From Emotion Recognition to Mind-Wandering Detection: A Comparative Analysis of Video-Based Emotion Foundation Models

Ekta Sood*
University of Colorado Boulder
ekta.sood@colorado.edu

Sebastian Ricke*
University of Colorado Boulder
sebastian.ricke@colorado.edu

Trisha Mittal
Dolby Laboratories
trisha.mittal@dolby.com

Sidney K. D'Mello
University of Colorado Boulder
sidney.dmello@colorado.edu

Abstract

Automated mind-wandering (MW) detection from educational video offers a potential path toward continuous and non-intrusive measurement of attentional state during learning. Recent work introduced a pragmatic starting point for video-based MW detection by transferring facial emotion recognition (ER) features to an in-lab reading dataset with MW labels, showing that an AffectNet-pretrained ResNet50 encoder can support above-chance prediction. In this work, we revisit this approach in light of recent ER foundation models by evaluating four frozen feature extractors—the AffectNet-pretrained ResNet50 baseline, MAE, VideoMAE, and the full Emotion-LLaMA representations—within the same downstream MW classification task. Across experiments, the AffectNet-pretrained baseline remains the strongest overall encoder, while none of the newer Emotion-LLaMA-based representations improves MW prediction despite greater architectural sophistication. To understand this gap, we analyze per-encoder error profiles, prediction-score separability, shared versus encoder-specific failures, hard versus easy subsets, and Emotion-LLaMA's predicted emotion labels. Results indicate that Emotion-LLaMA– a state-of-the-art foundation model across several ER benchmarks–produces more ambiguous MW decision scores, over-predicts MW more frequently and differs only weakly across MW-relevant error cases – that stronger emotion recognition models do not necessarily provide useful features for mind-wandering detection. Our findings showcase limitations of "emotion to mind wandering" transfer, highlighting the need for development of encoders that capture learning-specific signals.[1]

Picture a fourth-grade student working through an online reading task at home. A sound from outside draws her gaze to the window, and soon her thoughts follow—drifting away from the text toward the possibility of playing outside. This shift away from the task at hand toward task-unrelated thought is known as *mind wandering* (MW) [23]. Prior work has shown that MW occurs frequently during learning and is associated with poorer learning outcomes, including reading comprehension and lecture retention [7, 21]. It is also especially prevalent in online learning settings, where sustained attention to video-based instruction can be difficult to maintain [1]. Consequently, the automatic detection of MW is an important problem for intelligent tutoring systems, adaptive educational tools, and other human–computer interactive systems that aim to monitor and support learning in real time [7, 20].

Automated MW detection is particularly useful when it can be performed continuously and non-intrusively. Much prior work has relied on eye tracking or physiological sensors such as EEG, which provide rich process information but are costly, difficult to scale, and often require controlled settings [14]. In contrast, video-based MW detection from consumer-grade webcams offers a more practical path toward deployment in natural educational environments such as classrooms and homes [3]. This makes video an appealing modality for both large-scale research and future intervention-based systems.

At the same time, video-based MW detection faces a core representation problem. Labeled MW datasets are relatively small, self-reports are expensive to collect, and task-native encoders for learning-related cognitive states are still largely unavailable. A recent line of work addressed this limitation by transferring facial emotion recognition (ER) features to MW prediction, showing that facial representations extracted using an AffectNet-pretrained ResNet50 encoder, combined with a temporal decoder, can achieve above-chance performance on an in-lab reading dataset with

[1] *Equal Contribution

MW annotations [4]. This result is promising because it suggests that large-scale ER pretraining may provide a useful starting point for MW modeling when direct supervision for cognitive states is limited.

However, it remains unclear whether improvements in ER modeling translate into better MW representations. Recent ER foundation models promise stronger robustness and broader generalization than earlier facial encoders, but stronger performance on canonical ER benchmarks does not necessarily imply better alignment with learning-related cognitive states. In educational settings, facial behavior is shaped not only by affect, but also by task demands, concentration, effort, reading dynamics, and fatigue. As a result, representations optimized for emotion recognition may remain poorly matched to the signals required for MW detection.

In this paper, we revisit ER-to-MW transfer in the context of recent foundation models by evaluating four pretrained encoders in a frozen-encoder MW pipeline on the in-lab dataset of Bosch and D'Mello [3]. Our primary baseline is the AffectNet-pretrained ResNet50 encoder used by Bühler et al. [4], which has yielded the strongest reported video-based MW results to date. We compare this baseline against MAE [24], VideoMAE [26], and Emotion-LLaMA [5]. Because Emotion-LLaMA integrates multiple pretrained visual encoders, we evaluate both the fused Emotion-LLaMA representation and key backbone components used within its pipeline.

Our results show that none of the newer or more architecturally sophisticated representations outperforms the AffectNet baseline. We then analyze per-encoder error profiles, confidence distributions, shared versus encoder-specific failures, hard versus easy subsets, and Emotion-LLaMA's predicted emotion labels to analyze why these representations fail to transfer effectively. Our findings suggest that progress in automated MW detection may require a shift in representation: we argue that facial expressions may carry different semantics in educational contexts than in standard emotion recognition tasks. Consequently, moving beyond modest performance gains will likely require the development of task-specific encoders designed for cognitive state inference in classroom settings.

1. Related Work

Successful approaches to detecting MW have relied on tracking explicit features such tracking eye movement – through eye gaze trackers[13], as well as facial features — including head pose, facial textures, facial action units, and upper body movement [3] and finally physiological features – heart rate variability [17], EGG [8], EDA [2], amongst others.

