Evaluating Web-trained Facial Expression Recognition in Collaborative Problem-Solving

Anonymous CVPR submission

Paper ID 21

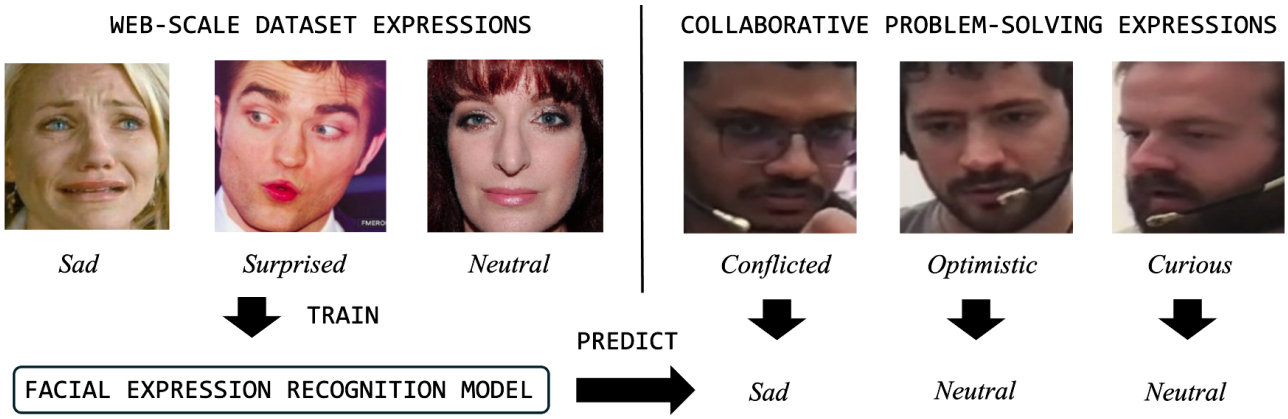

Figure 1. Popular facial expression recognition (FER) models are trained to predict canonical basic emotions from prototypical facial expression images, often obtained from web queries and annotated with external observation or query-matching. In collaborative problem-solving (CPS), epistemic emotions such as *curiosity*, *optimism*, and *conflict* are often more relevant but manifest more subtly in facial behavior and are subjective in their nature. We examine the alignment between these model outputs and epistemic affect observed in collaborative problem-solving settings. Samples from the left group are from AffectNet [1] and from the right are from our CPS dataset.

Abstract

Facial expression recognition (FER) have been adopted extensively in educational research. These models are usually trained on web-scale datasets optimized for canonical basic emotions. However, in collaborative learning, the affective states of interest are often epistemic (learning-relevant) in nature—such as confusion, curiosity, and frustration—rather than prototypical basic emotions like disgust or anger. Here, we evaluate popular FER systems trained using web-scale datasets on small-group collaborative problem-solving sessions. We annotate epistemic emotions using a retrospective cued-recall approach, where participants watch a video of a collaborative session immediately after the group task and individually self-report at different segments. We analyze cross-taxonomy alignment between epistemic emotions and predicted basic emotions, cross-model agreement in collaborative learning, and dimensional valence–arousal structure with respect to basic and epistemic emotions. Across models, categorical predictions overwhelmingly collapse to Neutral for all epistemic labels, and valence–arousal outputs fail to distinguish the different epistemic states. Furthermore, low agreement between models indicates some degree of instability in both web-scale and naturalistic collaborative contexts. Our findings suggest that FER systems trained for canonical emotion recognition on web-scale datasets do not directly generalize to the more subtle epistemic emotions experienced during collaboration, which has implications for education researchers deploying these tools in classrooms. Our code is available at https://anonymous.4open.science/r/cvpr-edu-affect-2026-6EF5.

1. Introduction

Recent advances in computer vision have increased the adoption of facial expression recognition (FER) tools in

CVPR #21

education. Open-source frameworks [2, 3], large-scale datasets [1, 4], and modern architectures [5] have lowered the barrier to deploying FER systems in applications such as classroom sensing and intelligent tutoring systems [6–9]. Despite this rapid adoption, relatively little empirical work has examined how web-trained FER systems perform in naturalistic multi-party learning environments such as collaborative problem-solving (CPS). Most FER models are trained to predict canonical basic emotions—such as happiness, sadness, anger, fear, disgust, and surprise—using datasets curated primarily from scraping or querying the web [1, 10] and annotated by external coders with subsets including one annotator per image [11] or computer vision algorithms that are trained on similar manual annotations [4]. Facial expressions from these datasets tend to be more prototypical and exaggerated compared to naturalistic collaborative data since many are stock images of posed expressions or celebrities at televised events, and have been criticized for being inconsistent in their labels due to their unreliable annotation process [11]. Recent affective computing research has also expanded to dimensional representations, most notably the valence–arousal (V–A) circumplex model [12]. Many modern FER datasets provide both categorical and continuous annotations [1, 11], enabling models to predict valence and arousal alongside discrete emotions. Such datasets typically also include soft-labels (the probability vectors for discrete labels) with each instance, and offers a potentially richer representational space [13].

