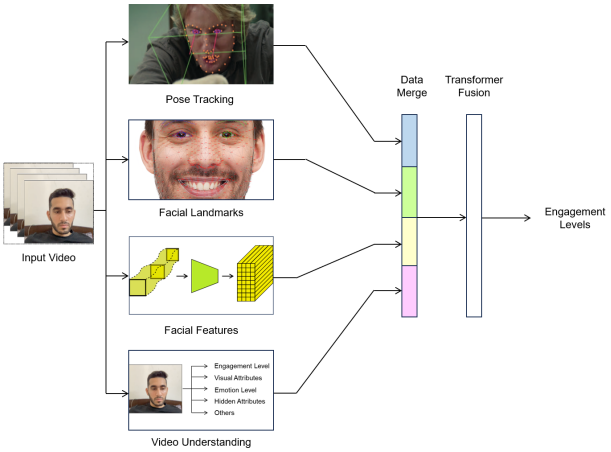

Figure 2. **Proposed Fusion Model** Our model implements a Transformer-Fusion to concatenate features from pretrained independant modalities.

4.1. Preprocessing

Each modality underwent feature extraction to ensure high-quality input for fusion: (i) Pose Tracking: Extracted gaze direction and head orientation using OpenFace 2.0. (ii) Facial Landmarks: Tracked 478 3D keypoints via MediaPipe for improved spatial accuracy. (iii) Facial Features: Used MARLIN for spatio-temporal representation of facial expressions. (iv) Video Understanding: Employed Video-LLaVA to extract rich visual-text embeddings.

Figure 3 demonstrates the outputs of three key modalities used in our approach: pose tracking, facial landmark detection, and facial feature extraction.

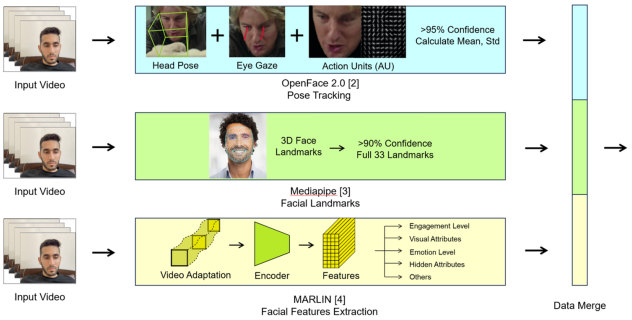

Figure 3. **Pose, Landmark, Facial**

Pose Tracking The OpenFace 2.0 toolkit implements head pose estimation, facial action unit (AU) recognition, and eye gaze estimation. OpenFace utilizes the Convolutional Expert Constrained Local Model (CE-CLM) to pinpoint facial landmarks and solves the perspective problem by mapping these landmarks onto the image via the orthographic projection method to accurately estimate the head pose. The Constrained Local Neural Field (CLNF) detector recognizes the pupil, iris, and eyelid for better estimation of eye gaze direction. In practice, we used three values: Eye Gaze, Head Pose and AU, as feature values to determine whether the participant in the video is focused or not.

Facial Landmarks We use Landmark function in MediaPipe to supplement the feature values, which is mainly used for detecting and tracking key points in an image or video. Although OpenFace provide face landmarks detection, it only offers 68 keypoints on the face in a 2D plane. MediaPipe's facial detection model is able to output 478 3D keypoints and provide blendshape scores for real-time rendering, which is more lightweight and easier to develop.

Facial Features The MARLIN functions as a framework to capture dynamic changes in facial movements and expressions. MARLIN extracts generic and transferable features by learning spatio-temporal variations in the face, using both unlabeled and labeled videos from the Internet for training. By applying weighted masking to different facial regions, the model focuses on learning subtle facial details. This approach effectively captures the variations and patterns of different facial regions, enhancing the expressiveness of the features.

