How you feelin'? Learning Emotions and Mental States in Movie Scenes
SUPPLEMENTARY MATERIAL

Dhruv Srivastava Aditya Kumar Singh Makarand Tapaswi
CVIT, IIIT Hyderabad, India
https://katha-ai.github.io/projects/emotx

We present supplementary material to support "How you feelin'? Learning Emotions and Mental States in Movie Scenes" work. In Sec. A we refer to Fig. 1 of the main paper (teaser) and share the hidden contexts in each scene reflecting upon the importance of individual modalities to capture the emotions in real-world environments. Sec. B present some statistics around emotions extracted from the MovieGraphs dataset. In Sec. C we share the character detection, tracking, and clustering pipeline used to extend the tracks provided in the MovieGraphs dataset. In Sec. D we visualize the class AP scores for top-10 and 25 emotions from MovieGraphs along with Emotic mapped emotions. Since there were several feature combinations in our work, an extended feature ablation is presented in Sec. E. Finally, Sec. F shares details of the modifications made to adapt EmotionNet [17], CAER [10], M2Fnet [4], and AttendAffectNet [15] for comparison with EmoTx. We end with another qualitative example showing the attention scores similar to Fig. 6 of the main paper in Sec. G.

A. The Stories behind Emotions in Fig. 1

We discuss some additional details from Fig. 1 of the main paper. Prior to this, note that the emotions are grouped into three tuples, each corresponding to the frame depicted in the example - however, this was for illustrative purposes and making it easy to match emotions to the frames. We do not explicitly generate frame-level predictions.

Scene A is taken from the movie "Sleepless in Seattle, 1993", scene number 087, where Suzy is narrating an incident from a classical movie "An affair to remember". While narrating, she gets sentimental and starts crying. The other characters, Sam and Greg listen curiously but feel neutral and mock her by faking a cry and narrating the scene from some war movie. This makes Suzy laugh, and she asks the duo to stop before the scene ends. The reflected emotions and mental states include *upset*, *calm*, *confused*, *excited*, *sad*, and *happy*. Observing the situation, it is evident that a single emotion label does not suffice and both the visual and dialog context taken over a longer duration is important to predict emotions with mental states.

Start	End	Speaker	Utterance
00:00	00:04	Sergeant Drill	Gump! What's your sole purpose in this Army?
00:04	00:06	Forrest Gump	To do whatever you tell me, Drill Sergeant!
00:06	00:10	Sergeant Drill	God damn it, Gump! You're a goddamn genius!
00:10	00:12	Sergeant Drill	That's the most outstanding answer I've ever heard.
00:12	00:15	Sergeant Drill	You must have a goddamn IQ of 160!
00:15	00:18	Sergeant Drill	You are goddamn gifted, Private Gump!
00:19	00:21	Sergeant Drill	Listen up, people!
00:21	00:25	Forrest Gump	Now, for some reason, I fit in the Army like one of them round pegs.
00:25	00:27	Forrest Gump	It's not really hard.
00:27	00:30	Forrest Gump	You just make your bed real neat and remember to stand up straight,
00:31	00:34	Forrest Gump	and always answer every question with, "Yes, Drill Sergeant!"
00:35	00:36	Sergeant Drill	Is that clear?
00:36	00:38	Everyone	Yes, Drill Sergeant!

Table 1. Subtitles from Scene 045 from movie Forrest Gump, 1993, corresponding to Scene B from Fig. 1. Note that the speaker names are added for improving the clarity and understanding, our model does not have access to them.

Scene B is taken from the movie "Forrest Gump, 1994", scene number 045. Forrest has joined the army and it is his introductory day. Sergeant Drill asks Forrest about his role in the army to which Forrest replies "To do whatever you tell me Sergeant Drill" which impresses him a lot. Then Sergeant Drill praises him by saying it is the best response he has ever heard! The original subtitles of this clip are shared in Table 1. We hope to show that the dialog modality is crucial in understanding the real emotions since visually

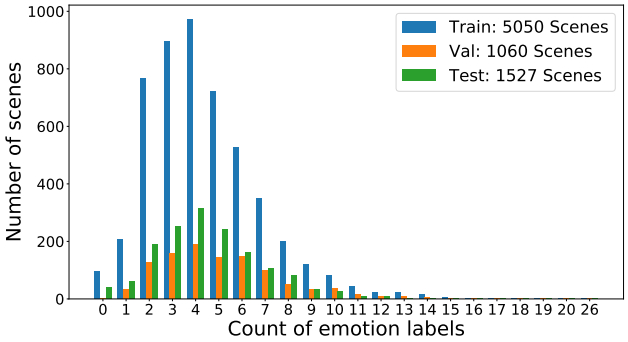

Figure 1. Number of scenes with a specific number of emotion labels in the train, val, and test splits.

it appears that both the characters are angry and screaming at each other but in reality Forrest is *determined*, *honest*, and *serious*, while the Sergeant is *excited*.