In contrast, the affective states more relevant to educational contexts such as CPS are epistemic in nature, including confusion, curiosity, frustration, and engagement, and they manifest in facial expression much more subtly. These emotions are closely tied to knowledge construction, uncertainty management, and collaborative coordination, and have been shown to influence learning outcomes in various educational contexts [14–16]. Consequentially, they are of particular interest for instructors and learning technologies that aim to monitor and support learning processes in real time [14]. To that end, researchers are using machine learning with behavioral data that can unobtrusively capture these states during authentic learning activities [17]. FER systems are becoming promising candidates for this purpose due to their increasing efficiency and ease of adoption [6].

These trends bring some important questions to attention: when web-trained FER models are used in collaborative learning settings, do they meaningfully align with epistemic affective states or collapse toward dominant basic emotion categories? Moreover, do the continuous valence–arousal outputs of VA models capture finer affective distinctions in this context, or do they largely reflect the same categorical structure embedded in their training data? Finally, do different FER systems/architectures even produce consistent signals under such a domain shift?

To address these questions, we evaluate multiple widely used pretrained FER systems on a dataset of CPS sessions annotated with self-reported epistemic states collected via retrospective cued recall [18, 19]. We analyze cross-taxonomy alignment, prediction concentration and inter-model agreement. Our goal is to examine how contemporary FER systems behave out-of-the-box when deployed in CPS settings.

2. Related Work

The last decade has seen significant interest in FER, primarily driven by web-scale datasets and deep learning. Early large-scale annotated datasets such as FER2013 [10] enabled supervised training of CNNs for basic emotion recognition. AffectNet [1] and RAF-DB [20] further expanded the scale of training data and introduced both categorical and valence–arousal annotations, with AffectNet+ [11] also introducing soft labels for facial expressions. Recent architectures have even leveraged attention mechanisms and transformer-style backbones to improve performance on these benchmarks. For example, dual-branch attention networks [21] and ViT based architectures such as POSTER++ [5] reported strong results on multiple in-the-wild emotion datasets. Concurrently, toolkits such as OpenFace [2, 3] and LibreFace [22] have been developed by researchers to provide user-friendly and efficient interfaces for end-to-end extraction of facial landmarks, action units, gaze, head pose estimation and facial expression prediction, facilitating adoption in domains such as education.

Most training and evaluation of FER models remain within the distribution of the web-scale benchmark datasets. Such datasets are often constructed by querying the web with emotion-related keywords [1, 4, 10] using tools such as WordNet [23]. This process have yielded posed or exaggerated expressions that approximate canonical emotion prototypes [13]. Such data differ substantially from spontaneous affect observed in real-world interaction in terms of intensity, dynamics, and contextual embedding. In collaborative learning settings, learners often regulate or mask visible affect during group interaction [24]. Negative epistemic states such as confusion or frustration may manifest subtly or co-occur with task-focused expressions. Group dynamics introduce additional layers of co-regulation and social monitoring [25]. As a result, the mapping between visible facial configurations and epistemic emotions may differ from that observed in web-scale datasets. In the learning sciences, affect is often studied not as isolated emotional expressions but as part of a broader cognitive–affective process that unfolds during knowledge construction. Epistemic emotions such as confusion, curiosity, and frustration are closely tied to processes such as impasse resolution, hypothesis gen-

eration, and collaborative sense-making [14, 15]. These states are frequently transient, low-intensity, and embedded within ongoing task activity, making them difficult to infer from facial cues alone.

Many web-scaled datasets also use the basic emotion taxonomies and/or the dimensional affect models to represent facial affect. Basic emotion theory characterizes affect using a small set of discrete categories—such as happiness, sadness, anger, fear, disgust, and surprise—that are assumed to correspond to recognizable facial configurations and are widely used in FER datasets and benchmarks [10, 26]. Datasets using this taxonomy often use a "Neutral" label if the annotator believes none of the basic emotions are present. Dimensional approaches, most notably the valence–arousal circumplex model, instead represent affect along continuous axes capturing positivity–negativity (valence) and activation level (arousal) [12]. These representations have enabled FER systems to predict both categorical and continuous affective states from facial images [1]. In educational contexts, it remains unclear whether the epistemic states of interest correspond directly to canonical facial prototypes associated with basic emotions, or map cleanly onto simple regions of valence–arousal space. Understanding how existing FER representations relate to these learning-relevant affective states is therefore an important step to consider before deploying such models in collaborative learning settings.

Another concerning factor is the annotation of affect in educational contexts. Many FER benchmarks rely on annotations from external observers who infer emotional states from static images or short clips [1, 20]. Although inter-rater agreement is often reported, observer-based labeling may not fully capture internal experiences, particularly in complex social contexts. Some subsets of popular datasets also only have one annotator assigned per image due to the sheer volume of images. Benitez-Quiroz et al. [4] have tried to improve this paradigm by generating annotations for millions of facial images using automated facial behavior analysis systems trained on manually coded Facial Action Units. While such approaches enable dataset scaling, subsequent work has raised concerns about the reliability of labels generated by automated FER systems, noting that algorithmic annotations may propagate biases and systematic errors present in the underlying models [27]. Educational research has therefore explored alternative annotation strategies, including self-reports, behavioral coding, and multimodal sensing approaches [7, 14]. Physiological signals such as electrodermal activity [28] and heart rate variability [29] have been used to study learning-related affect, but these methods are often intrusive and impractical in authentic learning environments. A recently adopted approach is retrospective cued-recall (RCR), in which participants review recordings of their own activity and report their internal states at specific moments in time [18]. This method preserves contextual grounding while allowing participants to reflect on their affective experiences during collaboration. Such annotations provide a complementary perspective to observer-based labels and offer a practical mechanism for studying epistemic emotions in naturalistic settings.