Recent advances in deep learning and computer vision have prompted a shift toward automatic feature representation learning [27]. In particular, facial emotion recognition models trained on large-scale datasets have attracted attention as potential sources of transferable knowledge for educational tasks such as mind wandering detection [18, 25]. The motivation is twofold: such models are trained on orders of magnitude more data than any mind wandering dataset, and the emotional expressions they capture — doubt, boredom, contemplation — may serve as indirect proxies for attentional states[22].

Promising results have been obtained by pairing features from a ResNet-50 [10] encoder trained on AffectNet dataset – containing 23,901 images annotated with seven distinct facial expressions (neutral, happy, sad, surprise, fear, disgust, anger) [16], with an LSTM-based classifier for mind wandering detection [4]. They evaluated on self reported mind wandering video recordings captured both in laboratory and in-the-wild learning settings, which revealed consistent performance trends achieving state of the art results compared to explicit features. Despite this progress, results leveraging transfer learning approaches from pretrained encoders for emotion recognition, are still under explored in their practical applicability to varied tasks and educational settings [19].

Multimodal emotion recognition models such as Emotion-LLaMa [5] have pushed state-of-the-art performance on common emotion recognition benchmarks. This approach integrates multiple modality-specific encoders into a LLaMA-based architecture, employing a multi-view strategy to capture complementary visual representations. Specifically, three visual encoders are employed: MAE [24], a ViT-based encoder pretrained via self-supervised masked pixel reconstruction; VideoMAE [26], which extends this masked reconstruction paradigm to the temporal domain; and EVA [9], a ViT-based encoder pretrained to reconstruct CLIP feature representations rather than raw pixels. EVA processes the *peak frame* — defined as the frame with the highest cumulative Action Unit (AU) intensity — at full resolution to capture global contextual information beyond facial expressions. An audio encoder, HuBERT [12], complements the visual stream by capturing vocal cues. Features from all four encoders are independently projected to a shared embedding space via trainable linear layers, which serve as input to a LLaMA model instruction-tuned for multimodal emotion recognition across nine categories: *happy, sad, neutral, angry, worried, surprise, fear, contempt*, and *doubt*.

This makes Emotion-LLaMA a natural candidate for transfer to mind wandering detection, where subtle facial expressions and their temporal dynamics may serve as informative cues for attentional state. In this work, we investigate whether foundation emotion recognition models can serve as effective feature encoders for downstream cognitive state inference tasks such as mind wandering detection.

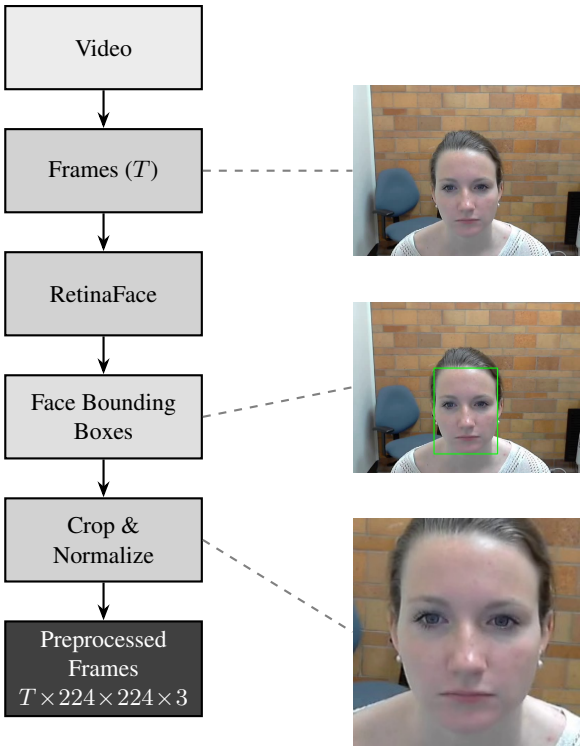


Figure 1. Face preprocessing pipeline. Raw video is sampled into T frames, from which faces are detected using RetinaFace implementation [28]. Detected bounding boxes are used to crop and normalise each face, yielding preprocessed frame tensors of size $T \times 224 \times 224 \times 3$.

2. Experiments

Dataset: Following the approach of Bühler et al. [4], we evaluate on the in-lab dataset introduced by Bosch & D'Mello [3], comprising 3,437 ten-second video recordings from 135 participants engaged in a reading task under controlled laboratory conditions. MW labels were self-caught, with positive instances drawn from a four-second window preceding each probe response. Negative instances were sampled from a thirty-second window prior to the MW report, ensuring temporal separation between classes [3]. The dataset exhibits notable class imbalance ($N = 1,031$ MW and $N = 2,406$ non-MW instances), which we address during training via class-weighted loss. The dataset is partitioned into 4 validation folds of approximately 860 samples each, which we use for cross-validated evaluation. MW is labeled as 1 (positive) and non-MW as 0 (negative).

Preprocessing: As illustrated in Figure 1, frames are first cropped to the detected face region and normalized. Face detection is performed using RetinaFace [6], with each frame cropped to the predicted bounding box. The normalization procedure follows the implementation of [4].

