



From Emotion Recognition to Mind-Wandering Detection

A Comparative Analysis of Video-Based Emotion Foundation Models



Institute for Cognitive Science
Emotive Computing Lab

Ekta Sood*, Sebastian Ricke*, Trisha Mittal, Sidney D'Mello
University of Colorado Boulder, University of Colorado Boulder, Dolby Laboratories, University of Colorado Boulder

University of Colorado, Boulder

Motivation

- Predicting cognitive states such as mind wandering during learning is important for adaptive educational technologies
- Video-based mind-wandering detection is promising → continuous and non-intrusive, but progress is limited by data scarcity
- Prior work leveraged pretrained emotion recognition model – transfer learning to MW pred → ER models may provide useful visual representations for MW detection
- Do more recent and stronger emotion-recognition foundation models improve frozen-transfer performance for video-based mind-wandering detection?

Experiments

- Building on SOTA transfer-learning pipeline, keep downstream MW classifier fixed
 - swap encoder w. newer pretrained ER
- Baseline: AffectNet-pretrained ResNet50
- Newer encoders: MAE, VideoMAE, Emotion-LLaMA

Evaluation

- Metrics: F1, Precision, Recall, and ROC-AUC
- Analysis: How each encoder fails
 - prediction confidence and score separability
 - FP vs. FN error profiles
 - shared vs. encoder-specific failures
 - predicted emotion labels vs. MW errors

Methodology

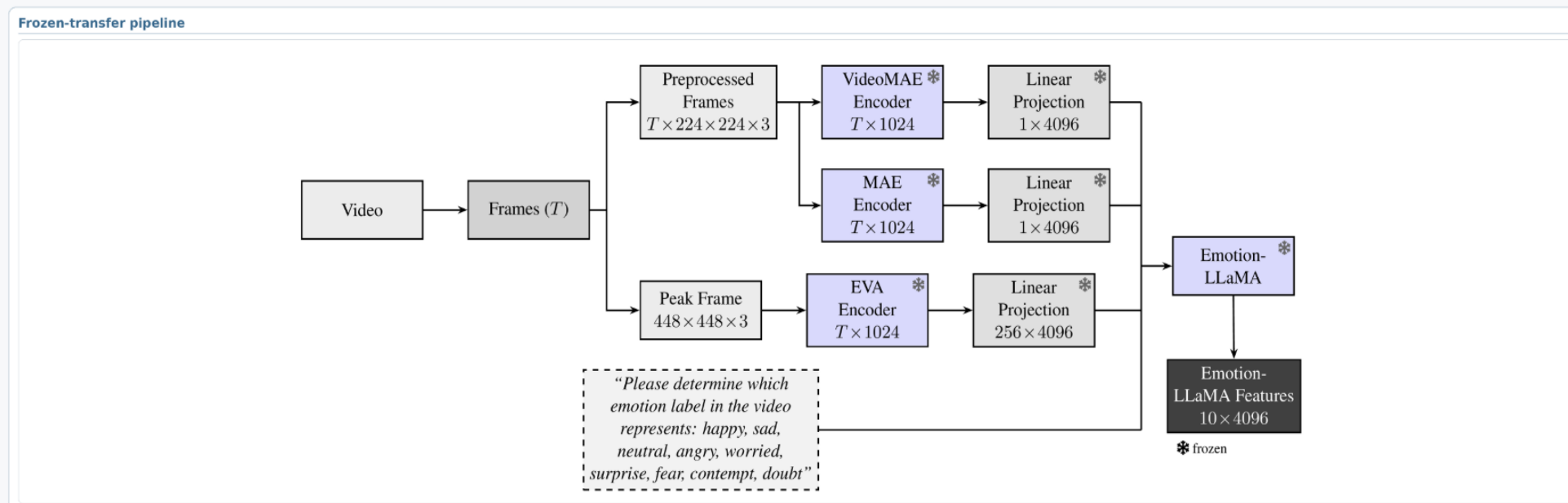
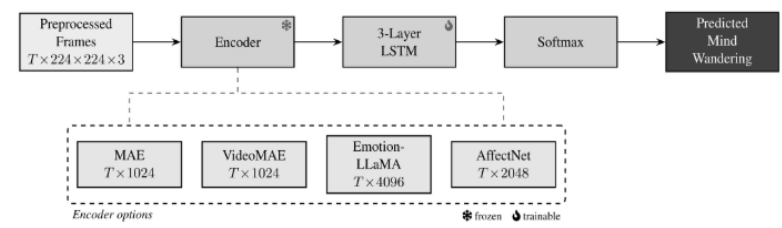


Figure 2: Emotion-LLaMA feature extraction pipeline (all encoders/model weights frozen)

Downstream MW classifier



Main result: stronger ER models do not win

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

AffectNet
SOTA avg. F1 + ROC-AUC
F1 0.462 | AUC 0.625

Emotion-LLaMA
Highest Recall, low Precision
Recall 0.743 | Prec. 0.298

Interpretation
many true MW clips caught, many false alarms
MW overprediction

Table 1: average performance across four folds

Takeaway The newer Emotion-LLaMA-based representations do not surpass the AffectNet baseline, despite greater architectural sophistication and stronger ER benchmark performance.

