Multi-Feature-Fusion Framework for Robust Learner Engagement Recognition in Virtual Environments

Eman Almotairi[1] and Danda B. Rawat[2]
Department of Electrical Engineering and Computer Science
Howard University, Washington, DC, USA
[1]aeman.almotairi@bison.howard.edu, [2]danda.rawat@howard.edu

***Abstract*—Recognizing inattentive students is more complex on online platforms such as Zoom. The widespread adoption of virtual learning has increased the demand for reliable and interpretable methods to detect learner engagement. Although several models have been proposed, many methods still struggle to accurately capture engagement, often misclassifying inattentive students as engaged. To address this limitation, we propose a framework guided by robust visual representations, temporal modeling, and explainability to enhance engagement recognition in virtual learning environments and support academic success. Our approach combines facial image features extracted using DINOv2 with head pose parameters (yaw, pitch, roll), fusing these features to achieve more holistic engagement modeling.DINOv2 provides highly transferable embeddings and inherent attention maps that highlight visual regions, making it particularly effective for identifying facial cues of engagement. Video frames are scored based on these facial regions, with the top-K frames aggregated into a compact representation that is temporally modeled to capture dynamic cues before being classified into binary engagement labels. To ensure interpretability, explainable AI (XAI) is applied in post-hoc analysis, validating that model predictions align with key facial cues. Experiments conducted on the DAiSEE and EmotiW2023-EngageNet datasets demonstrate that our approach achieves 88% accuracy on EmotiW2023 and 75% on DAiSEE, consistently outperforming existing models while enhancing transparency and trust.**

***Index Terms*—Student engagement, deep learning, explainable AI (XAI), Distillation with NO labels version 2(DINOv2), attention mechanisms, LIME, LRP, online learning, AI**

I. Introduction

The shift to online learning platforms such as Zoom and Microsoft Teams, accelerated by the COVID-19 pandemic, has introduced both opportunities and challenges in the education sector [1]. While these platforms offer flexibility and accessibility, they also make it more difficult for instructors to assess student engagement, as traditional cues such as eye contact and body language are often less visible. Student engagement is a critical indicator of academic success and learning outcomes [2], encompassing both behavioural and emotional dimensions, including attention, interest, and motivation [3]. Consequently, recognizing inattentive students in virtual classrooms is significantly more challenging than in traditional settings. Unlike face-to-face environments, where teachers can directly observe students' behaviour, online learning restricts visibility and control. Teachers often struggle to determine whether students are paying attention, as learners may be distracted by external tools, websites, or their surroundings. Additionally, the large number of participants in virtual classes makes effective monitoring even more difficult. This challenge underscores the need for recognition models that can accurately identify inattentive students in virtual classrooms, enabling timely intervention and improving the overall quality of online education. Although numerous models have been developed for engagement recognition, many still misclassify inattentive students as engaged, limiting their reliability. Our work addresses this limitation by enhancing detection precision and improving overall performance through a combination of state-of-the-art AI techniques.

Machine Learning (ML) and Deep Learning (DL), as subfields of Artificial Intelligence (AI), have been widely applied to engagement detection tasks [4]. Most existing studies rely on convolutional neural networks (CNNs), which remain the dominant approach in deep learning. While CNN-based models effectively extract visual features from facial images, they often fall short in capturing long-range dependencies and multi-scale contextual relationships, which are crucial in complex learning environments. Moreover, these models typically overlook directional information such as head pose, which is essential to understand learner attentiveness and gaze orientation.

Despite the success of Vision Transformers in image classification, their application to real-time engagement detection within educational contexts remains underexplored. To address this gap, we propose a framework that combines the powerful feature extraction capabilities of **DINOv2** [5], a self-supervised Vision Transformer, with head pose estimation. DINOv2 provides highly transferable embeddings and inherent attention maps that highlight salient visual regions, making it effective for identifying facial cues of engagement. Meanwhile, the head pose estimation module captures directional cues—including yaw, pitch, and roll—that signal a learner's gaze and attention orientation. By integrating these complementary features, the proposed approach enables more holistic engagement modeling.