Feature Extractors: We evaluate four pretrained encoders as frozen feature extractors. To extract deep learning-based facial expression features, we use the approach from Bühler et al. [4], namely a CNN with a ResNet-50 [10] architecture pretrained on the AffectNet dataset [16]; hereafter, the AffectNet-pretrained ResNet-50 encoder is referred to as AffectNet. This stands as our primary baseline with the best reported results on mind wandering detection [4]. We then evaluate Emotion-LlaMA, a novel multimodal foundation model that achieves state-of-the-art results in common emotion recognition datasets [5]. As described in Section 1, Emotion-LLaMA uses four different backbone encoders: MAE, VideoMAE, EVA and Hubert. As ablations, we evaluate on the two video encoders independently: MAE and VideoMAE. Both of these models have no emotion-specific pretraining which allows us to isolate whether emotion recognition objectives specifically drive any performance gains.

Figure 2 illustrates the Emotion-LLaMA encoding pipeline which we replicate from their work [5]. VideoMAE and MAE both receive as input a sequence of T face-cropped and aligned frames at resolution $224 \times 224 \times 3$, each producing an output of dimension $T \times 1024$. The 16 frames are sampled uniformly from T. In parallel, EVA processes the *peak frame* (see Section 1) at full resolution $448 \times 448 \times 3$, yielding an output of dimension 256×1024, as no temporal pooling is applied. Each output is then independently projected to the 4096-dimensional shared embedding space of Emotion-LLaMA via the frozen linear projection layers of the original model. For implementation details on feature extraction from MAE, VideoMAE, and Emotion-LLaMA, we refer the reader to our fork of the Emotion-LLaMA repository[2]

Mind Wandering Classifier: Following the best-performing configuration of [4], we employ a three-layer LSTM [11] with 100 hidden units as the downstream classifier, followed by a softmax layer mapping outputs to binary labels: 1 (mind-wandering) and 0 (non-mind-wandering). For a full description of the pipeline, see Figure 3.

Training and Evaluation: We train and evaluate using the same cross-validation protocol as [4]. All encoders are kept frozen; only the downstream classifier is trained, using the embeddings from each encoder. Hyperparameters were selected by replicating the best-performing configuration of [4], which employs dropout regularization of 0.2 at each layer, the Adam optimizer with a cosine learning rate schedule decaying from a maximum of $1e^{-3}$ to a minimum of $1e^{-6}$, and a batch size of 16. Models are trained for up to 50 epochs with early stopping based on F1 score with a patience of 10 epochs. Full configuration files are available in our repository[3]

Implementation details: Our preprocessing and train-

[2]Contact authors for code and other supporting material
[3]Contact authors for code and other supporting material

ing pipeline supports T frames as input. Through empirical evaluation, we found $T = 64$ frames sampled uniformly from the 125 frames of each clip to yield the best results. Emotion-LLaMA is treated differently from the other encoders, as it is a generative model producing hidden states rather than direct feature embeddings. Specifically, we extract the hidden states from the final output layer. For EVA, we use full-resolution images of 448×448 pixels, consistent with the original Emotion-LLaMA implementation [5], whereas VideoMAE and MAE operate on face-cropped frames at 224×224 pixels. For all three encoders, model weights are loaded from the checkpoints provided in the Emotion-LLaMA repository[4]. The *peak frame* used as input to EVA is extracted following the implementation in the MER-Factory repository [15], where facial action units are computed using OpenFace library and peak frame — highest cumulative Action Unit (AU) intensity frame — by computing the max sum. As the original video clips contain no audio, audio features aren't passed to the HuBERT encoder or as inputs for Emotion-LLaMA. Finally, we retain the original emotion recognition instruction prompt to preserve the pretraining signal of the model (see Figure 2 for illustration and prompt).

3. Results

Table 1 reports average MW prediction performance across all four validation folds using frozen feature extractors. Across encoders, performance remains modest, with ROC-AUC values ranging from near-chance to moderately above chance. AffectNet provides the strongest overall baseline, achieving the highest average ROC-AUC, while none of the newer Emotion-LLaMA-based representations consistently improves over it. As Emotion-LLaMA integrates multiple pretrained vision encoders, we additionally report MAE and VideoMAE independently to disentangle which components of the pipeline contribute a useful signal for MW prediction. When comparing MAE, VideoMAE, and Emotion-LLaMA – the representations from MAE provide the strongest results, whereas Emotion-LLaMA attains the highest recall, yet lowest precision and near-chance ROC-AUC, indicating a strong tendency to overpredict mind wandering. Simply put – Emotion-LLaMA is predicting a lot of clips as"mind wandering". While that helps to catch many of the true MW clips (high recall), it also means many of its MW predictions are incorrect (low precision). Across folds, the relative ranking of the newer encoders varies somewhat by split, but the overall trend remains stable: the AffectNet baseline remains a strong reference point, and none of the Emotion-LLaMA-based representations surpasses it.

Table 1 reports average MW prediction performance across all four validation folds using frozen feature extractors. Across newer encoders, performance remains modest, with ROC-AUC values ranging from near-chance to moderately above chance. Because Emotion-LLaMA integrates multiple pretrained vision encoders, we report MAE and VideoMAE independently in order to disentangle which components contribute useful signal for MW prediction. AffectNet provides the strongest overall baseline, achieving the highest average F1, Precision, Recall, and ROC-AUC. None of the Emotion-LLaMA-based representations consistently improves over it despite greater architectural sophistication and stronger ER benchmark performance. Importantly, Emotion-LLaMA attains relatively high Recall but substantially lower Precision and near-chance ROC-AUC, indicating that it predicts mind wandering too often: it captures many true MW clips, but also generates many false alarms. The diagnostic analyses below are shown for one representative fold and are intended to explain the error patterns underlying these average trends.

