Deep Learning for Vision and Multimodal Data

Fall 2024

Engagement Prediction: A Multimodal Fusion Approach

Yichen Kang, Yanchun Zhang, Jun Wu

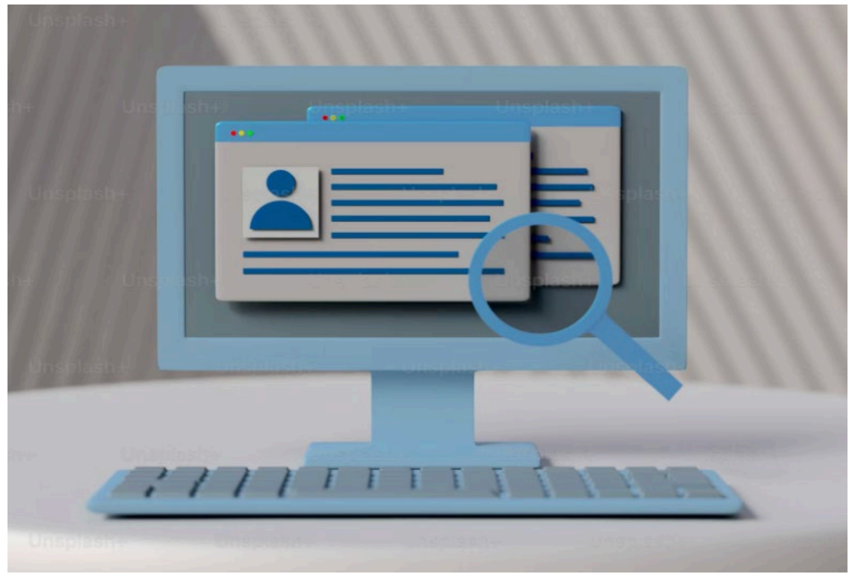
Code: https://github.com/yc-kang/EmotiW2024

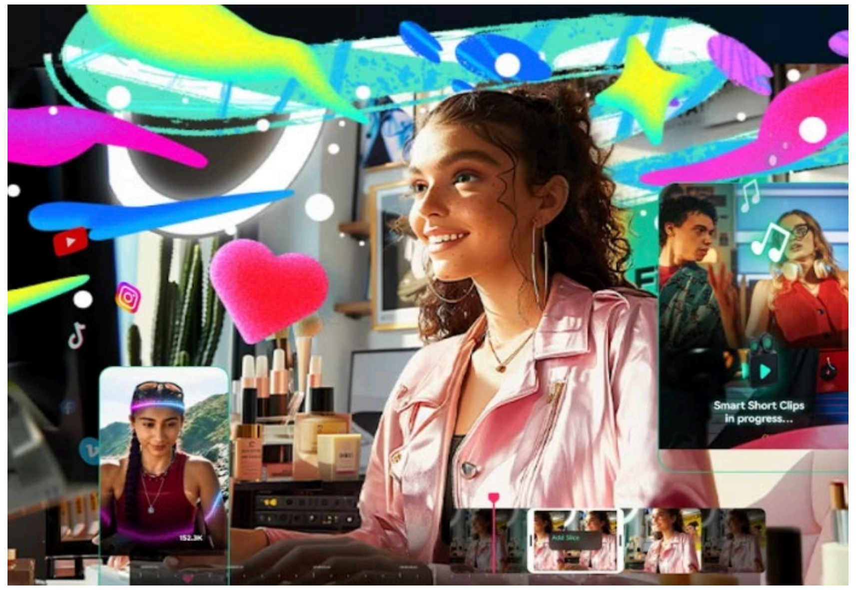
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