To improve robustness, we remove irrelevant or low-

quality samples and focus on selecting only the most informative frames, where facial cues of engagement are clearly visible. Each frame is processed using **DINOv2** to obtain embeddings and attention maps, which are then used to score frames based on their relevance. The top-K frames are aggregated into a compact video-level representation, and the fused features are classified using a lightweight multilayer perceptron (MLP). To ensure interpretability, **Explainable AI (XAI)** [6], [7] is employed in post-hoc analysis, confirming that the model's predictions align with semantically relevant facial cues. The proposed framework is evaluated on two benchmark datasets, **DAiSEE** [8] and **EmotiW2023-EngageNet** [9], demonstrating improved accuracy over prior approaches while providing transparent and trustworthy insights for real-world online learning settings.

The main contributions of this work are summarized as follows:

- We aim to enhance student engagement recognition in virtual learning environments.
- We leverage state-of-the-art AI methods, specifically **DINOv2**, a self-supervised Vision Transformer, to obtain transferable embeddings and attention maps, enabling effective identification of relevant cues for engagement recognition.
- We employ **explainable AI (XAI)** in post-hoc analysis to validate that the model's predictions align with key facial cues, thereby enhancing interpretability and trust.
- We evaluate the proposed framework on two benchmark datasets, **DAiSEE** and **EmotiW2023-EngageNet**, demonstrating improved generalization across diverse contexts.
- The results highlight the framework's improvement over prior work, clearly demonstrating the model's effectiveness in robust engagement detection.

The remainder of this paper is organized as follows: Section II reviews related work. Section III presents the proposed method. Section IV discusses experimental results. Section V concludes the paper and outlines future directions.

II. Related Work

Learner engagement detection has undergone significant evolution in recent years. Early studies primarily relied on convolutional neural networks (CNNs) to extract spatial features from facial images. While effective in capturing local patterns, these models struggled to represent long-range dependencies and multi-scale contextual cues. To overcome these limitations, research shifted toward hybrid designs incorporating temporal modeling, multimodal fusion, and attention mechanisms. More recently, transformer-based architectures—particularly Vision Transformers (ViTs)—have gained prominence due to their ability to capture global dependencies and generate interpretable attention maps. This progression from CNNs to transformer-driven approaches reflects a broader trend in computer vision toward more robust and explainable frameworks for engagement recognition. Initial research in learner engagement detection focused on CNN-based architectures. For instance, Gupta et al. [8] introduced a baseline CNN as a fundamental benchmark for spatial feature extraction. Building on this, Abedi et al. [10] combined ResNet with Temporal Convolutional Networks (TCN) to jointly capture spatial and temporal dependencies, while Tanwar et al. [11] explored deeper CNN configurations for improved representation learning. The work in [12] extended pre-trained CNNs with attention mechanisms, enabling models to focus on salient regions within the input. In another direction, the study in [13] proposed a multidimensional feature fusion strategy, and the work in [14] integrated geometrical facial descriptors with Long Short-Term Memory (LSTM) networks to incorporate temporal dynamics.

Recent studies have introduced hybrid designs combining spatial and motion cues. For example, Almotairi et al. [15] merged CNN-based spatial features with optical flow to capture fine-grained motion patterns, whereas the work in [16] employed an ensemble of ConvNeXt-Large and Gated Recurrent Units (GRU) for robust temporal modeling. The approach in [17] proposed a lightweight hybrid of KNN and CNN to reduce computational overhead, and [18] developed the BiusFPNICCSA framework, combining a Feature Pyramid Network (FPN) with coordinate self-attention for enhanced multi-scale representation. Transformer-based approaches have also emerged. For instance, [19] introduced MSC-Trans, which integrates CNNs, Transformers, and a temporal encoder–decoder structure for multimodal feature fusion on DAiSEE, while [20] combined a Vision Transformer (ViT) with LSTM for spatial-temporal modeling. Xiong et al. [21] proposed a CNN–CNN-Transformer model that jointly leverages coarse-grained body cues and fine-grained facial expressions for classroom engagement detection. Dresvyanskiy et al. [22] introduced a cross-multi-modal fusion approach enhanced with affective embeddings and cross-attention, enabling efficient integration of facial, body, and emotional cues for real-time engagement recognition. In summary, the field has evolved from early CNN baselines to increasingly sophisticated designs that integrate spatial, temporal, and multimodal information. This evolution highlights a clear trend toward attention-driven and transformer-based methods that enhance both accuracy and interpretability in engagement detection.