To explain these average trends, we next analyze decision behavior and error structure on one representative fold. The following error analyses help understand *why* the newer encoders fail to improve over AffectNet and *what kinds of error* they make.

Confidence and separability. Figure 4 examines how well each encoder separates MW from non-MW clips on the representative fold. This analysis is useful because aggregate metrics alone do not show whether a model is assigning clearly different scores to the two classes or producing heavily overlapping predictions. AffectNet exhibits clearer separation between true positives and true negatives (consistent with its stronger average ROC-AUC), whereas Emotion-LLaMA's MW score distributions overlap heavily across TP/TN/FP/FN buckets, indicating weaker separability and greater ambiguity in the decision scores. MAE and VideoMAE show different confidence profiles than the full Emotion-LLaMA representation, suggesting that the individual visual backbones contribute complementary signal, but that their integration within Emotion-LLaMA does not automatically yield a stronger MW representation.

Error profile: how each encoder fails. Beyond aggregate metrics, to quantify the types of mistakes made by each encoder, we compute per-encoder confusion statistics (TP/TN/FP/FN) and derived rates on the same representative fold used in Figure 4. Unlike Table 1, which summarizes average performance across folds, Table 2 is diagnostic: it shows how errors are distributed within a given split. This matters because two models can achieve similar Recall while making very different kinds of mistakes. AffectNet achieves the most favorable error profile, with the highest balanced accuracy. Emotion-LLaMA, by contrast, exhibits the highest FP rate and the lowest specificity, in-

[4] https://github.com/ZebangCheng/Emotion-LLaMA

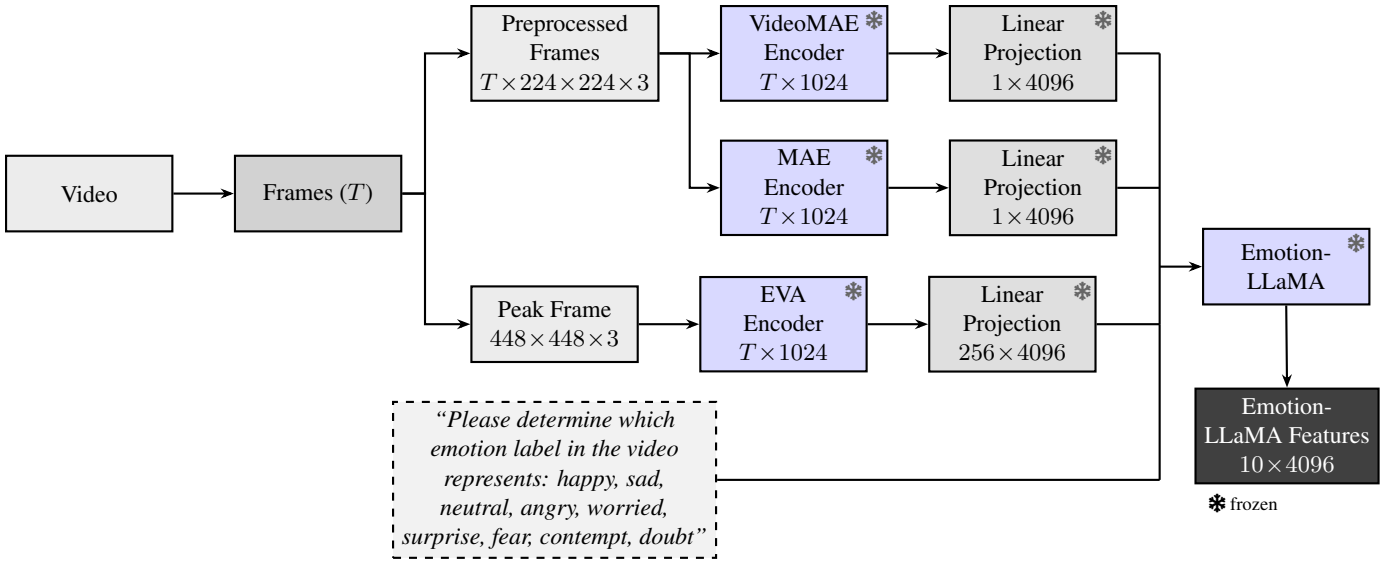


Figure 2. Emotion-LLaMA feature extraction pipeline. Raw video is sampled into T frames, preprocessed following the pipeline described in Figure 1, yielding face-cropped frame tensors of size $T \times 224 \times 224 \times 3$. These are encoded in parallel by VideoMAE and MAE. Simultaneously, the *peak frame* at resolution $448 \times 448 \times 3$ is encoded by EVA. All encoder outputs are of dimension $T \times 1024$. Each output is then projected to a 4096-dimensional shared space via the linear layers of Emotion-LLaMA. The original emotion recognition instruction prompt from Emotion-LLaMA is retained and encoded alongside the visual features. All inputs are passed jointly through the underlying LLaMA model. The first 10 hidden states from the final layer are extracted and concatenated as the feature representation for downstream classification, as this configuration yielded the best empirical results. For clarity, the tokenization step is omitted from the figure. All encoder and model weights are kept frozen throughout feature extraction.