Table 1. Demographic information of participants in the EECPS-WT dataset [19]

Gender	Male (18)	Female (7)	Non-binary (2)
Age	18–24 (14)	25–34 (11)	35+ (2)
Ethnicity	Asian (13)	White (12)	Hispanic (2)

Figure 2. Three team members collaborating during an experimental task from the EECPS-WT dataset

3. Methods

3.1. Dataset

We use a collaborative learning dataset, Epistemic Emotions in Collaborative Problem Solving: Weights Task (EECPS-WT). This dataset is a contribution from previous work [redacted for review] which we build upon in this paper. The data is collected from 27 college students who were over 18 years old and fluent in English. Table 1 shows their demographic information. Participants were organized into nine groups of three and were video recorded while completing a collaborative problem-solving (CPS) activity, adapted from Khebour et al [30]. In this activity, groups used a balance scale to infer the relative weights of five blocks, one of which has a known reference weight. Each group was required to submit a single shared answer via survey upon completion of the phase. Interactions were recorded using consumer-grade laptop webcams positioned in front of participants, resulting in naturalistic, low-resource video data similar to what may be encountered in classroom or remote learning settings. Figure 2 shows an example frame from the dataset, where three team members are collaborating on the task.

Following task completion, participants individually reviewed a recording of their collaborative session using a retrospective cued recall (RCR) procedure. During playback, participants reported their own affective states using an interactive survey interface. Reports could be self-initiated at any time. Additionally, if no report was submitted for 60 seconds, the system automatically issued a probe requesting a report. This design captures both spontaneous and probe-caught affective states. Participants selected one or more affective states from a predefined set of seven epistemic labels: Confused, Curious, Frustrated, Disengaged, Optimistic, Surprised, and Conflicted, adopted from prior research [31]. Each selected label was timestamped according to video playback time. All reports were standardized and consolidated into a single dataset, with each affective label represented as a separate instance. Reports were additionally tagged as self-caught or probe-caught based on their trigger mechanism. Statistical information such as frequency distribution and temporal distribution of labels is reported in Sec 1 of the supplementary material.

3.2. Data Preprocessing

To evaluate facial expression recognition (FER) models at the level of individual reports, we construct face-based image instances aligned with self-reported timestamps. All videos were processed frame-by-frame using RetinaFace for face detection [32]. For each detected face, a bounding box was extracted and cropped from the original frame. Since participants were seated in fixed positions during the collaborative task, with participants 1 through 3 seated from left to right from the camera's point of view, detections were organized per participant by spatial consistency across frames using the x-coordinate of detections to map them to participants. Cropped faces were stored at one-second resolution, with each second containing approximately 30 frames corresponding to the recorded frame rate. All subsequent model inference operates exclusively on these cropped face images.

Each report is associated with a specific playback timestamp. To account for potential temporal misalignment between when an affective state was experienced and when it was reported, we adopt a temporal neighborhood sampling strategy. For each affect report at time t, we collect all cropped face images within a ± 5 second window, i.e., the interval $[t-5,\ t+5]$. From this pool of candidate frames, we randomly sample 10 face images to represent the instance. This strategy reduces sensitivity to single-frame noise and accommodates slight reporting delays inherent in retrospective annotation. Each affect report is therefore represented as a set of 10 independently processed face images. Models are applied independently to each of the 10 frames, and predictions are aggregated to produce an instance-level output (described in Section 3.3).

3.3. Facial Expression Recognition Models

We benchmark FER models that output (i) categorical basic emotions or (ii) continuous valence–arousal representations. We perform inference independently on each sampled face crop, and predictions are aggregated at the instance level. For a given affect report at time t, let $\{x_i\}_{i=1}^{K}$ denote the set of $K = 10$ sampled face images within the temporal window described in Section 3.2. Each image is resized to 224×224 pixels and normalized according to the preprocessing requirements of the corresponding model.

3.3.1. Categorical Basic Emotion Models

We evaluate multiple pretrained FER models that predict discrete basic emotion categories. These include both multi-task pipelines OpenFace 3.0 [3] and LibreFace [22] and off-the-shelf deep learning FER models POSTER++ [5], EmotiEffLib [33], and DDAMFN [21]. In addition, we include Qwen2.5-VL-7B [34], a vision language model, as a general purpose comparison point. OpenFace-3.0 outputs class probabilities (also known as soft labels [11]) and the other five outputs hard labels. All six models are used with publicly released pretrained weights.

For models that output soft labels, each image x_i produces a probability vector:

$$\mathbf{p}_i = f(x_i) \in \mathbb{R}^C, \quad \text{with} \quad \sum_{c=1}^{C} p_{i,c} = 1.$$

To obtain an instance-level prediction, we compute the mean probability vector across sampled frames:

$$\bar{\mathbf{p}} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{p}_i,$$

and assign the predicted label as

$$\hat{y} = \arg\max_c \bar{p}_c.$$

Averaging probabilities preserves confidence information across frames and reduces sensitivity to single-frame noise.