III. Methodology

The proposed framework enhances engagement recognition by combining robust visual feature extraction, attention- and pose-guided frame selection, lightweight temporal modeling, and multi-level explainability using XAI. Unlike prior works that uniformly aggregate all

frames, our method focuses on the most relevant frames and produces interpretable attributions that bridge raw pixel information with meaningful human-level cues. The overall pipeline, illustrated in Figure 1, consists of four main components: feature extraction, temporal modeling, model training, and explainability.

A. Feature Extraction with DINOv2

We adopt **DINOv2**, a self-supervised Vision Transformer [5], as the feature extractor. For each uniformly sampled frame f_t, DINOv2 outputs:

- The **embedding** $h_t \in \mathbb{R}^{768}$, representing global facial features.
- The **attention map** $A_t \in \mathbb{R}^{H \times W}$, highlighting salient regions such as the eyes and mouth.

An attention-based score quantifies ROI concentration:

$$s_t = \frac{\sum_{(i,j)\in \text{ROI}} A_t(i,j)}{\sum_{(i,j)\in \Omega} A_t(i,j)}, \tag{1}$$

where ROI corresponds to the eyes and mouth. To reduce false positives (e.g., frames where the student looks away), we refine the score with head pose:

$$s'_t = s_t \cdot \exp(-\alpha \cdot |yaw_t|). \tag{2}$$

The frames are ranked by s'_t, and the top-K are selected. For each, the fused embedding is:

$$\hat{h}_{t_k} = [\, h_{t_k} \parallel p_{t_k}], \quad p_{t_k} = [yaw, pitch, roll] \in \mathbb{R}^3. \tag{3}$$

Visual embeddings (h_t) capture fine-grained appearance and expression cues, while geometric head pose features (p_t) encode directional and attentional information. These two modalities are complementary rather than redundant, enabling the model to combine expressive facial details with spatial orientation for more holistic engagement recognition. Such fusion of visual and geometric modalities has also been shown in prior multimodal frameworks to improve robustness and generalization in behavioral analysis tasks [19], [21].

Feature Alignment: Prior to fusion, all features are normalized to zero mean and unit variance to ensure balanced contributions. The concatenated representation $\hat{h}_{t_k} \in \mathbb{R}^{771}$ integrates 768-dimensional visual features with 3-dimensional geometric cues within a unified space.

B. Temporal Modeling

Selected embeddings $\{\hat{h}_{t_1}, \ldots, \hat{h}_{t_K}\}$ are aggregated using attention-weighted pooling:

$$h_{\text{att}} = \sum_{k=1}^{K} \tilde{w}_k \hat{h}_{t_k}, \quad \tilde{w}_k = \frac{\exp(s'_{t_k}/\tau)}{\sum_{j=1}^{K} \exp(s'_{t_j}/\tau)}. \tag{4}$$

To capture micro-dynamics (e.g., frequent blinking), we compute weighted deltas:

$$\bar{\Delta}_{\text{att}} = \sum_{k=2}^{K} \frac{\tilde{w}_k + \tilde{w}_{k-1}}{2} (\hat{h}_{t_k} - \hat{h}_{t_{k-1}}). \tag{5}$$

The final video representation is:

$$h_{\text{temp}} = [\, h_{\text{att}} \parallel \bar{\Delta}_{\text{att}}]. \tag{6}$$

C. Model Training

The temporal representation h_{temp} is fed into a lightweight multilayer perceptron (MLP) classifier:

$$\hat{y} = \text{MLP}(h_{\text{temp}}), \tag{7}$$

trained with cross-entropy loss on DAiSEE [8] and EmotiW2023-EngageNet [9].

D. Explainability with XAI

Explainability is a core component of our framework, integrated at two levels: during frame scoring and in post-hoc validation.