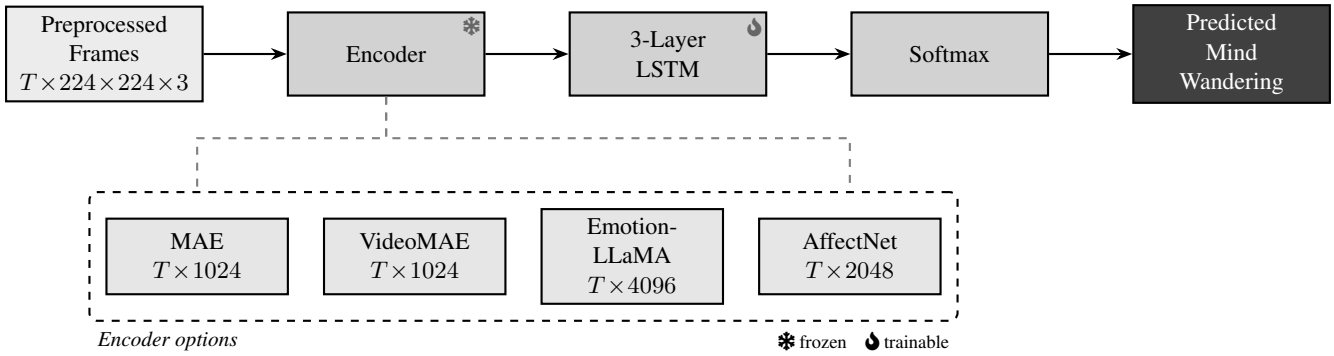


Figure 3. Classification pipeline for mind-wandering detection. Preprocessed video frames ($T \times 224 \times 224 \times 3$) are encoded by one of four frozen backbone networks and passed through a 3-layer LSTM and then a softmax layer to produce the final prediction. Encoders derived from Emotion-LLaMA pipeline are described in 2. Backbone output dimensionalities are shown beneath each option.

dicating a strong tendency to overpredict MW. MAE and VideoMAE also overpredict MW relative to AffectNet, but to a lesser extent than Emotion-LLaMA. Together, Figure 4 and Table 2 show that the weaker average performance of Emotion-LLaMA is driven not only by lower separability, but also by a less favorable balance between positive and negative errors.

Are failures shared across encoders? We next ask whether the encoders fail on the same clips or different subsets of the data. Figure 6 quantifies overlap between FP and FN sets using Jaccard similarity. This analysis helps distinguish shared representational weaknesses from encoder-specific failure modes: high overlap means two encoders tend to fail on the same clips, whereas low overlap indicates more distinct decision behavior. The general visual encoders, MAE and VideoMAE, share highly similar FP sets, while Emotion-LLaMA overlaps more strongly with MAE and VideoMAE than with AffectNet. AffectNet's lower overlap with the other encoders suggests that its er-

Encoder	F1	Precision	Recall	ROC-AUC
Affectnet	**0.4622**	**0.3657**	**0.6834**	**0.6249**
MAE	0.4283	0.3371	0.6492	0.5319
VideoMAE	0.4096	0.3128	0.6354	0.5254
Emotion-LLaMA	0.4232	0.2977	0.7425	0.5002

Table 1. **Average** performance across all four folds using extracted features from different frozen pretrained encoders. AffectNet is the strongest average baseline, achieving the highest F1 and ROC-AUC. Emotion-LLaMA attains the highest Recall, but at the cost of substantially lower Precision and near-chance ROC-AUC, indicating a strong tendency to overpredict MW.