Video Understanding Figure 4 illustrates the output of the video understanding modality, which extracts rich

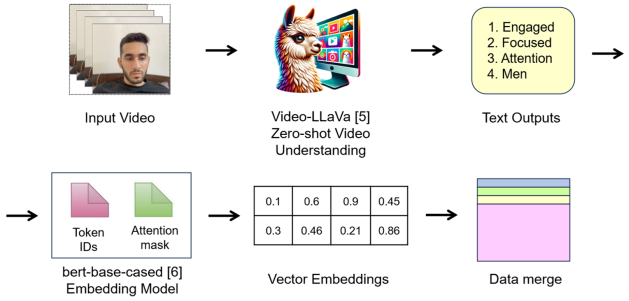

Figure 4. **Video Understanding**

visual-text embeddings from video clips. Video-LLaVA ensures the accuracy of facial feature extraction by integrating image and video features into a unified feature space. It performs well on various video quizzing and image understanding. Video-LLaVA has been open-sourced to provide a complete code implementation and pre-training weights, and we locally load the pre-trained Video-LLaVA model to extract additional facial visual features, which facilitates further processing and analysis of the new feature vector matrix.

4.2. Ensembling

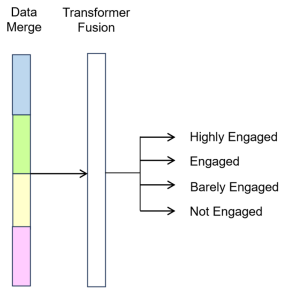

Figure 5. **Ensembling**

A Transformer Fusion model was employed to integrate features from all modalities. The model utilized two encoder layers with eight attention heads, ensuring robust feature interaction. Dropout and fully connected layers were added to prevent overfitting and enhance feature representation.

5. Results

5.1. Qualitative Results

Figures 6 and 7 illustrate the effectiveness of pose tracking and video understanding. For instance, in Figure 6, a 24-degree turn of head orientation classifies the person as "Not-Engaged," highlighting the effectiveness of pose and gaze analysis.

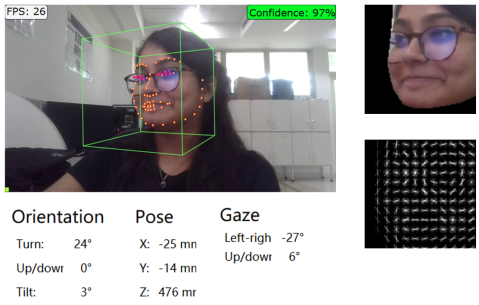

Figure 6. **Pose Modality**,"Not-Engaged" video, pre-processed via Pose Modality.

In Figure 7, in order to train the video understanding model, we tested with various customized prompts, ultimately adopting the one shown on the left. The model generates detailed descriptions of the person's pose, movements, and behaviors in the video. For example, the output here captures focused and attentive behavior, such as eye contact and hand gestures, aligning well with the "Highly-Engaged" label, showcasing the prompt's effectiveness.

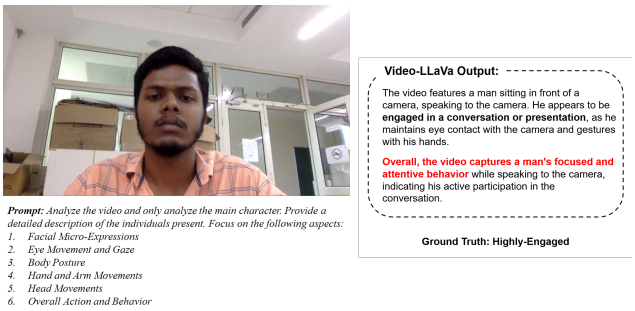

Figure 7. **Video Understanding Modality**,"Highly-Engaged" video, pre-processed via Video Understanding Modality.

5.2. Quantitative Metrics

To evaluate our models, we considered the accuracy and F1-score of each of the individual modalities and ensemble model on the test dataset of EngageNet.

5.3. Independent Classifier Performance

To better understand the contribution of each modality, we evaluated the performance of independent classifiers for Pose Tracking, Facial Landmarks, Facial Features, and Video Understanding. Table 1 highlights the performance metrics for each modality.

- **Pose Tracking:** Achieved the highest accuracy (69.8%) and F1 score (0.69), indicating that head orientation and gaze direction are strong indicators of engagement.

Model	Accuracy	F1 Score
Pose	**0.698**	**0.69**
Landmark	0.614	0.58
Face	0.689	0.67
Video	0.652	0.61

Table 1. **Modality Comparison** Summary of the best model performance on test dataset of each modality.

- **Facial Features:** Scored second-highest with an accuracy of 68.9% and an F1 score of 0.67, showcasing the value of spatio-temporal representation of facial expressions.
- **Video Understanding:** Reached an accuracy of 65.2% and an F1 score of 0.61, reflecting its ability to capture contextual visual features.
- **Facial Landmarks:** Performed the lowest with an accuracy of 61.4% and an F1 score of 0.58, suggesting redundancy in features when combined with other modalities.

