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

Supplementary Materials

Sifatul Anindho Videep Venkatesha Nathaniel Blanchard
Colorado State University, USA
c837200008@colostate.edu

1. EECPS-WT dataset additional statistics

This section provides additional descriptive statistics for the EECPS-WT dataset to contextualize the annotation distribution and temporal characteristics of the reported cognitive-affective states. Table 1 summarizes the total number of reported labels across the dataset and separates them by reporting mechanism. Probe-caught reports are more frequent than self-caught reports, indicating that the probing mechanism contributes substantially to overall coverage of cognitive-affective state annotations. Table 2 reports descriptive statistics for the time difference between the consecutive reports, providing an estimate of the density of affective reports in the dataset.

Table 1. Statistics on the number of labels of cognitive-affective states reported.

	All Labels	Self-Caught	Probe-Caught
Total Count	359	129	230
Mean	13.30	4.78	8.52
SD	4.67	4.81	5.69

Table 2. Descriptive statistics of the time difference between the experience and report of cognitive-affective states.

Metric	Value
Average Time Difference (seconds)	40.43
Standard Deviation (seconds)	25.45
Standard Error (seconds)	1.34

Figure 1 shows the frequency distribution of the affective state labels across the full dataset. These frequencies provide additional context for the label imbalance discussed in the main paper. Figure 2 further breaks down these frequencies by reporting mechanism, allowing comparison between self-caught and probe-caught reports. Finally, Figure 3 visualizes when the affective states were reported over normalized task time. This figure provides a coarse view of how cognitive-affective states are distributed across the progression of the collaborative task and complements the descriptive statistics reported in Tables 1 and 2.

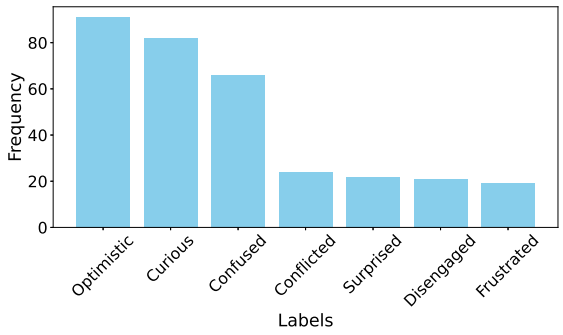

Figure 1. Frequency distribution of the cognitive-affective state labels overall.

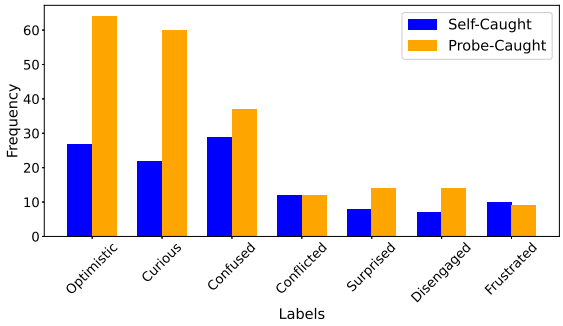

Figure 2. Frequency distribution of the cognitive-affective state labels across reporting mechanisms.