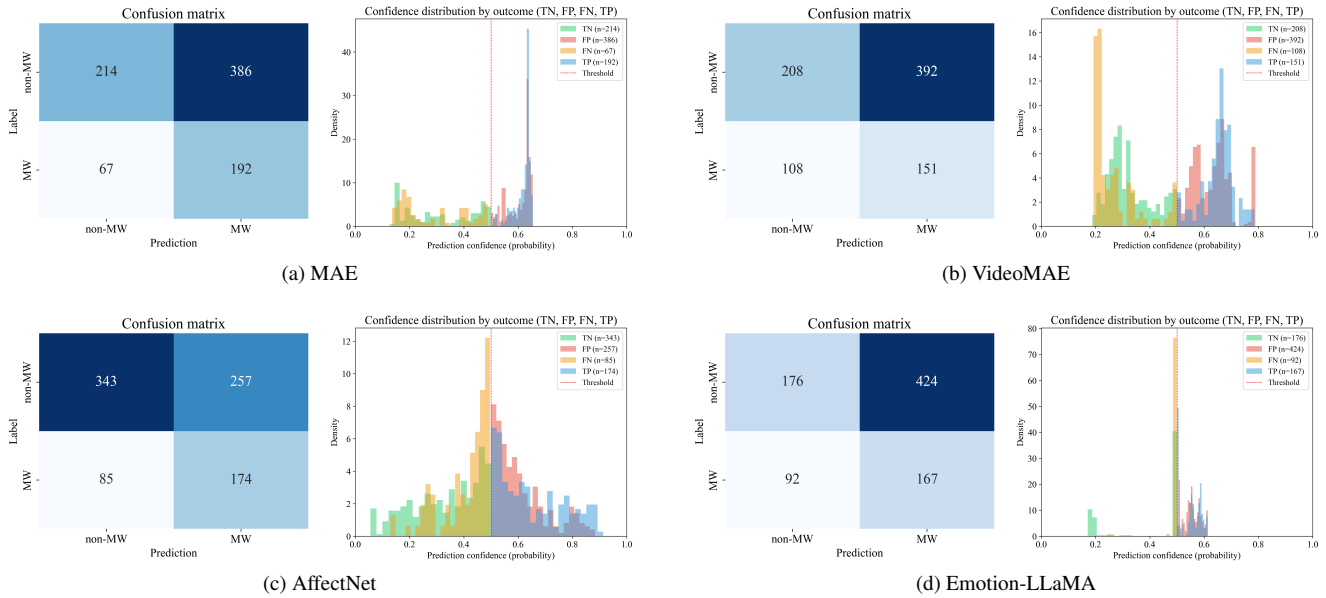


Figure 4. Confusion matrix and prediction confidence distribution (single fold). Within each sub-figure, the confusion matrix (left) and confidence histogram (right) correspond to the same model. The dashed red line denotes the classification threshold. All models exhibit a tendency to overpredict MW. AffectNet's negative-class distribution clusters near the decision threshold, reflecting uncertainty in non-MW predictions but better separability in positive cases. Emotion-LLaMA shows high confidence distributions in FN and FP, suggesting poor class separability.

rors arise from a somewhat different bias, which may help explain why it remains the strongest baseline on average.

Hard versus easy subsets. To further disentangle shared ambiguity from encoder-specific weakness, we partition clips into *easy* cases (correctly classified by all encoders), *hard* cases (misclassified by all encoders), and *encoder-specific hard* cases (misclassified by only one encoder). This analysis separates clips that are intrinsically difficult for all models from clips that are difficult only for a particular representation. Table 3 shows that a sizable subset of clips is hard for all encoders, suggesting possible ambiguity or systematic confounds in portions of the dataset. At the same time, each encoder also exhibits its own distinct hard subset, indicating sensitivity to representational choice. Although AffectNet is the strongest overall baseline, its relatively large "Hard only AffectNet" subset does not contradict Table 1; rather, it indicates that *in this fold*, AffectNet's remaining mistakes are more distinct from those made by the other encoders. This interpretation is consistent with the lower overlap of AffectNet's error sets in Figure 6.

Do predicted emotions explain MW? Finally, we examine whether Emotion-LLaMA's predicted emotion labels align with MW labels and MW-relevant error cases. Figure 5 provides the broadest view of this question by showing the distribution of top-1 predicted emotions across MW and non-MW clips. If Emotion-LLaMA's emotion taxonomy were strongly aligned with attentional state, we would expect clearly different emotion distributions for these two groups. Instead, the distributions are highly similar, suggesting that the model's top-1 emotion labels are only weakly informative for attentional state in this setting. We then sharpen this analysis by conditioning on MW errors rather than on MW labels alone.

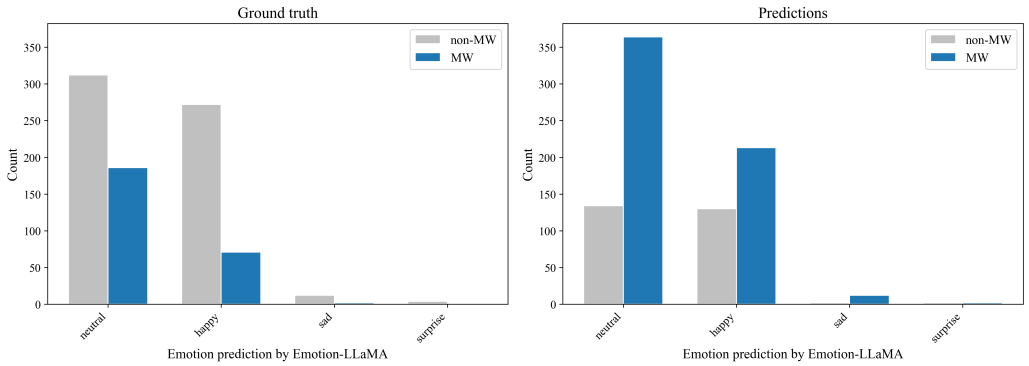


Figure 5. Emotion-LLaMA predicted emotion distribution grouped by mind wandering label. We extract the top-1 emotion prediction into four affective bins (neutral, happy, sad, and surprise) and compare their frequency across ground-truth MW labels and predicted MW labels. The proportions of MW and non-MW clips are highly similar between neutral, happy, sad and surprise, suggesting that the model's top-1 emotion taxonomy is only weakly aligned with attentional state.

Encoder	TPR (Recall)	Precision	FPR	Specificity	Bal. Acc.	F1
AffectNet	0.672	0.404	0.428	0.572	0.622	0.504
MAE	0.741	0.332	0.643	0.357	0.549	0.459
VideoMAE	0.583	0.278	0.653	0.347	0.465	0.377
Emotion-LLaMA	0.645	0.283	0.707	0.293	0.469	0.393

Table 2. Error profile by encoder (single fold). TPR=Recall. FPR and Specificity quantify MW overprediction tendencies; Balanced Accuracy is the mean of TPR and Specificity. Emotion-LLaMA exhibits the highest FPR and lowest Specificity, indicating strong MW overprediction.