- **Critical Frame Attribution:** Attention scores $\{s'_t\}$ highlight the most influential frames, ensuring that predictions are grounded in behaviorally relevant segments.
- **Spatial Attribution:** XAI maps derived from DINOv2 and post-hoc methods (e.g., LIME, LRP) validate that the model focuses on key facial regions such as the eyes and mouth.
- **Behavioral Attribution:** Temporal dynamics reveal interpretable behavioral cues—frequent blinks (linked to drowsiness) and head pose variations (linked to distraction).
- **Concept-Level Attribution:** Inspired by concept bottleneck models [23], engagement is decomposed into interpretable concepts such as smile intensity, brow furrow, eye openness, and head tilt. This forms a transparent reasoning chain: raw pixels → concepts → engagement prediction.

By combining intrinsic explainability (via attention mechanisms) with post-hoc interpretability methods, our framework ensures that engagement predictions are not only accurate but also transparent and trustworthy.

IV. Experiments and Results

In our experiment, we evaluated the performance of the framework on two benchmark datasets, EmotiW2023-EngageNet and DAiSEE. The performance was measured in terms of classification accuracy. In addition to quantitative evaluation, we employed XAI techniques to analyze the behavior of the model. Post-hoc visualizations highlight the facial regions most influential in predictions, providing insight into how the model attends to engagement-related cues such as eye gaze, blinking, and head orientation.

A. Model Performance Metrics

We used standard classification metrics, including precision, recall, accuracy, and F1-score, to evaluate the performance of the model in detecting engagement levels. These metrics provide a comprehensive assessment of the model's effectiveness, as outlined by Sokolova and Lapalme [24].

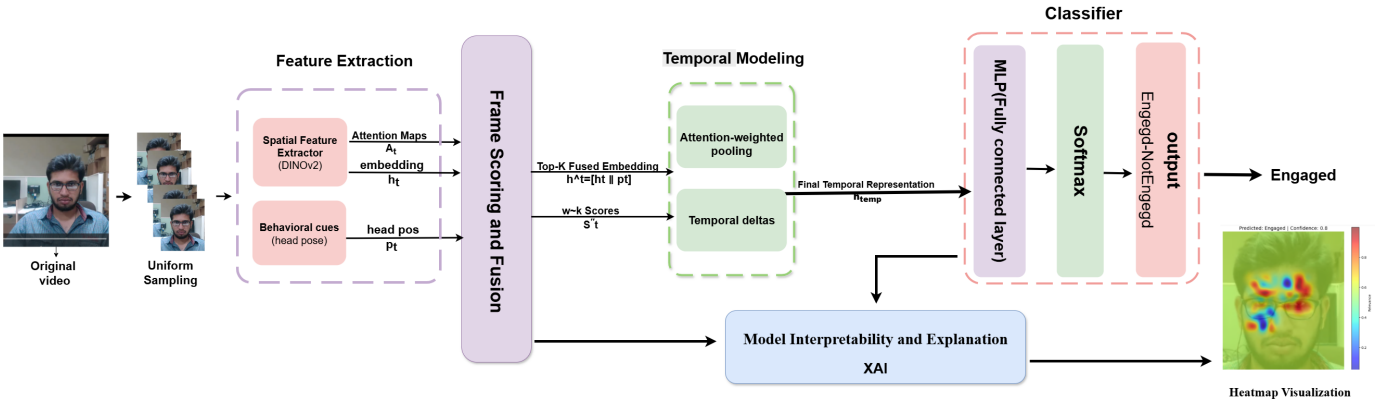

Fig. 1: The proposed framework.

B. Datasets and Preprocessing

We trained a model on two publicly available datasets to investigate engagement. **EmotiW2023-EngageNet** [9]: A dataset of 11,206 video clips with four engagement levels. To address the imbalance, the two higher levels were merged as 'Engaged' and the two lower levels as 'Not-Engaged.' **DAiSEE** [8]: A dataset of 9,086 video clips labeled with engagement, frustration, confusion, and boredom. For balance, the last three were merged into a single 'Not-Engaged' class, yielding a binary Engaged/Not-Engaged distribution. To prepare the data for training, we selected balanced subsets from each dataset to ensure equal representation of engagement classes. This strategy maintained fairness and consistency in evaluation. Subsets were randomly sampled to capture diverse engagement scenarios without compromising feasibility.

Each video frame was resized to 224×224 pixels and normalized to the range [0, 1], producing consistent inputs suitable for the training. In our experiments, we used 70 training videos and 20 validation videos with identical class distributions. From each video, 16 frames were uniformly sampled, and the DoVx framework further refined these by selecting the top-8 most informative frames using attention and head pose cues.

