1. Appendix

1.1. Extended analysis of perceived VSE experiments

In Tab. 1, we present the results for perceived VSE detection, while using Resnet152 features. In the main manuscript, we had only presented the results for VGG19 based backbone. We observe similar trends as the direct end-to-end method performs better than the the purely social signal based method (SS) in terms of average precision. However, the direct+SS method, closes this performance gap, while still enabling interpretation of detected perceived VSE in terms of underlying social signals. Please note that the AP in table 4 of the main manuscript and in Tab. 1 here, is weighted mean of Average Precision over classes.

Table 1. Performance on detecting intensity of interaction with Resnet152 features.

Dataset	Method	Accuracy (%)			AP (%)
		High	Low	No	
R3	Social Signal (SS)	**86.00**	**50.00**	80.25	80.58
	Direct	80.00	4.17	**92.59**	**86.52**
	Direct+SS	80.00	0.00	**92.59**	86.01
FID	Social Signal (SS)	80.28	**25.00**	66.67	69.28
	Direct	**94.29**	15.38	73.46	**82.97**
	Direct +SS	92.86	23.08	**83.33**	80.24

We present an example decision tree that explains the detected perceived VSE in terms of underlying social signals as shown in Fig. 1. Here the decision tree is constructed based on the the detected social signals to perform the classification of perceived VSE. Please note that the original decision tree for perceived VSE detection (whose results were presented in the main manuscript) is constructed with 50 leaf nodes and depth of 11. For easier visualization purpose, we have presented a simpler decision tree that is derived with 4 leaf nodes and depth of 2 in Fig. 1. For constructing the decision tree, criterion for split is set as entropy and the class weight is set to balanced. As shown in the figure, a cascade of decisions are made using social signals such as conversation, engagement and emotion to detect the classes of 'High', 'Low', and 'No' interaction. For example, high scores for conversation and high scores for emotion lead to a decision of 'High' perceived VSE.

Apart from class-wise accuracy and average precision, we also checked other common performance metrics such as precision, recall, F1-score etc. for perceived VSE detection. We have presented these results in Tab. 2. Here again, the measures are class-weighted according to the number of samples in each class. In this table, we can see that the direct+SS method comes very close to the performance of the direct method on the R3 dataset. On the FID dataset which has significantly higher distribution of interaction episodes, the direct+SS outperforms the direct method. Kindly note that we have not accounted for all possible social signals in our work. Including more social signals may enhance this performance further.

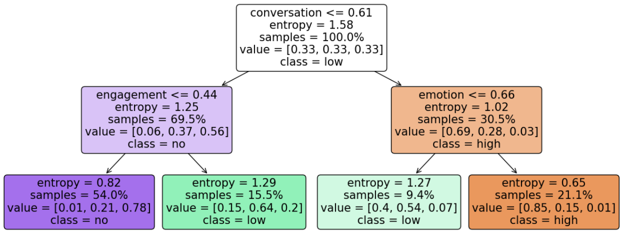


Figure 1. Decision tree for interaction intesity detection for R3 dataset with VGG19 backbone and LSTM model. The colors - orange, green and purple represent the high, low and no interaction classes, respectively.

Table 2. Extended evaluation on detecting intensity of interaction with more metrics. Feature used – VGG19.

Dataset	**Method**	**Prec. (%)**	**Recall (%)**	**F1 (%)**	**MAE**
R3	Social Signal (SS)	76.2	70.8	72.9	0.394
	Direct	**81.4**	**82.2**	**80.4**	**0.250**
	Direct+SS	79.4	81.4	79.3	0.267
FID	Social Signal (SS)	74.5	70.5	72.1	0.326
	Direct	78.3	80.0	79.1	0.221
	Direct +SS	**80.1**	**81.1**	**80.5**	**0.211**

1.2. Extended Analysis of AMR Experiments

The AMR prediction results presented in the main manuscript used the original train-test data split that were provided in [Xu et al., 2021]. This split included all images in the R3 dataset irrespective of whether they contained a face in it. Since our approach is interaction based, we investigated the performance of the models when they are trained and tested using only the subset of images that involved a human face. Specifically, using this subset we investigated whether social signals can be used to predict AMR effectively.

We developed the following set of models for this purpose. **M1:** social signal (SS) based model, which takes the output from the CNN-LSTM model for each social signal as input and uses an MLP or a Random Forest model to predict AMR. **M2:** Social Signals + perceived VSE (SS+Inter.) based model, which takes the output from the CNN-LSTM model for each social signal and the detected perceived VSE label and the score as input, and uses an MLP or Random Forest model to predict AMR.

However, since the subset of data samples with human faces is much smaller than the complete R3 dataset, we adopted a 5-fold cross-validation approach for training and testing. The results are shown in Table 3. Here, we found that using just social signal scores (without using any other image features), we are able to predict AMR better than chance (AUC $> 50\%$) resulting in a weighted F1-score of 51.8% for three AMR categories. This shows that the information available in the social signals are useful for predicting AMR. We also see that the use of perceived VSE in M2 models improves the performance further (F1 score of 53.6%).

Table 3. Prediction of AMR using only the detected Social Signals and perceived VSE

Input	Method	Precision(%)	Recall(%)	F1(%)	MAE
M1 : SS	RandomForest	49.0	57.9	51.8	0.511
	MLP	47.7	40.0	41.9	0.849
M2: SS + Inter.	RandomForest	**52.2**	**60.1**	**53.6**	**0.492**
	MLP	50.5	40.6	43.9	0.795

1.3. More details on Model Training

Our models are trained and evaluated using PyTorch Framework on Machine with Intel Xeon W-2123 CPU and NVIDIA GeForce GTX 1070 Ti GPU.