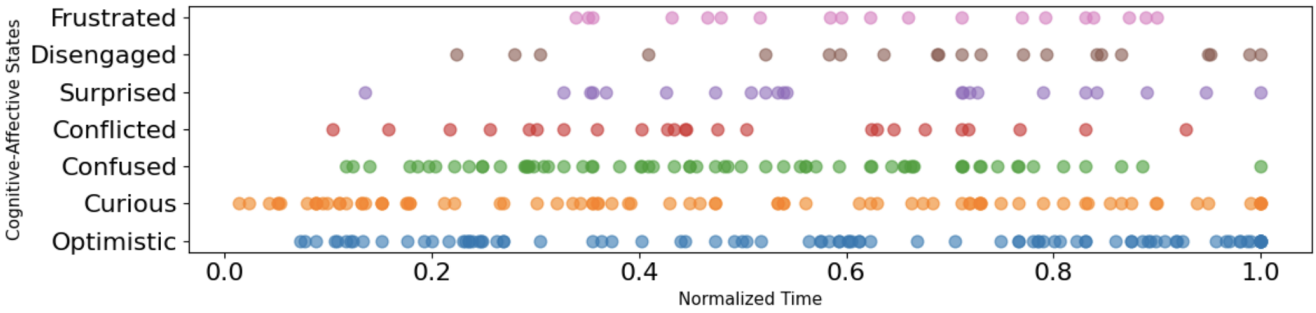

Figure 3. Distribution of the 7 most frequent cognitive-affective state reports over normalized task time. Each point represents a reported label positioned according to its normalized occurrence time within the task.

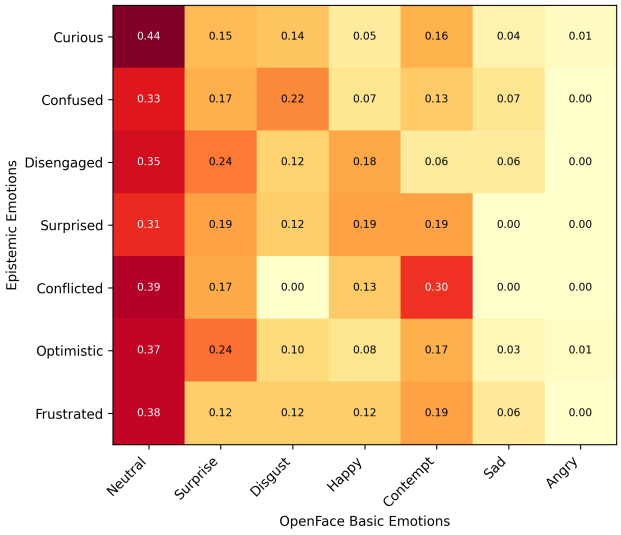

Figure 4. Row-normalized confusion matrix between epistemic states and OpenFace 3.0 predictions using a $\pm$5s temporal window. Predictions remain concentrated in a small subset of basic emotion categories, most notably Neutral.

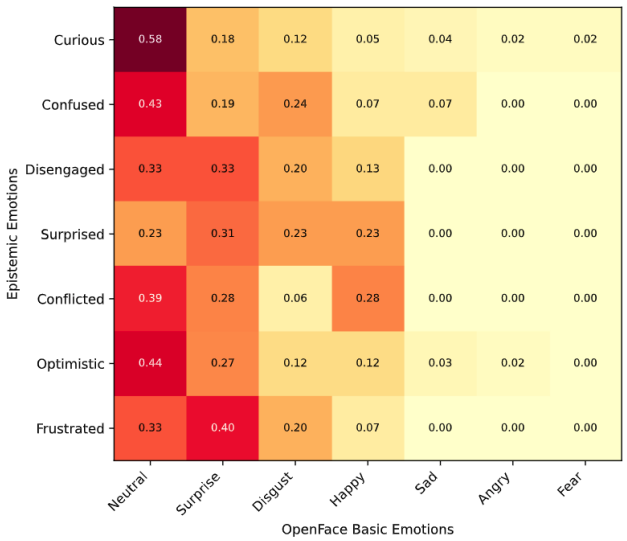

Figure 5. Row-normalized confusion matrix between epistemic states and OpenFace 3.0 predictions using a $\pm$3s temporal window. The overall pattern is similar to the $\pm$5s setting, indicating that the categorical collapse is not highly sensitive to moderate changes in temporal alignment.