For models that return only a discrete label per image, we aggregate predictions via majority vote:

$$\hat{y} = \text{mode}\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_K\}.$$

In the case of ties, a deterministic class ordering is used. Majority voting is appropriate when per-frame confidence scores are not available or not directly comparable across models.

For Qwen 2.5-VL, we provide the 10 frames along with this text prompt: *These images are frames from a single video. Classify the emotion being displayed. Choose exactly one from the universal emotions. Output ONLY the emotion name, and nothing else.*

Because emotion labels vary slightly across different model outputs (e.g., "Happiness" vs. "Happy"), predicted categories are mapped to a standardized set: {*Neutral, Happy, Sad, Angry, Surprise, Fear, Disgust, Contempt*}. *Contempt* was removed during cross-model analysis since some categorical models does not use it as one of their outputs [35]. This normalization enables cross-model comparison under a unified taxonomy.

3.3.2. Dimensional Valence–Arousal Model

To examine whether continuous affect representations better align with collaborative learning contexts, we evaluate a pretrained system that jointly predicts discrete emotions and continuous valence–arousal values: EmotiEffLib [33]. For each image x_i, the model outputs:

$$(v_i, a_i) = f_{\text{VA}}(x_i),$$

where $v_i \in [-1, 1]$ denotes valence and $a_i \in [-1, 1]$ denotes arousal.

Instance-level dimensional predictions are computed by averaging across sampled frames:

$$(\bar{v}, \bar{a}) = \frac{1}{K} \sum_{i=1}^{K} (v_i, a_i).$$

This produces a single valence–arousal coordinate for each affect report instance.

We evaluate a single representative dimensional model rather than benchmarking multiple architectures. Because valence–arousal models are typically trained on the same web-scale datasets [1, 36] and share similar objective functions, we picked one model as our representative to examine whether continuous affect representations meaningfully differentiate epistemic states in collaborative learning.

3.4. Analysis

Our experiments focus on understanding how FER models perform on collaborative learning data. Because epistemic emotions and basic emotions/valence-arousal represent different conceptual spaces, we do not treat this as a conventional supervised classification problem. Instead, we analyze model behavior through distributional alignment, prediction concentration, and cross-model agreement.

3.4.1. Categorical Affect Alignment

We first examine how predicted basic emotion categories distribute across epistemic labels. For each categorical model, we compute a cross-taxonomy confusion matrix whose rows correspond to epistemic emotions (e.g., Curious, Confused, Frustrated) and columns correspond to predicted basic emotions. To account for the different distributions of labels in the dataset, we report the row-normalized version of this matrix, representing the conditional distribution of predicted basic emotions given each epistemic state. This analysis allows us to assess whether specific epistemic states map consistently onto particular basic emotions, or whether predictions collapse toward dominant categories. To maintain clarity and conciseness, we only report the row-normalized confusion matrix for OpenFace 3.0 as the representative categorical model and summarize results across all other evaluated models in Sec 2.1 of the supplementary material. We also compute the proportion of instances assigned to each basic emotion category and report the most frequent predicted basic emotion for each epistemic label across all categorical models.

3.4.2. Dimensional Affect Structure

We next analyze continuous valence–arousal predictions produced by the dimensional model. For each epistemic label, we compute the mean and standard deviation of valence and arousal values and visualize their distributions using scatter plots in the two-dimensional affect space. This analysis addresses whether dimensional representations provide separable structure for epistemic states that categorical models fail to capture. We examine if epistemic states form clusters in valence–arousal space, the degree of overlap between states, and the stability of dimensional predictions across instances. In addition, we analyze valence–arousal distributions grouped by predicted basic emotion categories to verify internal coherence of the dimensional outputs and how the basic emotions complement the dimensional predictions. Here, we report scatter plots in the valence-arousal space to represent the distributions from both epistemic labels and basic emotion labels. Detailed statistical reports of these distributions and box plots are in Sec 2.2 of the supplementary material.

3.4.3. Cross-Model Agreement

To assess robustness across architectures under domain shift, we computed pairwise confusion matrices between categorical predictions generated by two FER models on the same instances. This analysis was conducted across random samples from EECPS-WT and AffectNet [1], a large-scale, in-the-wild facial expression benchmark. Pairwise confusion matrices were then constructed by treating the Model 1 predictions as the reference axis and the Model 2 predictions as the comparison axis. We normalize the counts of predictions from Model 1 for each label to counter the influence of label imbalance in evaluating cross-model agreement. This allows us to directly compare inter-model agreement between these FER models in two different datasets, one that is web-scaled and the other educational. Here, we report confusion matrices representing agreement between OpenFace 3.0 and LibreFace as the representative models. Confusion matrices between all other combinations of models is reported in Sec 2.3 of the supplementary material.