Scene C is taken from movie "Slumdog Millionaire, 2008", scene number 076. The scene represents a Television Show "Who wants to be a millionaire?" where Jamal is being asked some question. He has given the response and is waiting for the confirmation from the anchor. The frames used in the figure reflect the moment when the anchor excitedly reveals that the answer given by Jamal is correct. However, by only looking at the faces, it appears as if Jamal is tense and he anchor is scolding him, whereas in reality, everyone is clapping and cheering for him. We show that looking at the visual frame is necessary to correctly predict the wider perspective of emotions, here corresponding to the transition from *nervous* and *curious* to *surprised* and *amused* for Jamal, and *excited* for the anchor.

B. MovieGraphs-Emotions: Dataset Features

The MovieGraphs dataset [16] contains graph-based annotations for each scene within a movie. The nodes of these graphs include characters and their details such as relationships, interactions, emotions, and other physical attributes, along with movie scene-level labels such as the overarching situation, place (scene), and a few sentence natural language description. There are a total of 51 movies divided into 7637 clips with associated graphs. The MovieGraphs dataset is provided with train, validation and test splits which contain 33/7/11 movies with 5050/1060/1527 clip graphs respectively. These clips have an average duration of 41.7s at 23.976 fps (frames per second). For each clip, we focus on characters and their emotion attributes. As the dataset consists of free-text annotations, this amounts to massive 509 unique emotion labels in the dataset, which however, can be mapped to a smaller set.

Label distributions. We analyze the dataset from various perspectives and highlight some statistics.

Fig. 1 shows the number of scenes that have a certain

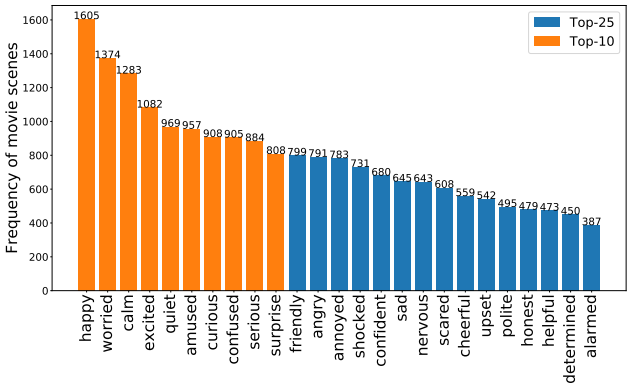

Figure 2. Number of movie scenes containing top-10 and 25 emotions. Note, the top-25 label set includes top-10.

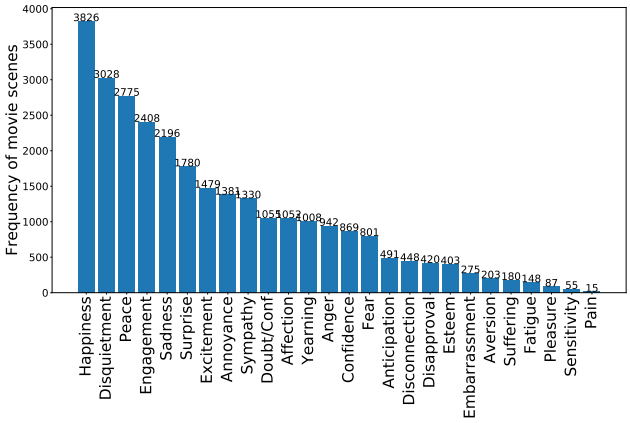

Figure 3. Number of movie scenes depicting each of the 26 Emotic mapped emotions.

number of emotions. We observe that most scenes have 2-7 emotions, and the train, val, and test distributions are relatively similar. The absolute counts are expected to be lower due to smaller val/test sizes.

Fig. 2 presents the number of instances for top-10 (orange) and top-25 (orange + blue) label sets. We see a classic long-tail effect, however, by selecting the top-25, we ensure that there are sufficient instances for all labels to learn a decent representation.

Fig. 3 shows the same distribution after mapping 181 emotions from MovieGraphs to the 26 emotion labels of the Emotic dataset [9]. We used a similar mapping as shared by [12] and show the details in Table 2. Recall that we report results on this label set in our SoTA experiments in Section 4.4 of the main paper.

We assign the character index 1, 2, ... to the most frequent, second most frequent character, and so on. The plot in Fig. 4 shows the average number of scenes in which a character appears, or rather, has an annotated emotion from the MovieGraphs dataset. This provides interesting avenues for future research, to track emotions across the entire movie.

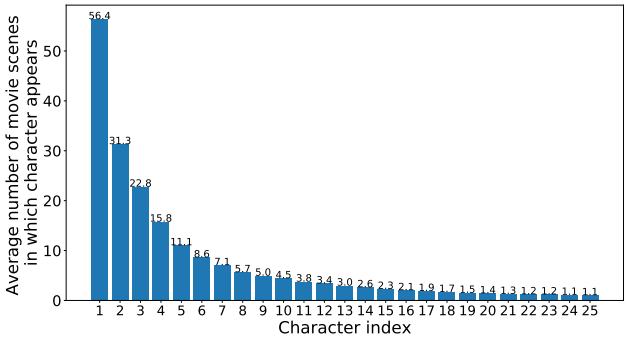

Figure 4. Average number of scenes in which characters appear. Character 1 corresponds to the most frequently occurring character in the movie, character 2 to the second most frequent, and so on. For our work, a character is considered as present if there is at least one emotion annotated in the scene.