5.4. Ensembling Performance

The ensembled models demonstrated a significant improvement in performance compared to individual classifiers. Table 2 compares the performance of different ensembling strategies:

Models	Accuracy
Late-Fusion (Hard Voting)	0.676
Late-Fusion (Soft Voting)	0.718
Late-Fusion (Weighted)	0.694
Early-Fusion (Transformer Fusion)	**0.744**

Table 2. **Ensemble Comparison** Performance on the test dataset for different ensembling strategies. Each ensemble strategy used the model outputs from all four modalities.

- **Late-Fusion (Hard Voting):** Achieved an accuracy of 67.6%, showing moderate gains by relying on majority voting among modality predictions.
- **Late-Fusion (Soft Voting):** Improved performance to 71.8%, leveraging probability-weighted predictions.
- **Late-Fusion (Weighted):** Achieved 69.4% accuracy by assigning different weights to each modality predictions.
- **Early-Fusion (Transformer Fusion):** Outperformed all late-fusion strategies with an accuracy of 74.4%, highlighting the benefits of integrating features at an earlier stage for robust interaction and representation.

5.5. Ablation Results

In order to identify the most important modality in ensembling, we perform an ablation study, the results shown in Table 3, by removing individual modalities from the early-fusion framework, the results suggests that the pose modality was necessary to get a higher accuracy. On the other hand, we also noticed that, when landmark modality is removed, it outperformed the accuracy of 4 modalities ensembling, by 0.3%.

Models	Accuracy
Pose-Land-Face	0.743
Pose-Land-Vid	0.740
Pose-Face-Vid	**0.747**
Land-Face-Vid	0.695

Table 3. **Ablation Comparison** Performance on the test dataset after removing different modalities from the early-fusion ensemble.

5.6. Compare with the state-of-the-art

As shown in Table 4, our proposed model outperformed baseline and state-of-the-art methods, achieving 74.74% accuracy compared to Ordinal ST-GCN's 71.2%.

This comparison highlights the robustness of the proposed Transformer Fusion model in effectively integrating multimodal features for engagement prediction, particularly in real-world scenarios.

Models	Accuracy
Baseline	0.665
VisioPhysioENet [23]	0.631
TCCT-Net [26]	0.689
Ordinal ST-GCN [1]	0.712
This work	**0.747**

Table 4. **Comparative analysis** of SOTA methods, on EngageNet test dataset.

The results indicate that: (1) Pose Tracking and Video Understanding are the most critical modalities for engagement prediction, offering complementary information about user behavior and contextual understanding. (2) Facial Landmarks provide limited additional value, as their features overlap with those extracted by other modalities. (3) Early-fusion strategies outperform late-fusion methods by enabling deeper interaction between features from different modalities.

6. Discussion

The results highlight the importance of early modality fusion for engagement prediction in real-world scenarios. Pose tracking and video understanding were critical contributors, while facial landmarks provided marginal gains. Notably, removing certain modalities during ablation tests suggested redundancy in features. Future work could focus on incorporating additional modalities, such as audio cues, and exploring advanced ensembling techniques to further improve performance.

7. Conclusion

This study proposed a novel multimodal engagement prediction framework using a Transformer Fusion model. By integrating pose tracking, facial landmarks, and video understanding features, we achieved a test accuracy of 74.74%, improving the baseline by 8.24%. Future work could explore advanced ensembling strategies and additional modalities for further enhancement.

8. Code

The code for this project is open-sourced on GitHub: https://github.com/yc-kang/EmotiW2024, including detailed instructions for replication.

9. Contributions

Yichen Kang

- Assembled the baseline model and implemented the baseline ensemble.
- Implemented MediaPipe-based landmark detection and enhanced 3D keypoint tracking.
- Contributed to debugging and refining the data pipeline for pose tracking features.

Yanchun Zhang

- Performed research and literature review on multimodal engagement prediction.
- Led the integration of the Video-LLaVA framework for video understanding.
- Analyzed results and proposed improvements for transformer-based fusion.

Jun Wu

- Designed and implemented data augmentation techniques to address dataset imbalance.
- Enhanced facial features extraction using MARLIN.
- Contributed to training the Transformer Fusion model and model optimization.