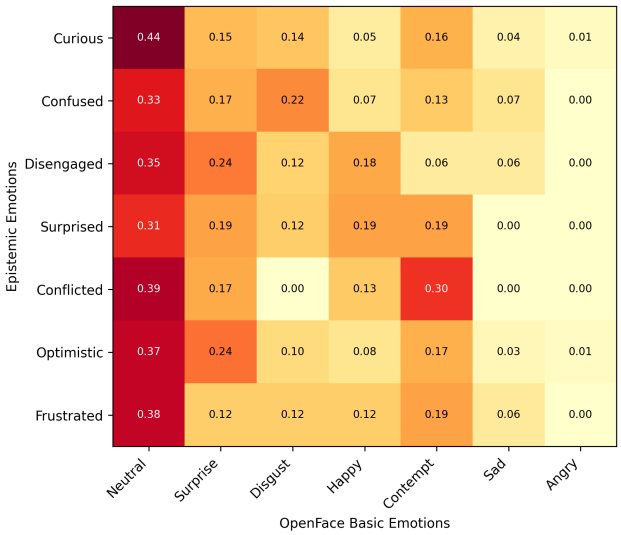

Figure 3. Row-normalized confusion matrix between epistemic states and OpenFace 3.0 predictions.

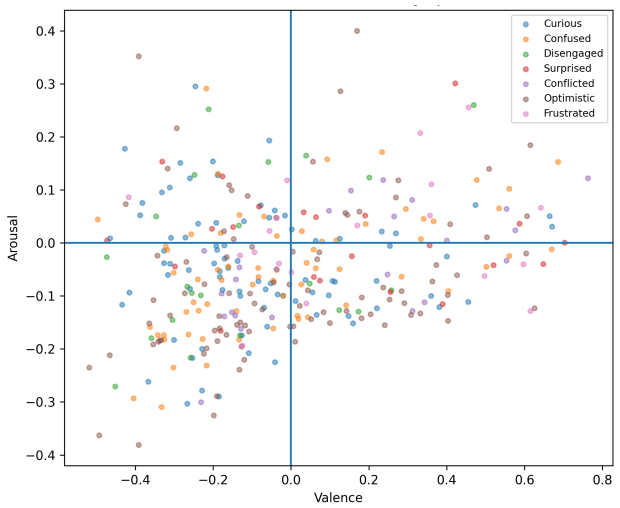

Figure 4. Valence–arousal scatter plot grouped by epistemic labels. Substantial overlap and high within-label variance are observed, with no clearly separable clusters.

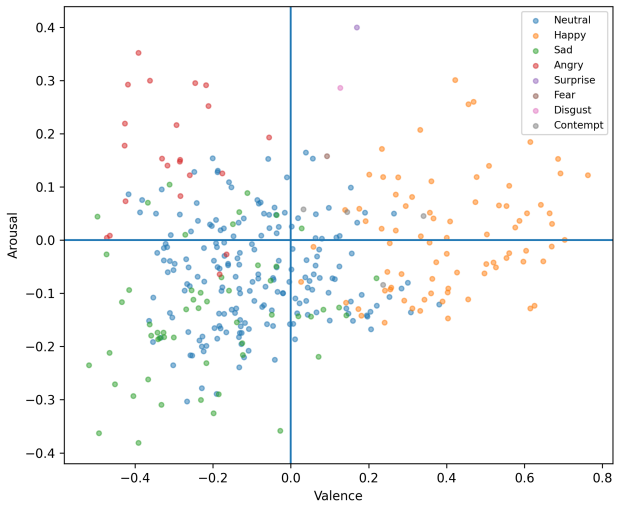

Figure 5. Valence–arousal scatter plot grouped by predicted basic emotion categories. Distinct clusters emerge (e.g., Happy in high valence, Angry in low valence/high arousal), indicating internal coherence of the dimensional model with respect to its training taxonomy.

4. Results

We report results for our experiments in: (1) cross-taxonomy alignment between predicted basic emotions and epistemic labels, (2) dimensional valence–arousal structure, and (3) cross-model agreement under domain shift.

4.1. Cross-Taxonomy Alignment

Figure 3 shows the row-normalized confusion matrix between epistemic labels and OpenFace 3.0 predictions. We see a consistent pattern across all epistemic states: *Neutral* is the most frequent prediction. For example, 43.8% of Curious instances, 33.3% of Confused instances, and 37.5% of Frustrated instances are predicted as Neutral. No epistemic label shows a strong one-to-one mapping to any basic emotion category. This pattern persists across all evaluated categorical models. Table 2 summarizes, for each epistemic label, the most common predicted basic emotion and its corresponding proportion. Across models, Neutral is overwhelmingly the dominant prediction for nearly all epistemic states. In the few cases where another label appears most frequently (e.g., Happy or Sad), the mapping is inconsistent across models and lacks an interpretable semantic relationship to the epistemic construct.

4.2. Valence–Arousal Structure

We next analyze continuous valence–arousal (V–A) outputs. Figure 4 visualizes the distribution of epistemic labels in V–A space. We observe that epistemic states do not form clearly separable regions, as the scatter plot reveal substantial overlap across labels, showing large within-label variance relative to between-label differences. When grouped by predicted *basic emotion* categories, we observe a more clear structure (Figure 5). For example, Happy predictions cluster in high-valence regions, Sad in low-valence and lower-arousal regions, and Angry in low-valence but higher-arousal regions.

4.3. Cross-Model Agreement

Finally, we assess robustness across architectures under domain shift by analyzing pairwise confusion matrices between categorical predictions generated by OpenFace 3.0 and LibreFace as representative models on subsets of the

Table 2. Most frequent predicted basic emotion for each epistemic label and their proportion across categorical models.