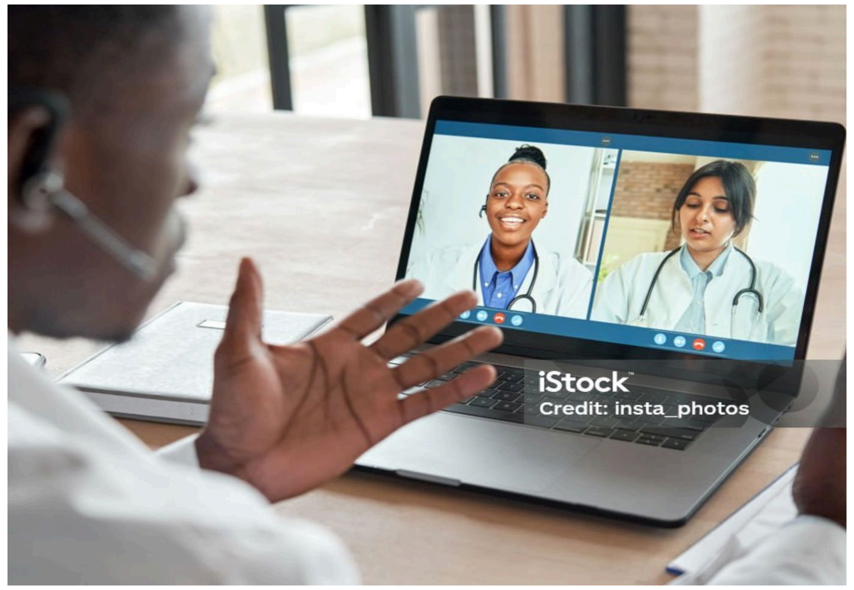
Engagement Prediction

What is the problem? & Why is it interesting?

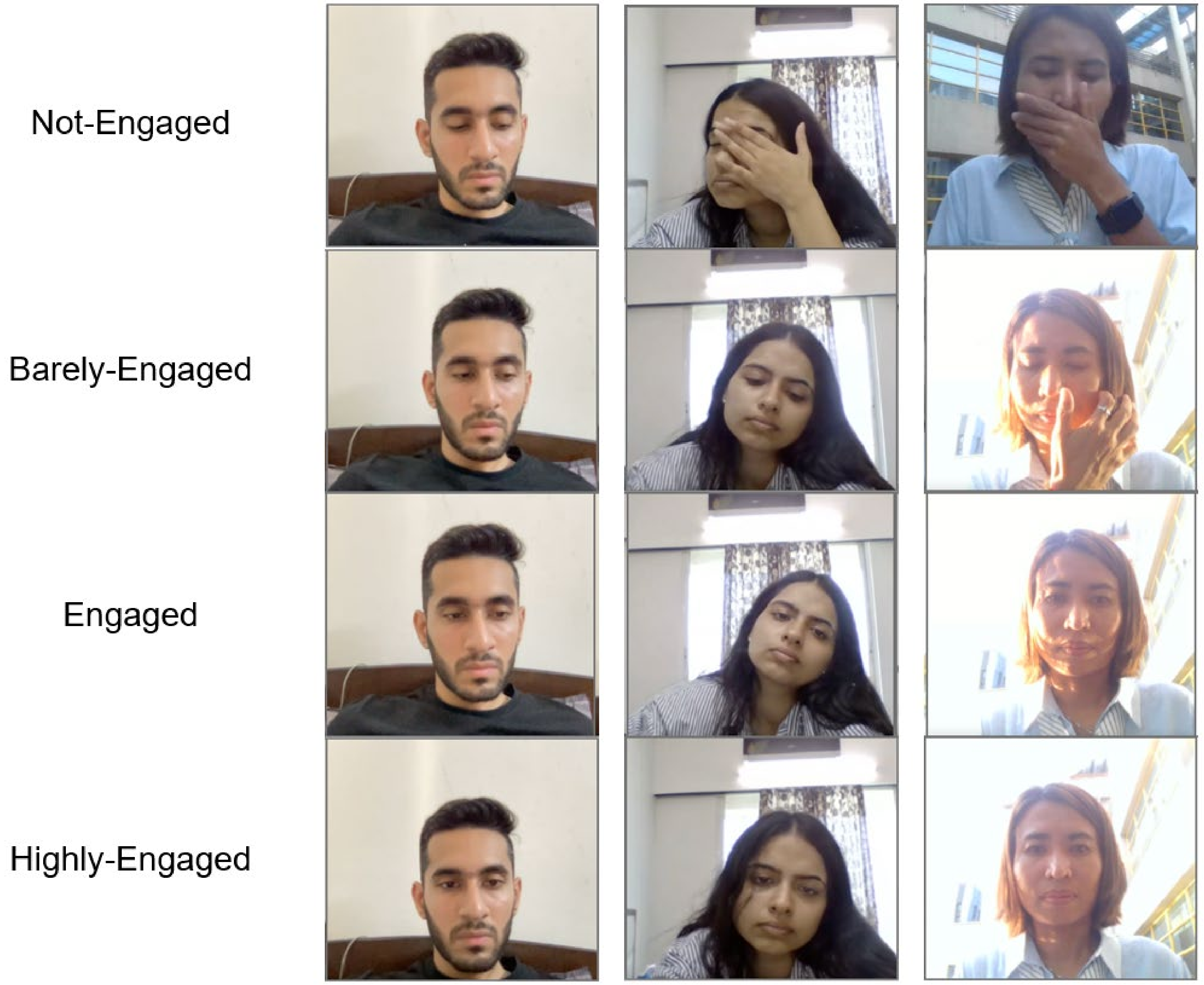
- Predicting user engagement levels in online videos (e.g., MOOCs).
- Engagement is a critical factor in UX, digital marketing, healthcare, etc.