2. Additional results

2.1. Temporal window robustness

To assess whether our findings depend strongly on the temporal sampling window used to associate face crops with retrospective reports, we repeated the analysis using three window sizes centered on the report timestamp: $\pm$5s, $\pm$3s, and $\pm$1s. We report row-normalized confusion matrices for OpenFace 3.0 as a representative categorical model (see Figures 4, 5, and 6 . Across all three conditions, the qualitative pattern remains stable: predictions remain concentrated in a small subset of basic emotion categories, with no emergence of a clear one-to-one mapping between epistemic states and categorical outputs. These results suggest that the main findings in the paper are not driven solely by the choice of temporal window.

2.2. Categorical predictions

As noted in the main paper, we report the row-normalized confusion matrix for OpenFace 3.0 in the main text as a representative categorical model. Here, we provide the corresponding confusion matrices for other evaluated categorical FER systems to show that the observed pattern is not specific to a single architecture (see Figures 7, 8 and 9. Across models, the same broad trend persists: epistemic states do not map cleanly onto basic emotion categories, and predictions are often concentrated in a small subset of outputs such as Neutral, Sad, or Happy.

2.3. Dimensional predictions

As described in the main paper, we visualize valence–arousal distributions both by epistemic labels and by pre-

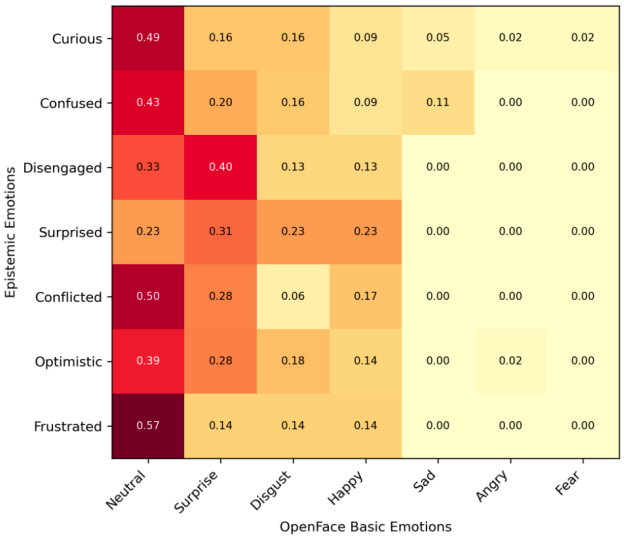

Figure 6. Row-normalized confusion matrix between epistemic states and OpenFace 3.0 predictions using a $\pm$1s temporal window. Even under the narrowest window, predictions remain concentrated and do not reveal stable, label-specific mappings from epistemic states to basic emotions.

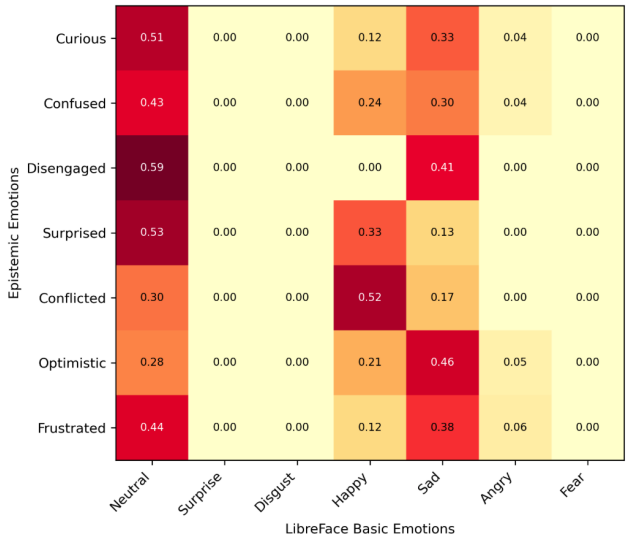

Figure 7. Row-normalized confusion matrix between epistemic states and LibreFace predictions. LibreFace exhibits the same general mismatch pattern, although the dominant predicted category differs for some epistemic states.