Epistemic Label	OpenFace 3.0	LibreFace	POSTER++	DDAMFN	EmotiEffLib	Qwen 2.5-VL
Curious	Neutral (0.44)	Neutral (0.51)	Neutral (0.73)	Neutral (0.73)	Neutral (0.70)	Neutral (0.96)
Confused	Neutral (0.33)	Neutral (0.43)	Neutral (0.54)	Neutral (0.54)	Neutral (0.44)	Neutral (0.93)
Disengaged	Neutral (0.35)	Neutral (0.59)	Neutral (0.71)	Neutral (0.71)	Neutral (0.48)	Neutral (0.81)
Surprised	Neutral (0.31)	Neutral (0.53)	Neutral (0.38)	Happy (0.50)	Neutral (0.41)	Neutral (0.95)
Conflicted	Neutral (0.39)	Happy (0.52)	Neutral (0.43)	Neutral (0.43)	Neutral (0.46)	Neutral (0.88)
Optimistic	Neutral (0.37)	Sad (0.46)	Neutral (0.60)	Neutral (0.60)	Neutral (0.55)	Neutral (0.97)
Frustrated	Neutral (0.38)	Neutral (0.44)	Neutral (0.56)	Neutral (0.56)	Neutral (0.47)	Neutral (0.84)

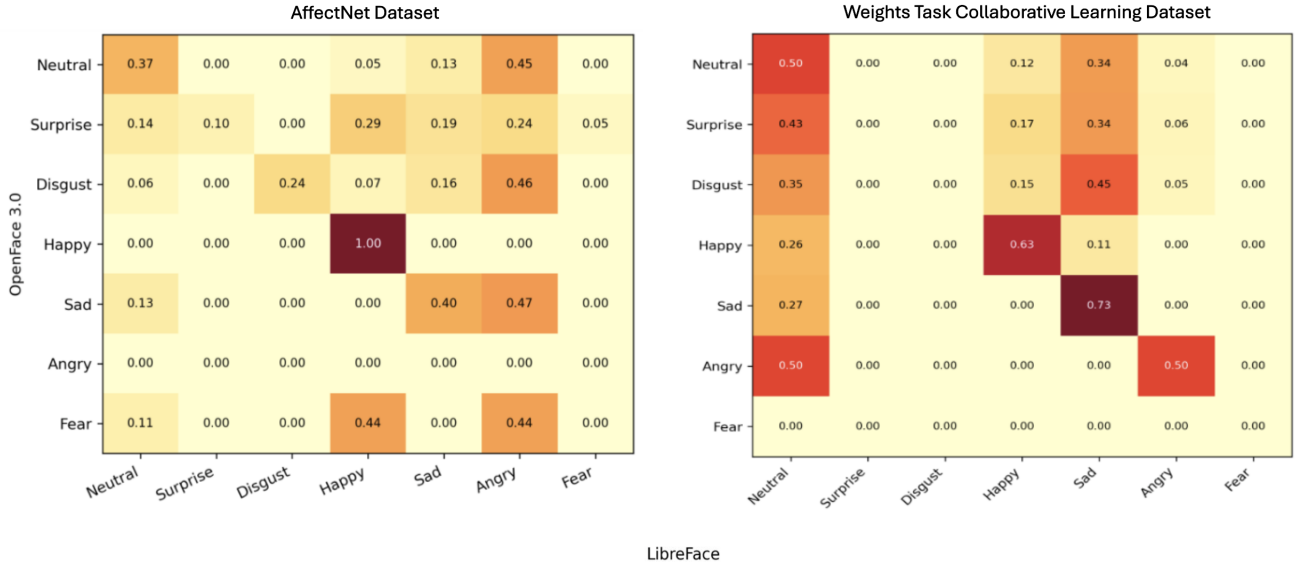

Figure 6. Cross-model confusion matrices between OpenFace 3.0 and LibreFace predictions on AffectNet (left) and EECPS-WT (right).

AffectNet and EECPS-WT dataset. Figure 6 presents the two confusion matrices.

For AffectNet, agreement between models varies substantially across emotion categories. The strongest alignment occurs for *Happy*, where predictions coincide almost perfectly between the two models. Moderate agreement is also observed for *Sad*. However, other categories exhibit substantial disagreement. For example, images predicted as *Neutral* by OpenFace 3.0 are frequently labeled as *Angry* or *Sad* by LibreFace, and *Disgust* predictions often correspond to *Angry*. Despite these inconsistencies, the AffectNet matrix exhibits a relatively diverse distribution of predicted labels, including *Fear*, *Disgust*, and *Angry*, reflecting the presence of more prototypical facial expressions within the benchmark dataset.

In contrast to AffectNet, predictions in EECPS-WT are heavily concentrated in a small subset of categories, particularly *Neutral*, *Sad*, and *Happy*. Other categories such as *Fear*, *Disgust*, and *Angry* appear rarely or not at all. Additionally, cross-model agreement is limited, with predictions frequently distributed across multiple categories rather than aligning along the diagonal.

5. Discussion

Our findings highlight three interrelated challenges in deploying pretrained FER systems in collaborative educational settings.