EngageNet Dataset

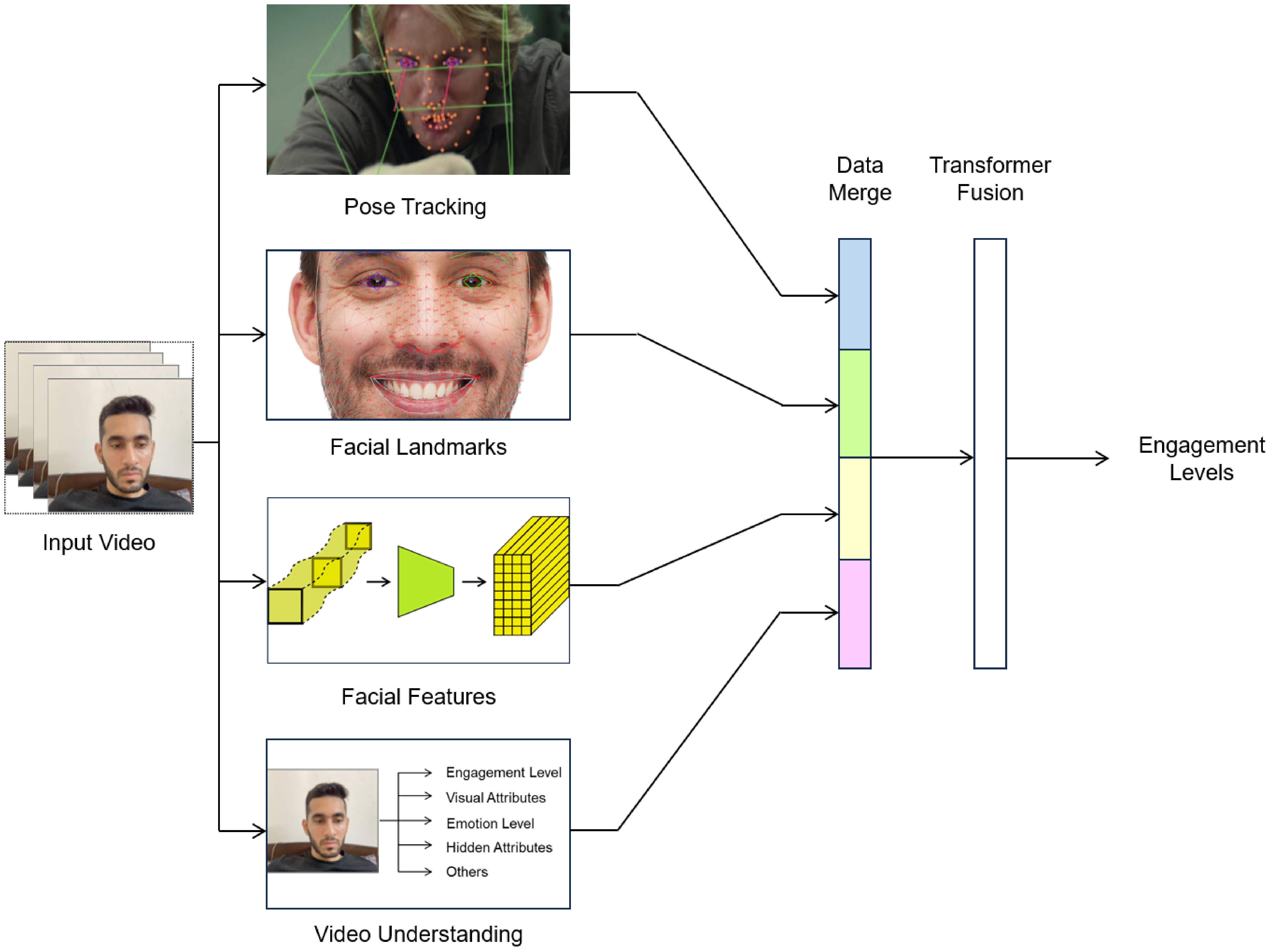


EngageNet [1] Statistics:

- 31 hours (11311 videos), 127 subjects
- Video duration: 10s
- 4 Classes: “Highly-Engaged”, “Engaged”, “Barely-Engaged” and “Not-Engaged”
- In the wild settings, while watching MOOCs

[1] Do I Have Your Attention: A Large Scale Engagement Prediction Dataset and Baselines, Monisha et al, ICMI 2023.