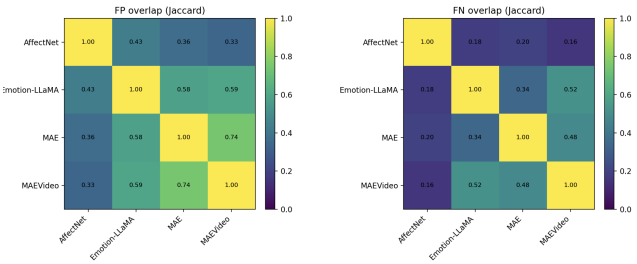


Figure 6. Error overlap across encoders. Left: false-positive (FP) overlap (Jaccard). Right: false-negative (FN) overlap. Higher overlap indicates shared failure cases; lower overlap indicates divergent error sets.

Group	#clips	#MW	#non-MW
Easy all correct	118	81	37
Hard all wrong	153	10	143
Hard only AffectNet	78	39	39
Hard only Emo-LLaMA	53	8	45
Hard only MAE	7	2	5
Hard only VideoMAE	23	14	9

Table 3. Hard/easy subset sizes across encoders (single fold). "Easy" clips are correctly classified by all encoders; "Hard" clips are misclassified by all encoders; "Hard only X" clips are misclassified only by encoder X while all others are correct.

Emotion labels conditioned on MW errors (FP vs TN). Figure 7 compares the distribution of Emotion-LLaMA's top-1 emotion labels for MW false positives (FP) versus true negatives (TN). This analysis asks whether systematic MW false alarms coincide with particular emotion predictions, or whether the emotion labels remain broadly non-discriminative even in error cases. We observe only modest shifts between the two distributions: false positives are slightly enriched for neutral and sad predictions, whereas true negatives are more frequently labeled happy. Importantly, the label space in this setting is dominated by a small set of categories (neutral/happy/sad/surprise), and the overall FP–TN differences remain limited. Thus, while Emotion-LLaMA's top-1 labels are coherent as affect descriptors, they do not provide a strong explanatory handle for distinguishing attentional state or preventing MW false alarms in the frozen-transfer setting.

Exploratory qualitative analysis of hard misclassified cases. To complement the quantitative analyses, Figure 8 presents representative hard clips that were misclassified by all evaluated models. This analysis is intended as an exploratory case study of the kinds of ambiguous facial behav-

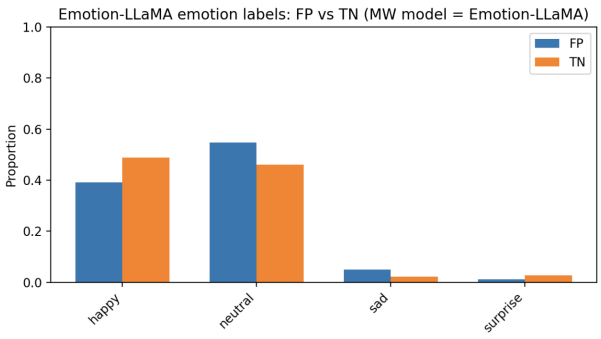


Figure 7. Emotion-LLaMA top-1 emotion labels conditioned on MW outcome buckets for the Emotion-LLaMA MW predictor. FP (false positives) vs TN (true negatives) shows only modest differences, suggesting limited alignment between predicted emotion categories and MW errors in this setting.

iors that may contribute to systematic errors. In these cases, Emotion-LLaMA's predicted emotion labels and visual descriptors often correspond to facial activity such as blinking, jaw drop, brow lowering, or momentary eye closure. These cues may plausibly reflect fatigue, effort, or context-dependent concentration rather than an overt affective state that cleanly distinguishes MW from on-task reading. The qualitative examples therefore reinforce the broader quantitative pattern: emotion-centric descriptors may capture meaningful facial activity, but they do not map cleanly onto MW labels in educational video.

Taken together, these analyses suggest that the newer Emotion-LLaMA-based representations do not fail for a single reason. Rather, they combine weaker score separability, stronger MW overprediction, shared error patterns with other general visual encoders, and emotion descriptors that seem to align only weakly with MW-relevant behavior. This helps explain why the older AffectNet baseline remains more effective on average despite the greater architectural sophistication and robust training paradigm of the newer ER models.

4. Conclusion

In this work, we revisited transfer from facial emotion recognition to mind-wandering detection in the context of recent foundation models. Using a frozen-encoder setting on an in-lab MW dataset, we compared the AffectNet baseline against representations extracted from three encoders of the pretrained Emotion-LLaMA framework. Across experiments, AffectNet representations remained the strongest baseline, and none of the Emotion-LLaMA-based features improved MW prediction.

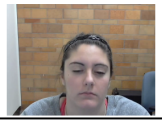
Our analyses indicate that while the newer encoders are architecturally more sophisticated and stronger on emotion

MW Label	Positive
Model prediction	False
Emotion-LLaMA conf.	0.4621
AffectNet conf.	0.3346
Predicted emotion	Happy (0.7768)
Facial expressions	Brow lowerer, Cheek raiser, Lid tightener, Lips apart, Jaw drop

(a) FN1

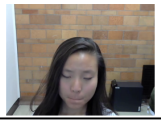
MW Label	Positive
Model prediction	False
Emotion-LLaMA conf.	0.4911
AffectNet conf.	0.4833
Predicted emotion	Neutral (0.7566)
Facial expressions	Dimpler, Blink

(b) FN2



MW Label	False
Model prediction	True
Emotion-LLaMA conf.	0.5819
AffectNet conf.	0.6613
Predicted emotion	Neutral (0.5913)
Facial expressions	Brow lowerer, Cheek raiser, Lid tightener, Lips apart, Blink