We adopted a fixed 20% validation split rather than k-fold cross-validation to simplify evaluation and avoid overlap across folds. To ensure independence and prevent data leakage, all splits were performed at the subject level so that frames or videos from the same student never appear in more than one set (training, validation, or testing).

C. Experimental Results on EmotiW2023

This section presents the experimental evaluation of the framework on the EmotiW2023-EngageNet dataset. We first illustrate the frame selection mechanism, followed by a quantitative comparison against baseline models, detailed performance metrics of our method, and a comparative analysis with prior approaches in the literature. Figure 2 illustrates an example of the frame selection process. The top-ranked frames are shown with their corresponding importance scores, while the histogram visualizes the overall distribution of frame importance. These results confirm that our approach prioritizes frames containing meaningful facial cues, ensuring that the classifier bases its decision on relevant evidence. Notably, frames where the learner maintains a frontal head pose with clear visibility of the eyes and mouth consistently receive higher scores, whereas frames with strong yaw deviations or partial occlusions are down-weighted. This behavior demonstrates that the scoring mechanism effectively distinguishes between informative and distracting frames, thereby reducing noise introduced by irrelevant samples.

TABLE I: Performance of baseline CNN, ViT, and DINOv2 models on EmotiW2023.

Model	Accuracy	Precision	Recall	F1-score
MobileNetV2	0.53	0.55	0.53	0.49
ResNet101	0.53	0.69	0.54	0.42
InceptionV3	0.63	0.68	0.62	0.63
EfficientNet	0.62	0.68	0.62	0.59
VGG19	0.62	0.63	0.62	0.61
ViT	0.66	0.65	0.65	0.66
DINOv2	**0.75**	**0.77**	**0.76**	**0.75**

a) Baseline Performance.: Table I reports the performance of CNN-based models, ViT, and DINOv2 on EmotiW2023. These results highlight the relative strengths and weaknesses of conventional architectures compared to self-supervised transformer representations.

b) Proposed Framework Performance.: Table II presents the detailed classification metrics of our framework on EmotiW2023. The model achieves strong precision, recall, and F1-scores, with an overall accuracy of 88%.

Fig. 2: Illustration of frame selection. Top-ranked frames are shown with their corresponding importance scores, and the histogram depicts the distribution of frame importance.

TABLE II: Performance of the proposed framework on EmotiW2023 dataset.

Metric	Precision	Recall	F1-score
Accuracy		**0.88**	
Macro Avg	0.88	0.90	0.87
Weighted Avg	0.91	0.88	0.88

c) Comparison with Prior Work: Table III compares our model with previous engagement recognition approaches. Earlier models based on handcrafted or hybrid features (e.g., OpenFace 2.0 + TCN, CNN + Optical Flow) report accuracies around 65–66%, while transformer-based ViT + LSTM improves to 74%. Our DINOv2-based framework with explainability-guided frame selection achieves 88%, outperforming all prior approaches.

TABLE III: Comparative evaluation of engagement recognition approaches on EmotiW2023.

Reference	Method	Accuracy
Singh et al. [9]	OpenFace 2.0 + TCN	65.6%
Almotairi et al. [15]	CNN + Optical Flow	66.0%
Alarefah et al. [20]	ViT + LSTM	74.0%
Proposed (Ours)	DINOv2 + Head Pose	**88%**

Overall, the results demonstrate that the proposed design not only improves predictive accuracy but also provides interpretable insights into the behavioural cues contributing to engagement predictions (Figure 3).

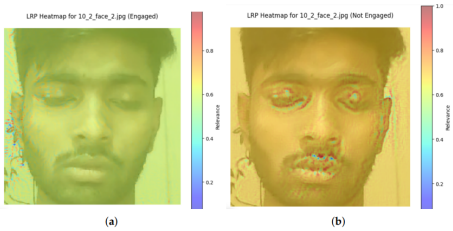

Fig. 3: LRP heatmaps: (**a**) simple DINOv2 baseline; (**b**) proposed model focusing on critical facial regions.