dicted basic emotion labels. This section provides additional visualizations and descriptive statistics to contextualize the dimensional analysis. Grouping by epistemic labels reveals substantial overlap and weak separation in valence–arousal space, whereas grouping by predicted basic emotions yields distributions that are more internally coherent with the canonical emotion taxonomy on which the model

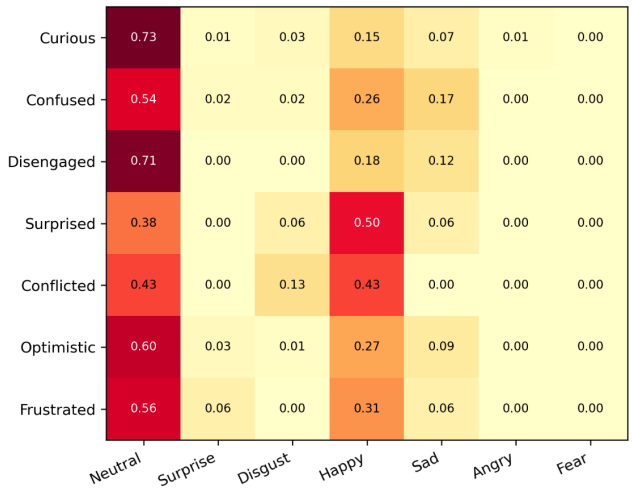

Figure 8. Row-normalized confusion matrix between epistemic states and POSTER++ predictions. Despite architectural differences, POSTER++ does not recover clear label-specific structure between epistemic and basic emotion categories.

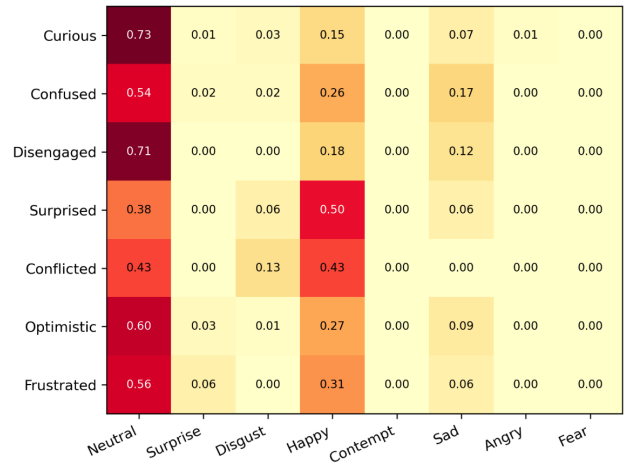

Figure 9. Row-normalized confusion matrix between epistemic states and DDAMFN predictions. Similar to the other categorical models, DDAMFN shows substantial overlap and limited discrimination among epistemic states.

was trained.

Consistent with the scatter plots in the main paper, the descriptive statistics in Table 3 and the corresponding box plot in Figure 10 show relatively small differences in means compared with the within-label variability, reinforcing the lack of strong separation among epistemic states. In contrast, Table 4 and the corresponding box plot in Figure 11 shows that the same dimensional model produces more interpretable regions when grouped by its own predicted categorical outputs, further suggesting that the learned representation is better aligned with basic emotion structure.

Table 3. Descriptive statistics of valence and arousal grouped by epistemic labels.

Label	n	Valence Mean	Valence SD	Arousal Mean	Arousal SD
Curious	82	-0.078676	0.245169	-0.042169	0.112119
Confused	66	-0.011643	0.284990	-0.048052	0.111944
Disengaged	21	-0.115182	0.245743	-0.025882	0.151067
Surprised	22	0.116545	0.329009	0.011059	0.097129
Conflicted	24	0.134300	0.292307	-0.041270	0.103854
Optimistic	91	-0.007026	0.279774	-0.072525	0.140265
Frustrated	19	0.166680	0.298367	0.017158	0.108912

Table 4. Descriptive statistics of valence and arousal grouped by predicted basic emotion labels.