(c) FP1



MW Label	False
Model prediction	True
Emotion-LLaMA conf.	0.6135
AffectNet conf.	0.6746
Predicted emotion	Neutral (0.613)
Facial expressions	Brow lowerer, Dimpler, Blink

(d) FP2

Figure 8. Qualitative examples of hard misclassified cases. Top row: false negatives (FN1, FN2); bottom row: false positives (FP1, FP2). For each case, we show the participant image, the groundtruth MW label, the shared MW prediction across models, and the corresponding MW confidence scores obtained using features extracted from the Emotion-LLaMA and AffectNet encoders. We additionally report the predicted emotion label and the associated facial-expression descriptors to illustrate how ambiguous visual cues may contribute to misclassification.

recognition benchmarks, they produced more ambiguous MW decision scores, overpredicted mind wandering more frequently, and exhibited error patterns that overlapped substantially with other general visual encoders.

Taken together, these findings suggest that emotion-pretrained representations can provide a useful starting point for video-based MW detection, but that stronger emotion recognition models do not automatically yield better frozen-transfer features for educational cognitive-state inference. Progress beyond modest above-chance performance will likely require moving beyond emotion-centric transfer toward lightweight, task-specific encoders trained on classroom-relevant behavioral signals.

Acknowledgements

This research was supported by the National Science Foundation (DRL 1920510 and 2019805). The opinions expressed are those of the authors and do not represent views of the funding agency.

References

[1] Vagner Beserra, Miguel Nussbaum, and Macarena Oteo. On-task and off-task behavior in the classroom: A study on mathematics learning with educational video games. *Journal of Educational Computing Research*, 56(8):1361–1383, 2019. 1

[2] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. Automated physiological-based detection of mind wandering during learning. 2014. 2

[3] Nigel Bosch and Sidney K. D'Mello. Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing*, 12(4):974–988, 2021. 1, 2, 3

[4] Babette Bühler, Efe Bozkir, Patricia Goldberg, Ömer Sümer, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. From the lab to the wild: Examining generalizability of video-based mind wandering detection. *International Journal of Artificial Intelligence in Education*, 35, 2024. 2, 3

[5] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems*, pages 110805–110853. Curran Associates, Inc., 2024. 2, 3, 4

[6] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild, 2019. 3

[7] Sidney D'Mello and Arthur C. Graesser. Feeling, thinking, and computing with affect-aware learning technologies. In *The Oxford Handbook of Affective Computing*. Oxford University Press, 2015. 1

[8] Henry W. Dong, Caitlin Mills, Robert T. Knight, and Julia W. Y. Kam. Detection of mind wandering using eeg: Within and across individuals. *PLOS ONE*, 16(5):1–18, 2021. 2

[9] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2, 3

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 3

[12] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *CoRR*, abs/2106.07447, 2021. 2

[13] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R. Brockmole, and Sidney K. D'Mello. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, 29(4):821–867, 2019. 2

[14] Christina Yi Jin, Jelmer P Borst, and Marieke K Van Vugt. Predicting task-general mind-wandering with eeg. *Cognitive, Affective, & Behavioral Neuroscience*, 19(4):1059–1073, 2019. 1

[15] Yuxiang Lin and Shunchao Zheng. MER-Factory, 2025. 4

[16] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. 2, 3

[17] Phuong Pham and Jingtao Wang. Attentivelearner: Improving mobile mooc learning via implicit heart rate tracking. In *Artificial Intelligence in Education*, pages 367–376, Cham, 2015. Springer International Publishing. 2

[18] Jamie Pordoy, Haleem Farman, Nevena Kostadinova Dicheva, Aamir Anwar, Moustafa M Nasralla, Nasrullah Khilji, and Ikram Ur Rehman. Multi-frame transfer learning framework for facial emotion recognition in e-learning contexts. *IEEE Access*, 12:151360–151381, 2024. 2

[19] Ikram Qarbal, Nawal Sael, and Sara Ouahabi. Student's engagement detection based on computer vision: A systematic literature review. *IEEE Access*, 2025. 2

[20] David N. Rapp. The value of attention aware systems in educational settings. *Computers in Human Behavior*, 22(4): 603–614, 2006. Attention aware systems. 1

[21] Evan F. Risko, Nicola Anderson, Amara Sarwal, Megan Engelhardt, and Alan Kingstone. Everyday attention: Variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology*, 26(2):234–242, 2012. 1

[22] Krist Shingjergji, Deniz Iren, Corrie Urlings, and Roland Klemke. Affective computing in online higher education: A systematic literature review. *Computers and Education: Artificial Intelligence*, page 100499, 2025. 2

[23] Jonathan Smallwood and Jonathan W. Schooler. The restless mind. *Psychological Bulletin*, 132(6):946–958, 2006. 1

[24] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Maedfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 6110–6121, 2023. 2

[25] Licai Sun, Xingxun Jiang, Haoyu Chen, Yante Li, Zheng Lian, Biu Liu, Yuan Zong, Wenming Zheng, Jukka M Leppänen, and Guoying Zhao. Learning transferable facial emotion representations from large-scale semantically rich captions. *arXiv preprint arXiv:2507.21015*, 2025. 2

[26] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 2

[27] Aili Wang, Haibin Wu, and Yuji Iwahori. Advances in computer vision and deep learning and its applications, 2025. 2

[28] Elliot Zheng. batch-face. https://github.com/elliottzheng/batch-face, 2021. Accessed: 2026-01-10. 3