10. Acknowledgments

We would like to express our deepest gratitude to **Prof. Qifeng Chen** for his invaluable guidance and support throughout the course **EESM5900V**. His insights and expertise in the field of computer vision and deep learning were instrumental in shaping our project.

We also appreciate the constructive feedback and encouragement provided during the development process, which inspired us to strive for excellence in tackling the challenges of multimodal engagement prediction.

Finally, we would like to thank the extend our heartfelt thanks to **Prof. Abhinav, Monisha Singh and peers** for the dataset and baseline, their collaboration and helpfulness, greatly facilitated our research to a success.

References

[1] Ali Abedi and Shehroz S Khan. Engagement measurement based on facial landmarks and spatial-temporal graph convolutional networks. *arXiv preprint arXiv:2403.17175*, 2024. 5

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 2

[3] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846. IEEE, 2021. 1

[4] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, Xi'an, China, 2018. IEEE. 2, 3

[5] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, R. Haffari, and M. Hayat. Marlin: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1493–1504. IEEE, 2023. 3

[6] Sidney D'Mello and Arthur Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012. 1

[7] H. Fan, Z. Cao, Y. Jiang, Q. Yin, and C. Doudou. Learning deep face representation. *arXiv preprint arXiv:1403.2802*, 2014. 3

[8] B.T. Jin, L. Abdelrahman, C.K. Chen, and A. Khanzada. Fusical: Multimodal fusion for video sentiment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 798–806. ACM, October 2020. 1

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732. IEEE, 2014. 1

[10] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, December 2018. 1

[11] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012. 2

[12] Ajay Kumar, Amandeep Kaur, and Munish Kumar. Face detection techniques: a review. *Artificial Intelligence Review*, 52(2):927–948, 2019. 2

[13] D. Kumar, S. Madan, P. Singh, A. Dhall, and B. Raman. Towards engagement prediction: A cross-modality dual-pipeline approach using visual and audio features. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11383–11389. ACM, October 2024. 1

[14] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2852–2861, 2017. 2

[15] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 3

[16] Bolin Lin, Yang Ye, Biao Zhu, Jun Cui, Mengfei Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122v3*, 2024. 2

[17] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738. IEEE, 2015. 1

[18] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.L. Chang, M.G. Yong, J. Lee, and W.T. Chang. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 3

[19] J.J. Lv, X.H. Shao, J.S. Huang, X.D. Zhou, and X. Zhou. Data augmentation for face recognition. *Neurocomputing*, 230:184–196, 2017. 2

[20] Kshitij Sharma and Michail Giannakos. Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, 51(5):1450–1484, 2020. 2

[21] Kshitij Sharma, Patrick Jermann, and Pierre Dillenbourg. "with-me-ness": A gaze-measure for students' attention in moocs. In *Proceedings of International Conference of the Learning Sciences (ICLS)*, pages 1017–1022, 2014. 1

[22] C. Shorten and T.M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019. 2

[23] Alakhsimar Singh, Nischay Verma, Kanav Goyal, Amritpal Singh, Puneet Kumar, and Xiaobai Li. Visiophysioenet: Multimodal engagement detection using visual and physiological signals. *arXiv preprint arXiv:2409.16126*, 2024. 5

[24] M. Singh, X. Hoque, D. Zeng, Y. Wang, K. Ikeda, and A. Dhall. Do i have your attention: A large scale engagement prediction dataset and baselines. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 174–182. ACM, October 2023. 1, 2

[25] Yuting Tang, Bo Yang, Lei Li, and Min Lin. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35, 2022. 2

[26] Alexander Vedernikov, Puneet Kumar, Haoyu Chen, Tapio Seppänen, and Xiaobai Li. Tcct-net: Two-stream network architecture for fast and efficient engagement estimation via behavioral feature signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4723–4732, 2024. 5

[27] Yuxin Wang, Yi Cao, Shaoying Gong, Zhihong Wang, Na Li, and Lin Ai. Interaction and learning engagement in online learning: The mediating roles of online learning self-efficacy and academic emotions. *Learning and Individual Differences*, 94:102128, 2022. 2

[28] D. Yi, Z. Lei, S. Liao, and S.Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 3

[29] Bingbing Zhang, Wei Dong, Zhenwei Wang, Jianxin Zhang, and Qiule Sun. Second-order transformer network for video recognition. *Alexandria Engineering Journal*, 114:82–94, 2025. 2

[30] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016. 2