Architecture Overview



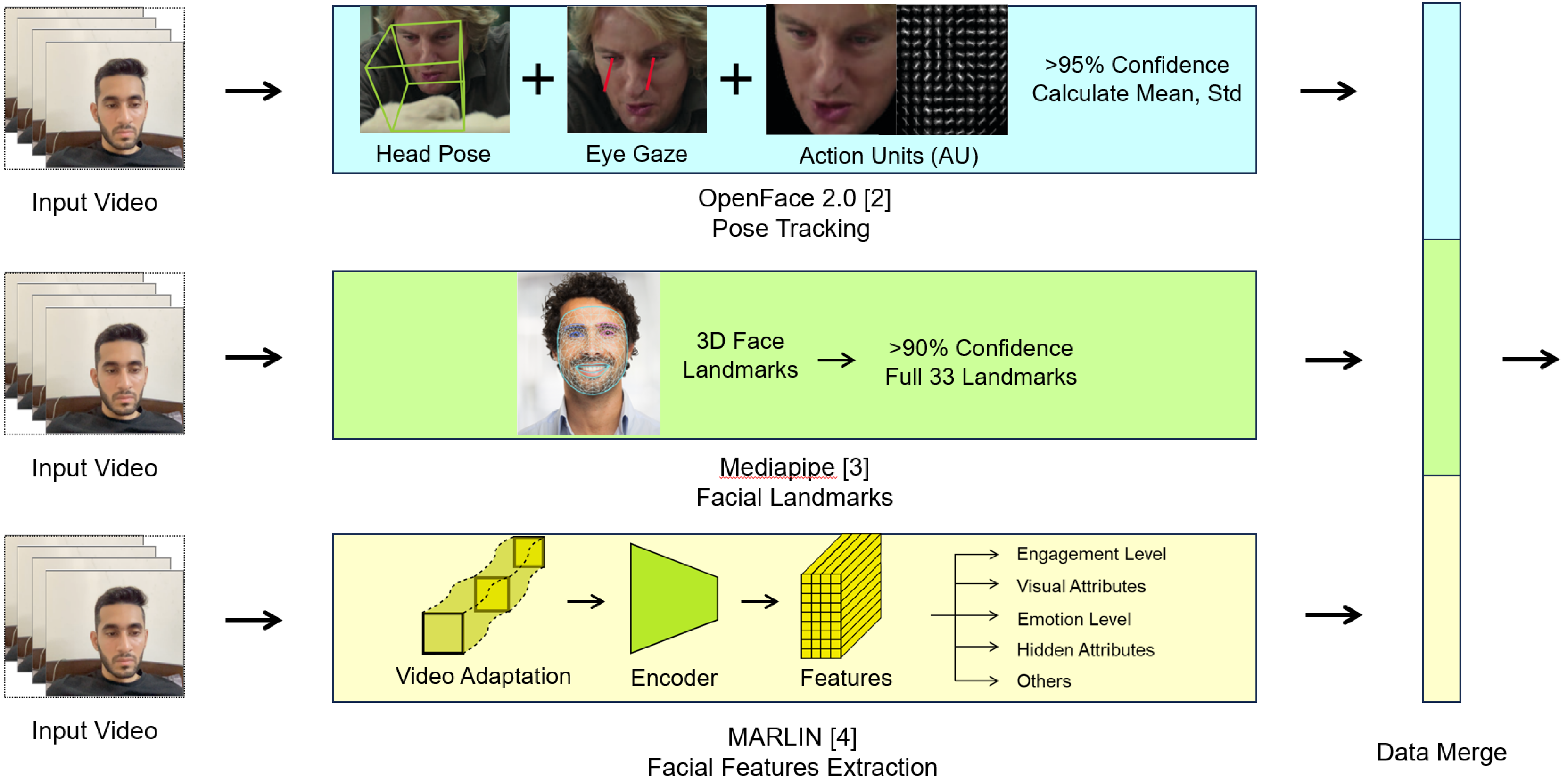
What method are we using?

- Multimodal Fusion Approach

Transformer Fusion Model:

- Integrates features from multiple modalities.
- Provides robust engagement level predictions.