D. Experimental Results on DAiSEE

This section evaluates the proposed framework on the DAiSEE dataset. We first compare backbone architectures and report classification metrics, followed by an interpretability analysis, and finally benchmark our results against previous work.

The results show that transformer-based models (ViT and DINOv2) outperform conventional CNN architectures, with DINOv2 achieving the highest precision of 62% (Figure 4).

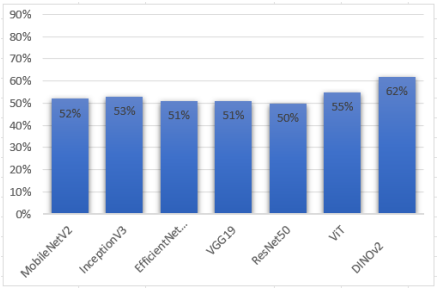

Fig. 4: Classification accuracy of different models on the DAiSEE dataset.

As shown in Table IV, the proposed framework achieves an overall accuracy of 75%, with balanced macro and weighted averages across precision, recall, and F1-score. This indicates consistent performance across both engaged and not-engaged classes.

TABLE IV: Classification report on DAiSEE dataset.

Metric	Precision	Recall	F1-score
Macro Avg	0.77	0.75	0.75
Weighted Avg	0.77	0.75	0.75
Accuracy		**0.75**	

To illustrate the performance of different backbone architectures, Figure 4 compares classification accuracies on DAiSEE. Conventional CNN models such as MobileNetV2, InceptionV3, and ResNet50 achieve moderate performance, while transformer-based models (ViT and DINOv2) provide notable improvements, with DINOv2 reaching the highest accuracy.

Finally, Table V compares our approach to previous work on DAiSEE.While early CNN-based baselines achieved around 57–61% accuracy, recent hybrid and transformer-based approaches report up to 68%. Our framework, by integrating attention-driven frame selection with explainability, achieves 75%, surpassing all prior methods.

To validate interpretability, we visualize the model's attention using LIME in Figure 5. Before refinement, attention maps were often scattered over irrelevant regions. After explanation-guided refinement, saliency maps

TABLE V: Comparison of classification accuracy across prior methods and the proposed framework on DAiSEE dataset.

Reference	Method	Accuracy
[8]	Baseline CNN	57.0%
[10]	ResNet + TCN Hybrid	59.0%
[11]	Deep CNN	60.8%
[12]	Pre-trained CNN + Attention	63.4%
[13]	Multi-dimensional Fusion	64.2%
[14]	Geometric + LSTM	60.1%
[15]	CNN + Optical Flow	61.0%
[16]	ConvNeXt-Large + GRU Ensemble	68.13%
[17]	Lightweight KNN + CNN	68.57%
[19]	MSC-Trans (CNN + Transformer)	61.86%
[18]	BiusFPN_ICCSA (Self-Attention)	68.16%
[25]	MS-ResNet-50 + Self-Attention	60.03%
Proposed (Ours)	**DINOv2 + Head Pose**	**75.0%**

consistently highlight the eyes and mouth, confirming that the model focuses on engagement-relevant cues.

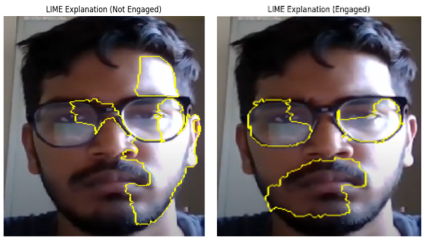

Fig. 5: Visualization of model attention using LIME.

Together, the results on **DAiSEE** and **EmotiW2023** confirm that the proposed framework achieves robust generalization across diverse benchmarks.

V. Conclusion

Recognizing learner engagement in virtual environments is challenging due to subtle behavioral cues and frequent misclassification by existing models. This paper introduced a framework combining **DINOv2** embeddings, attention maps, and head pose cues to capture both visual and directional indicators of engagement. By prioritizing informative frames and discarding irrelevant samples, the method improves robustness and interpretability. Experiments on **DAiSEE** and **EmotiW2023-EngageNet** show that our approach outperforms previous methods in accuracy and explainability. **Future Work:** Evaluate cross-dataset generalization by training on one dataset and testing on another. In addition, assess real-time feasibility through experiments on inference speed, memory usage, and model size.