5.1. Basic Emotions and Epistemic States

Across models, epistemic states are predominantly mapped to Neutral. This collapse suggests that models trained on web-scale benchmarks appear calibrated toward canonical and often exaggerated facial prototypes and default to Neutral when subtle or ambiguous expressions are encountered. In collaborative problem-solving (CPS) settings, where affect may be regulated, masked, or embedded within task-focused expressions, such calibration leads to a collapse of predictions into a small subset of categories. Such emo-

tions may instead manifest as mild brow furrows, brief gaze shifts, or micro-level muscular changes that fall below the decision thresholds learned from prototypical training data.

5.2. Valence–Arousal and Epistemic States

Dimensional valence–arousal modeling provides a richer representational space, yet epistemic states remain highly overlapping in V–A space. While the dimensional model behaves consistently with respect to canonical basic emotions, it does not yield separable structure for collaboratively reported epistemic labels. This contrast indicates that the dimensional model is internally coherent with respect to the basic emotion taxonomy on which it was trained. However, that structure does not transfer to epistemic states observed in CPS. Dimensional representations therefore inherit the same domain alignment limitations as categorical models.

5.3. Cross-Model Reliability

The cross-model agreement analysis reveals several important insights about the reliability of current FER systems when applied to educational contexts. First, even when evaluated on AffectNet images—data drawn from a large-scale facial expression benchmark—agreement between OpenFace 3.0 and LibreFace is inconsistent across several emotion categories. While both models reliably identify highly prototypical expressions such as *Happy*, predictions diverge for other emotions including *Disgust*, *Angry*, and *Neutral*. This suggests that different model architectures may learn distinct feature representations and decision boundaries despite sharing a common emotion taxonomy. Second, cross-model disagreement becomes even more pronounced when models are applied to CPS. In EECPS-WT, predictions from both models collapse into a small subset of categories—primarily *Neutral*, *Sad*, and *Happy*. Emotions such as *Fear*, *Disgust*, and *Angry*, which appear in AffectNet predictions, are rarely produced in the collaborative setting. This pattern suggests that facial expressions observed during CPS may not strongly resemble the canonical facial configurations present in benchmark datasets.

These results are particularly noteworthy given that both OpenFace 3.0 and LibreFace were trained using datasets derived in part from AffectNet. However, the two systems differ substantially in their training pipelines and additional training data. LibreFace also incorporates pretraining on the FFHQ [37] and EmotioNet [4] datasets and fine-tuning on DISFA [38], while OpenFace 3.0 employs a multi-task learning architecture trained to jointly predict facial landmarks, action units, and emotion logits. Differences in training objectives and auxiliary datasets may therefore influence how each model interprets facial configurations in naturalistic environments.

For educational applications, these findings raise important concerns about measurement reliability. If two widely used FER systems produce different predictions when applied to the same student behavior, it becomes difficult to interpret model outputs as stable indicators of learner affect. In CPS settings where emotional signals may be subtle, regulated, or embedded within task-focused activity, reliance on facial expressions alone may therefore provide an incomplete or inconsistent representation of learners' internal states. This highlights the need for multimodal approaches and domain-specific evaluation protocols when deploying affective computing systems in educational research.

6. Limitations

This study has several limitations. First, our analysis is based on a single collaborative problem-solving (CPS) dataset collected within one institutional context with a relatively small participant pool. Further validation across diverse classrooms, age groups, and cultural contexts is necessary to assess the generalizability of our findings. Second, we evaluate a limited set of FER models that primarily operate on static images. Future work should extend this analysis to multimodal and temporally-aware architectures, including video-based models and vision–language models fine-tuned for affect recognition [39]. Finally, our dataset relies on retrospective cued-recall annotations, which may be susceptible to recall bias. Future work should explore complementary annotation strategies, including hybrid approaches that integrate behavioral and physiological signals.

7. Conclusion

We evaluated web-trained FER systems in a collaborative problem-solving (CPS) setting annotated with epistemic states via retrospective cued-recall. Across multiple architectures, categorical models largely mapped epistemic states to *Neutral*, while valence–arousal representations failed to produce separable structure among labels. Low inter-model agreement further indicated instability under domain shift. Dimensional predictions remained internally consistent with canonical basic emotions, suggesting the misalignment arises not from model failure but from a mismatch between web-trained emotion taxonomies and collaboratively expressed epistemic affect. These results suggest that these modern FER systems trained on web-scale datasets may not directly transfer to CPS contexts, and their outputs should be interpreted cautiously when used in human-centered AI research.

References

[1] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal

computing in the wild," *IEEE transactions on affective computing*, vol. 10, no. 1, pp. 18–31, 2017. 1, 2, 3, 5

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10. 2

[3] J. Hu, L. Mathur, P. P. Liang, and L.-P. Morency, "Openface 3.0: A lightweight multitask system for comprehensive facial behavior analysis," in *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2025, pp. 1–11. 2, 4

[4] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5562–5570. 2, 3, 8

[5] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, and Y. Wang, "Poster++: A simpler and stronger facial expression recognition network," *Pattern Recognition*, vol. 157, p. 110951, 2025. 2, 4

[6] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal, "Edusense: Practical classroom sensing at scale," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–26, 2019. 2

[7] P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein, "Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction," *Educational Psychology Review*, vol. 33, no. 1, pp. 27–49, 2021. 3

[8] A. Joshi, D. Allessio, J. Magee, J. Whitehill, I. Arroyo, B. Woolf, S. Sclaroff, and M. Betke, "Affect-driven learning outcomes prediction in intelligent tutoring systems," in *2019 14th ieee international conference on automatic face & gesture recognition (fg 2019)*. IEEE, 2019, pp. 1–5.