Pose, Landmark, Facial

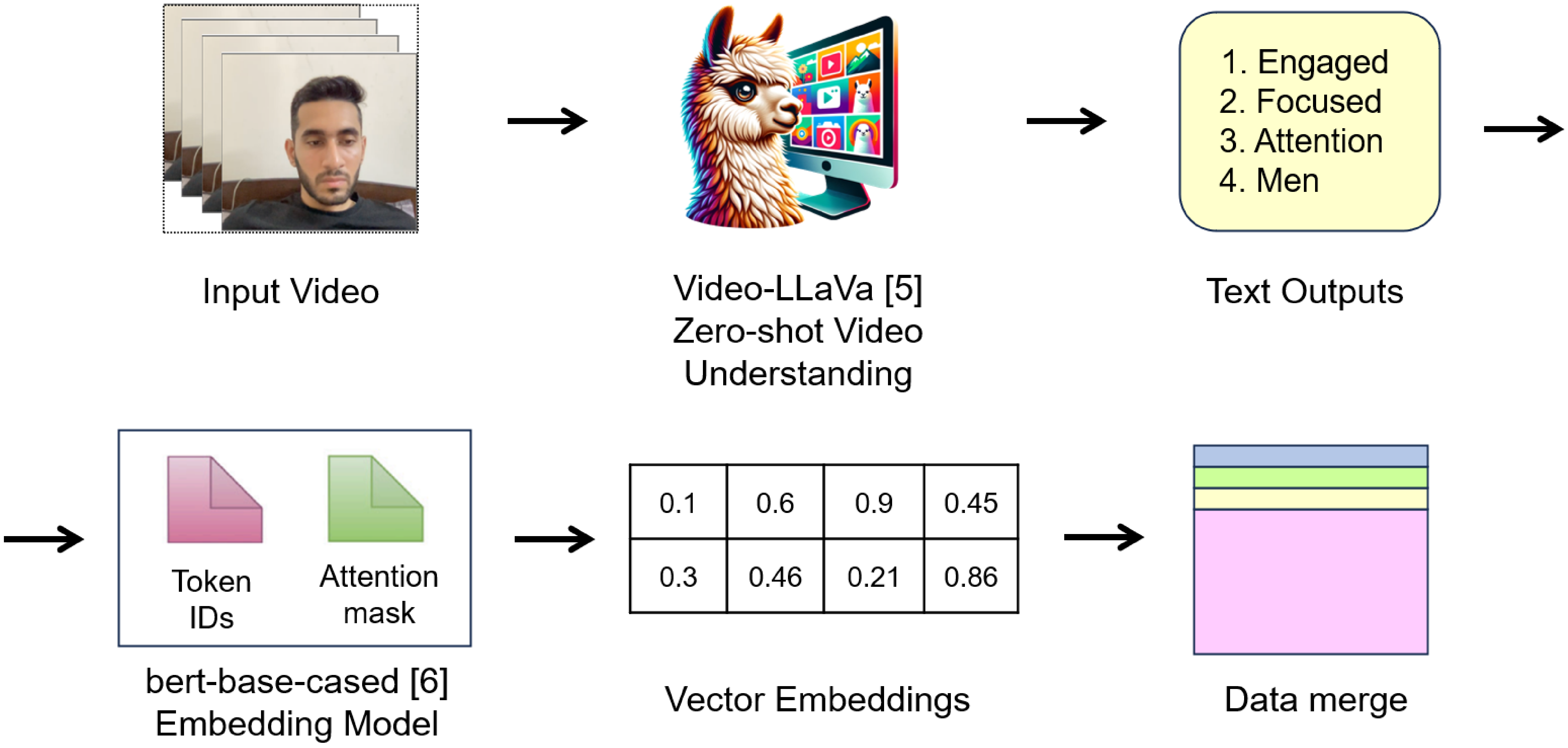


[2] OpenFace 2.0: Facial Behavior Analysis Toolkit, Tadas et al, IEEE FG 2018.

[3] Mediapipe: A framework for perceiving and processing reality, Lugaresi et al, CVPR 2019.

[4] Marlin: Masked autoencoder for facial video representation learning, Cai et al, CVPR 2023.