For Social Signal detection models and perceived VSE detection models, batch size are set as 16, exponential learning rate schedule with gamma 0.999 are set and models are trained for 200 epochs.

For AMR prediction, we used Resnet50 backbone for AMNet with batch size of 32 and learning rate 0.0001 and models are trained for 55 epochs.

1.4. Social Signal Detection Network

The complete social signal detection network is shown on the Figure 2. As mentioned, the perceived VSE detection network is derived from the social signal detection network by changing the sigmoidal function to softmax function to accommodate multiple outputs for the labels (high interaction, low interaction and no interaction).

For social signal detection, we compared our modular approach of having a separate network for each social signal versus having a single end-to-end neural network to detect all three social signals. These results are presented in Tab. 4. As expected, the combined approach performed poorly (in terms of F1 score) compared to the modular approach. This may be attributed to the fact that different social signals need different number of steps in the LSTM, due to their temporal characteristics. The combined approach also has the disadvantage of the need to retrain all over for different combinations of social signals. In contrast, our simple architecture shown in the main manuscript allows us to easily add or remove social signals that may or may not be available on a new dataset.

1.5. More details on Annotated Social Signals

In Tab. 5, we provide the details of agreement between the annotators for social signals on R3 dataset. For each social signal, the annotators have to decide the presence or absence of an individual social signal according to the definitions provided. This is a subjective human- intelligence task and there is scope for disagreement. However, we found that a majority of samples had total agreement between the annotators. Unanimous agreement among annotators was highest for conversation (953/1119) and the lowest for Face-to-face engagement (894/1119).

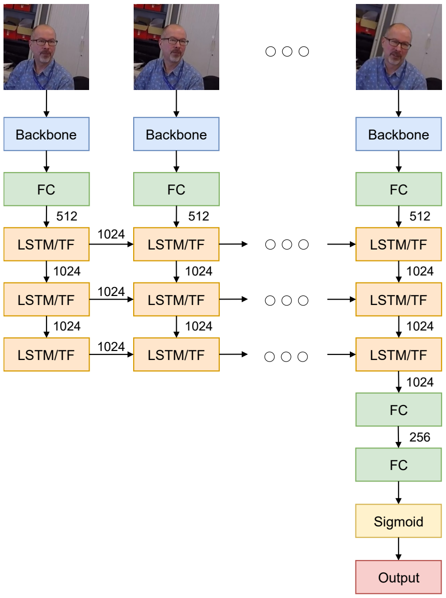


Figure 2. Overview of Social Signal Detection Network

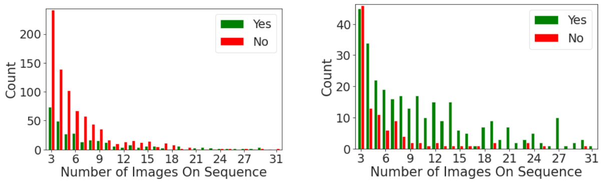
Table 4. Social signal detection using individual models versus a combined model. Feature – VGG19, Model – CNN-LSTM. The values mentioned in this table are percentages.

Social Signal		Elevated Emotion		Direct Conversation		Face-to-Face Engagement	
Method		Individual	Combined	Individual	Combined	Individual	Combined
R3	Recall (%)	74.14	**75.86**	82.35	**86.27**	77.97	**79.66**
	F1 (%)	**70.49**	67.18	**75.00**	72.13	**71.88**	70.15
	AP (%)	72.43	**74.09**	80.73	**81.93**	74.94	**78.46**
FID	Recall (%)	**93.15**	89.04	**90.28**	**90.28**	**95.06**	88.69
	F1 (%)	88.89	**89.04**	**89.66**	**89.66**	**96.25**	93.51
	AP (%)	**93.69**	93.09	**94.04**	92.86	**98.65**	98.21

Table 5. Agreement between annotators for social signals on R3 dataset.

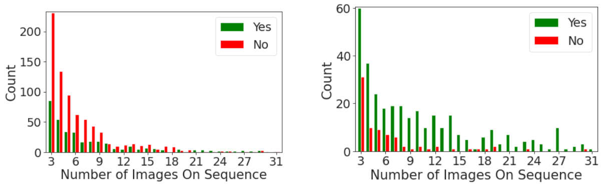
Social Signals	All agree as No	1/3 agree as Yes	2/3 agree as Yes	All agree as Yes
Emotion	706	70	96	247
Conversation	757	58	55	249
Engagement	630	139	86	264
Final decision	**No**		**Yes**	

In the main manuscript, we had shown the histogram of presence/absence of elevated emotion for different sequence-lengths and contrasted the difference in distribution for the FID and R3 datasets. Here, we provide a similar breakdown for the presence/absence of sustained conversation (Fig. 3) and face-to-face engagement (Fig. 4) across various interaction sequence lengths in both R3 and FID datasets. We observe that enacted FID dataset contains more social signals at all image sequence lengths compared to in-the-wild R3 dataset.



(a) (b)

Figure 3. Histogram of Sustained Conversation labels in (a) R3 dataset and (b) FID dataset



(a) (b)

Figure 4. Histogram of Face-to-Face Engagement labels in (a) R3 dataset and (b) FID dataset