Label	n	Valence Mean	Valence SD	Arousal Mean	Arousal SD
Neutral	189	-0.090604	0.162619	-0.063792	0.096550
Happy	79	0.410579	0.168118	0.008041	0.103393
Sad	55	-0.215394	0.180689	-0.140165	0.117975
Angry	22	-0.304407	0.108315	0.159531	0.111899
Surprise	1	0.169346	0.000000	0.400017	0.000000
Fear	1	0.092937	0.000000	0.158013	0.000000
Disgust	1	0.126887	0.000000	0.286436	0.000000
Contempt	4	0.188647	0.113637	0.018207	0.058968

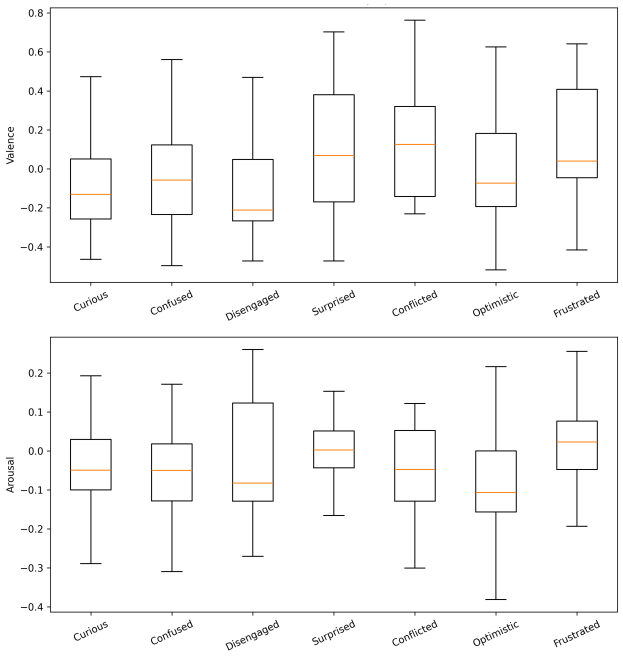

Figure 10. Distribution of valence (top) and arousal (bottom) grouped by epistemic labels. Large overlap across the seven epistemic states indicates that dimensional predictions do not form clearly separable structure for collaboratively reported epistemic affect.

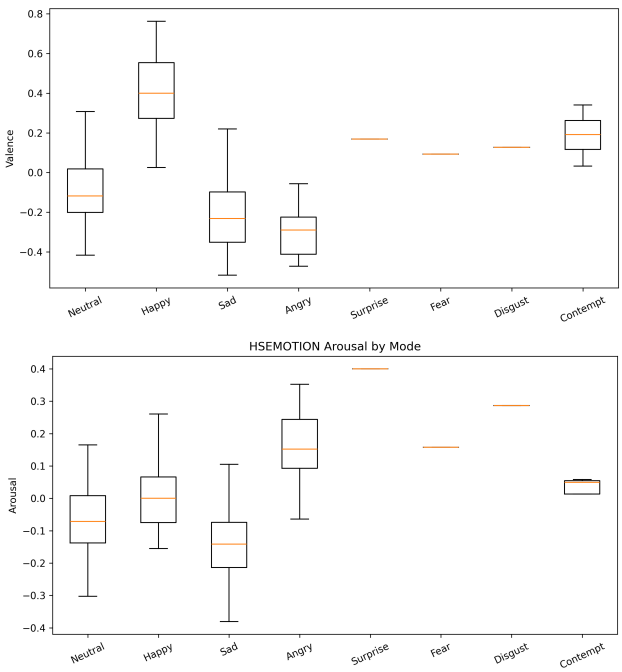

Figure 11. Distribution of valence (top) and arousal (bottom) grouped by predicted basic emotion labels. Compared with epistemic labels, dimensional values are more consistently organized with respect to canonical basic emotions, indicating stronger internal coherence than educational alignment.