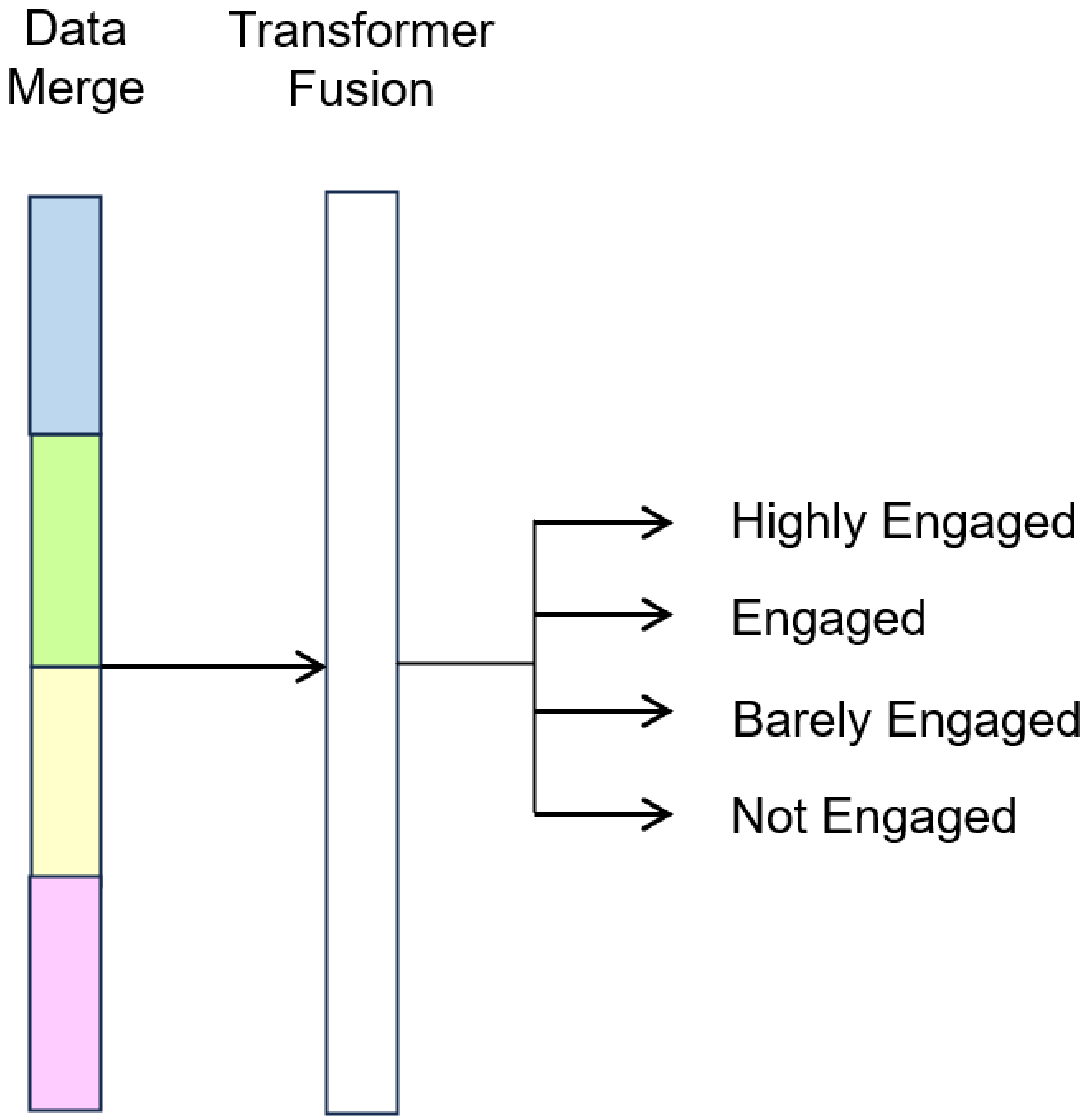
Video Understanding



[5] Video-llava: Learning united visual representation by alignment before projection, Lin et al, arXiv preprint 2023.

[6] https://huggingface.co/google-bert/bert-base-cased

Ensembling



Transformer Early-Fusion Ensemble

- Combines features, before making final decision
- Performed best F1-score, as it could learn relationships between modalities

Independent Classifier Performance

Models	Accuracy	F1-score
Pose	**0.698**	**0.69**
Landmarks	0.614	0.58
Facial	0.689	0.67
Video Understanding	0.652	0.61

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

Ensembling Performance

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.

Ablations Results

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.

Landmarks: Land
Facial: Face
Video Understanding: Vid

Compare with the state-of-the-art

Models	Accuracy
Baseline	0.665
VisioPhysioENet [7]	0.631
TCCT-Net [8]	0.689
Ordinal ST-GCN [9]	0.712
This work	**0.747**

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

[7] VisioPhysioENet: Multimodal Engagement Detection using Visual and Physiological Signals, Singh et al, arXiv preprint 2024.
[8] TCCT-Net: Two-Stream Network Architecture for Fast and Efficient Engagement Estimation via Behavioral Feature Signals, Vedernikov et al, CVPR 2024.
[9] Engagement Measurement Based on Facial Landmarks and Spatial-Temporal Graph Convolutional Networks, Abedi et al, arXiv preprint 2024.

Conclusion

- Propose a novel approach on addressing engagement classification, with a multimodal fusion approach.
- Propose a video-LLM classification pipeline, based on zero-shot video understanding.
- Achieving an overall test accuracy of 74.74%, 8.24% improvement from baseline.