[9] X. Tang, Y. Gong, Y. Xiao, J. Xiong, and L. Bao, "Facial expression recognition for probing students' emotional engagement in science learning," *Journal of Science Education and Technology*, vol. 34, no. 1, pp. 13–30, 2025. 2

[10] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International conference on neural information processing*. Springer, 2013, pp. 117–124. 2, 3

[11] A. P. Fard, M. M. Hosseini, T. D. Sweeny, and M. H. Mahoor, "Affectnet+: A database for enhancing facial expression recognition with soft-labels," *IEEE Transactions on Affective Computing*, 2025. 2, 4

[12] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980. 2, 3

[13] F. Zhou, S. Kong, C. C. Fowlkes, T. Chen, and B. Lei, "Fine-grained facial expression analysis using dimensional emotion model," *Neurocomputing*, vol. 392, pp. 38–49, 2020. 2

[14] S. D'Mello, R. W. Picard, and A. Graesser, "Toward an affect-sensitive autotutor," *IEEE Intelligent Systems*, vol. 22, no. 4, pp. 53–61, 2007. 2, 3

[15] R. S. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser, "Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments," *International Journal of Human-Computer Studies*, vol. 68, no. 4, pp. 223–241, 2010. 3

[16] E. B. Cloude, A. Munshi, J. A. Andres, J. Ocumpaugh, R. S. Baker, and G. Biswas, "Exploring confusion and frustration as non-linear dynamical systems," in *Proceedings of the 14th learning analytics and knowledge conference*, 2024, pp. 241–252. 2

[17] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010. 2

[18] D. M. Russell and M. Oren, "Retrospective cued recall: a method for accurately recalling previous user behaviors," in *2009 42nd Hawaii International Conference on System Sciences*. IEEE, 2009, pp. 1–9. 2, 3

[19] S. Anindho, V. Venkatesha, and N. Blanchard, "A methodological framework for capturing cognitive-affective states in collaborative learning," *arXiv preprint arXiv:2507.01166*, 2025. 2, 3

[20] S. Li, W. Deng, and J. 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*, 2017, pp. 2852–2861. 2, 3

[21] S. Zhang, Y. Zhang, Y. Zhang, Y. Wang, and Z. Song, "A dual-direction attention mixed feature network for facial expression recognition," *Electronics*, vol. 12, no. 17, p. 3595, 2023. 2, 4

[22] D. Chang, Y. Yin, Z. Li, M. Tran, and M. Soleymani, "Libreface: An open-source toolkit for deep facial expression analysis," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 8205–8215. 2, 4

[23] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995. 2

[24] M. K. Underwood, "Peer social status and children's understanding of the expression and control of positive and negative emotions," *Merrill-Palmer Quarterly (1982-)*, pp. 610–634, 1997. 2

[25] L. Isenbarger and M. Zembylas, "The emotional labour of caring in teaching," *Teaching and teacher education*, vol. 22, no. 1, pp. 120–134, 2006. 2

[26] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992. 3

[27] M. P. Cross, A. M. Acevedo, and J. F. Hunter, "A critique of automated approaches to code facial expressions: What do researchers need to know?" *Affective Science*, vol. 4, no. 3, pp. 500–505, 2023. 3

[28] A. Horvers, N. Tombeng, T. Bosse, A. W. Lazonder, and I. Molenaar, "Detecting emotions through electrodermal activity in learning contexts: A systematic review," *Sensors*, vol. 21, no. 23, p. 7869, 2021. 3

[29] J. W. Y. Chung, H. C. F. So, M. M. T. Choi, V. C. M. Yan, and T. K. S. Wong, "Artificial intelligence in education: Using heart rate variability (hrv) as a biomarker to assess emotions objectively," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100011, 2021. 3

[30] I. Khebour, R. Brutti, I. Dey, R. Dickler, K. Sikes, K. Lai, M. Bradford, B. Cates, P. Hansen, C. Jung *et al.*, "When text and speech are not enough: A multimodal dataset of collaboration in a situated task," *Journal of open humanities data*, vol. 10, 2024. 3

[31] S. Anindho, V. Venkatesha, M. Bradford, A. M. Cleary, and N. Blanchard, "An exploration of internal states in collaborative problem solving," in *International Conference on Human-Computer Interaction*. Springer, 2025, pp. 135–150. 4

[32] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212. 4

[33] A. Savchenko, "Facial expression recognition with adaptive frame rate based on multiple testing correction," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 30 119–30 129. [Online]. Available: https://proceedings.mlr.press/v202/savchenko23a.html 4, 5

[34] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024. 4

[35] S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2021, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9659697 5

[36] D. Kollias and S. Zafeiriou, "Aff-wild2: Extending the aff-wild database for affect recognition," *arXiv preprint arXiv:1811.07770*, 2018. 5

[37] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410. 8

[38] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013. 8

[39] A. Chaubey, X. Guan, and M. Soleymani, "Face-llava: Facial expression and attribute understanding through instruction tuning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 2648–2660. 8