Score behavior: separability and false alarms

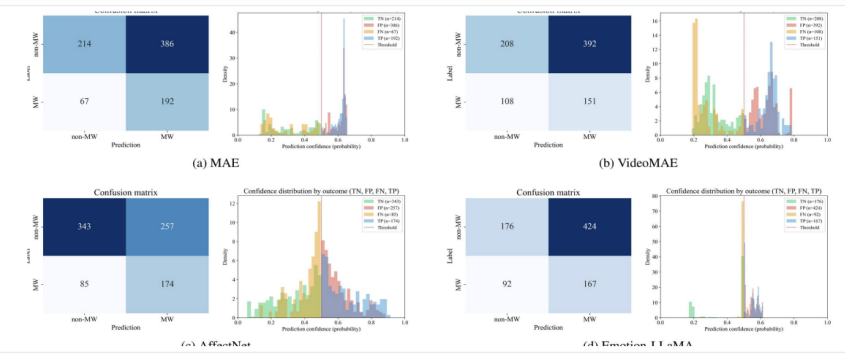


Figure 4: confidence distributions overlap; Emotion-LLaMA shows poor class separability and high-confidence errors

Analysis

1. Error type: overprediction

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: Emotion-LLaMA has the highest FPR and lowest specificity on the representative fold.

2. Shared vs. unique failures

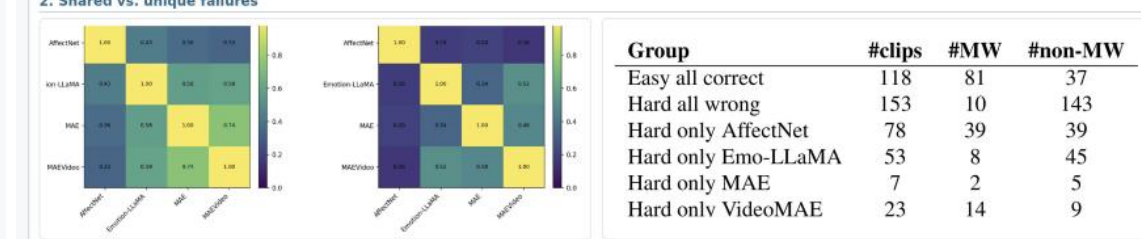


Figure 3: MAE, VideoMAE, and Emotion-LLaMA share more failure structure; AffectNet errors are more distinct.

3. Do predicted emotions explain MW?

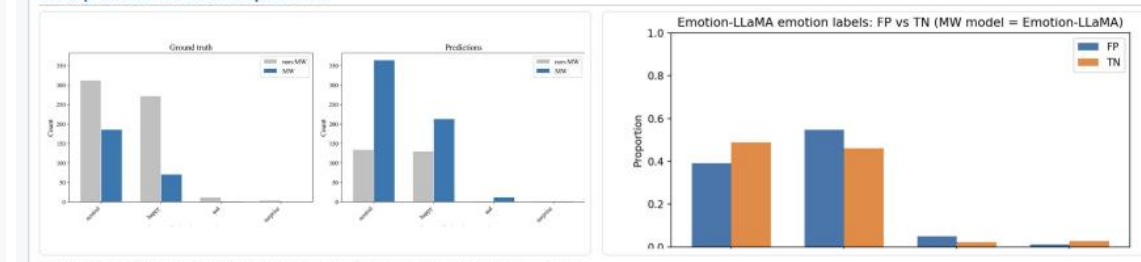


Figure 5 and 7: top-1 emotion labels are dominated by neutral/happy/sad/surprise and only rarely align with MW labels or false alarms.

4. Hard cases: ambiguous facial behavior

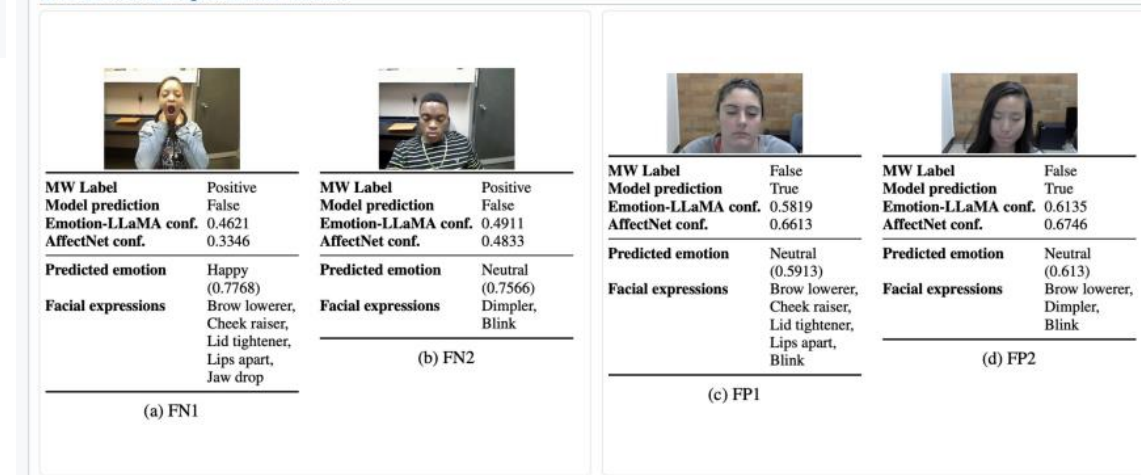


Figure 6: blinking, brow lowering, jaw drop, and eye closure may reflect fatigue, effort, or concentration – not a clear emotion-to-MW mapping.

Outlook

SOTA ER performance → poorer MW prediction

- emotion descriptors align weakly with MW-relevant behavior, weaker class separability, overpredict.
- Move beyond emotion-centric transfer → task-