Acknowledgments

This work was supported in part by the Center of Excellence in AI and Machine Learning (CoE-AIML) at Howard University under Contract W911NF-20-2-0277 with the U.S. Army Research Laboratory, as well as in part by Meta Research Gift Funds. The authors would also like to acknowledge the support of Shaqra University. However, any opinions, findings, conclusions, or recommendations expressed in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

References

[1] C. B. Hodges, S. Moore, B. B. Lockee, T. Trust, and M. A. Bond, "The difference between emergency remote teaching and online learning," 2020.

[2] A. Brown, J. Lawrence, M. Basson, and P. Redmond, "A conceptual framework to enhance student online learning and engagement in higher education," *Higher Education Research & Development*, vol. 41, no. 2, pp. 284–299, 2022.

[3] J. C. Richards, "Exploring emotions in language teaching," *Relc Journal*, vol. 53, no. 1, pp. 225–239, 2022.

[4] P. Bhardwaj, P. Gupta, H. Panwar, M. K. Siddiqui, R. Morales-Menendez, and A. Bhaik, "Application of deep learning on student engagement in e-learning environments," *Computers & Electrical Engineering*, vol. 93, p. 107277, 2021.

[5] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[8] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "Daisee: Towards user engagement recognition in the wild," *arXiv preprint arXiv:1609.01885*, 2016.

[9] 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*, 2023, pp. 174–182.

[10] A. Abedi, F. Shih, and Y. Chen, "Improving student engagement recognition through temporal modeling," *IEEE Transactions on Affective Computing*, 2021.

[11] S. Tanwar, P. Gupta, and A. Goel, "Engagement recognition using deep convolutional networks," in *Proceedings of the 2022 International Conference on Signal Processing and Communications (SPCOM)*, 2022, pp. 1–5.

[12] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[13] Y. Xie, K. Zhang, and Y. Zhao, "Student engagement recognition using multidimensional feature fusion," *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 123–140, 2023.

[14] E. Sherly, "Fostering learning with facial insights: Geometrical approach to real-time learner engagement detection," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*. IEEE, 2024, pp. 1–6.

[15] E. Almotairi, D. B. Rawat, S. Liu, and S. Rubaii, "Enhancing engagement detection using spatial and motion cues," *IEEE Access*, 2024.

[16] A. Shiri, L. Zhang, and W. Huang, "Recognition of student engagement using convnext and gru ensembles," *Neural Computing and Applications*, 2024.

[17] M. Malekshahi and M. Pourahmadi, "A general lightweight hybrid knn-cnn framework for engagement detection," in *Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[18] K. Naveen and S. Reddy, "Biusfpn_iccsa: Feature pyramid networks with coordinate attention for engagement detection," in *Proceedings of the 2025 IEEE International Conference on Computer Science and Artificial Intelligence (ICCSA)*. IEEE, 2025, pp. 1–7.

[19] N. Xie, Z. Li, H. Lu, W. Pang, J. Song, and B. Lu, "Msc-trans: A multi-feature-fusion network with encoding structure for student engagement detecting," *IEEE Transactions on Learning Technologies*, 2025.

[20] S. Alarefah, N. Ahmed, and F. Alotaibi, "Engagement recognition with vision transformers and lstm hybrids," in *Proceedings of the 2025 International Conference on Pattern Recognition (ICPR)*. IEEE, 2025, pp. 1–8.

[21] Y. Xiong, G. Xinya, and J. Xu, "Learning engagement recognition using a cnn–transformer model with coarse- and fine-grained cues," *Education and Information Technologies*, 2024.

[22] D. Dresvyanskiy, A. Karpov, and W. Minker, "A cross-multi-modal fusion approach for enhanced engagement recognition," in *International Conference on Speech and Computer*. Cham: Springer, 2024, pp. 3–17.

[23] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International conference on machine learning*. PMLR, 2020, pp. 5338–5348.

[24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[25] Q. Liu, J. Zhang, Q. Xu, Y. Wang, R. Chen, and H. Zhang, "A method for recognizing students' concentration based on self-attention mechanism," in *2025 7th International Conference on Computer Science and Technologies in Education (CSTE)*. IEEE, 2025, pp. 458–463.