Co-occurrence in the top-25 labels. Similar to Fig. 3 of the main paper, we show the row-normalized co-occurrence matrices for the top-25 labels in Fig. 5. From a cursory look, we observe that the movie scene labels (left) are denser than the per-character co-occurrence (right) - this is expected as the movie scene level labels contain a combination of multiple characters.

We present a few notable differences between the scene-level and character-level co-occurrences. Tuples here correspond to label1 selecting a row, and label2 selecting a column. (*friendly*, *polite*) seems to be applicable to different characters in a scene, but not for one. (*honest*, *curious*) shows similar characteristics. Interestingly while a single character is (*alarmed*, *worried*), in a scene, (*alarmed*, *serious*) also gets fairly high scores.

C. Character Processing Pipeline

The face tracks provided by the MovieGraphs dataset [16] occasionally miss the characters due to the quality of the face detection. By watching some clips, we observed that many face tracks were broken within a clip due to missed detections and multiple track IDs were provided for the same character within a single shot. In addition, some shots had 0 detections, but could be useful to provide a wider perspective on the emotions of that character and scene. Therefore, we extend the face tracks from MovieGraph dataset by first extending the sparse ground-truth tracks within a shot and then over multiple shots within a scene through clustering.

In summary, we first recompute face detections and tracks for the movie scenes. A subset of the new face tracks are assigned a name based on overlap with the original tracks present in the dataset. Then, we cluster all detections in a clip using hierarchical clustering and assign names to remaining unnamed tracks based on the clustering. Fig. 6 shows an example where original tracks did not have a single detection (due to the dark scene) for a scene in the "Forrest Gump, 1994" movie.

Emotic Label	MovieGraphs emotions
Affection	loving, friendly
Anger	angry, resentful, outraged, vengeful
Annoyance	annoyed, annoying, frustrated, irritated, agitated, bitter, insensitive, exasperated, displeased
Anticipation	optimistic, hopeful, imaginative, eager
Aversion	disgusted, horrified, hateful
Confidence	confident, proud, stubborn, defiant, independent, convincing
Disapproval	disapproving, hostile, unfriendly, mean, disrespectful, mocking, condescending, cunning, manipulative, nasty, deceitful, conceited, sleazy, greedy, rebellious, petty
Disconnection	indifferent, bored, distracted, distant, uninterested, self-centered, lonely, cynical, restrained, unimpressed, dismissive
Disquietment	worried, nervous, tense, anxious, afraid, alarmed, suspicious, uncomfortable, hesitant, reluctant, insecure, stressed, unsatisfied, solemn, submissive
Doubt/Conf	confused, skeptical, indecisive
Embarrassment	embarrassed, ashamed, humiliated
Engagement	curious, serious, intrigued, persistent, interested, attentive, fascinated
Esteem	respectful, grateful
Excitement	excited, enthusiastic, energetic, playful, impatient, panicky, impulsive, hasty
Fatigue	tired, sleepy, dizzy
Fear	scared, fearful, timid, terrified
Happiness	cheerful, delighted, happy, amused, laughing, thrilled, smiling, pleased, overwhelmed, ecstatic, exuberant
Pain	hurt
Peace	content, relieved, relaxed, calm, quiet, satisfied, reserved, carefree
Pleasure	funny, attracted, aroused, hedonistic, pleasant, flattered, entertaining, mesmerized
Sadness	sad, melancholy, upset, disappointed, discouraged, grumpy, crying, regretful, grief-stricken, depressed, heartbroken, remorseful, hopeless, pensive, miserable
Sensitivity	apologetic, nostalgic
Suffering	offended, insulted, ignorant, disturbed, abusive, offensive
Surprise	surprise, surprised, shocked, amazed, startled, astonished, speechless, disbelieving, incredulous
Sympathy	kind, compassionate, supportive, sympathetic, encouraging, thoughtful, understanding, generous, concerned, dependable, caring, forgiving, reassuring, gentle
Yearning	jealous, determined, aggressive, desperate, focused, dedicated, diligent

Table 2. Mapping MovieGraphs emotions to Emotic labels, adapted from Affect2MM [12].

Face and person detection and tracking. New face and person detections are extracted from every movie scene of the MovieGraphs dataset. We adopt MTCNN (Multi-Task Cascaded Convolutional Neural Networks) [18] for face detection and Cascade-RCNN pretrained on cast annotations of MovieNet [7] for person detection. Since the original tracks are only for faces, we first compute person boxes using the person detector and obtain face detections within the person box in order to define a mapping between face and

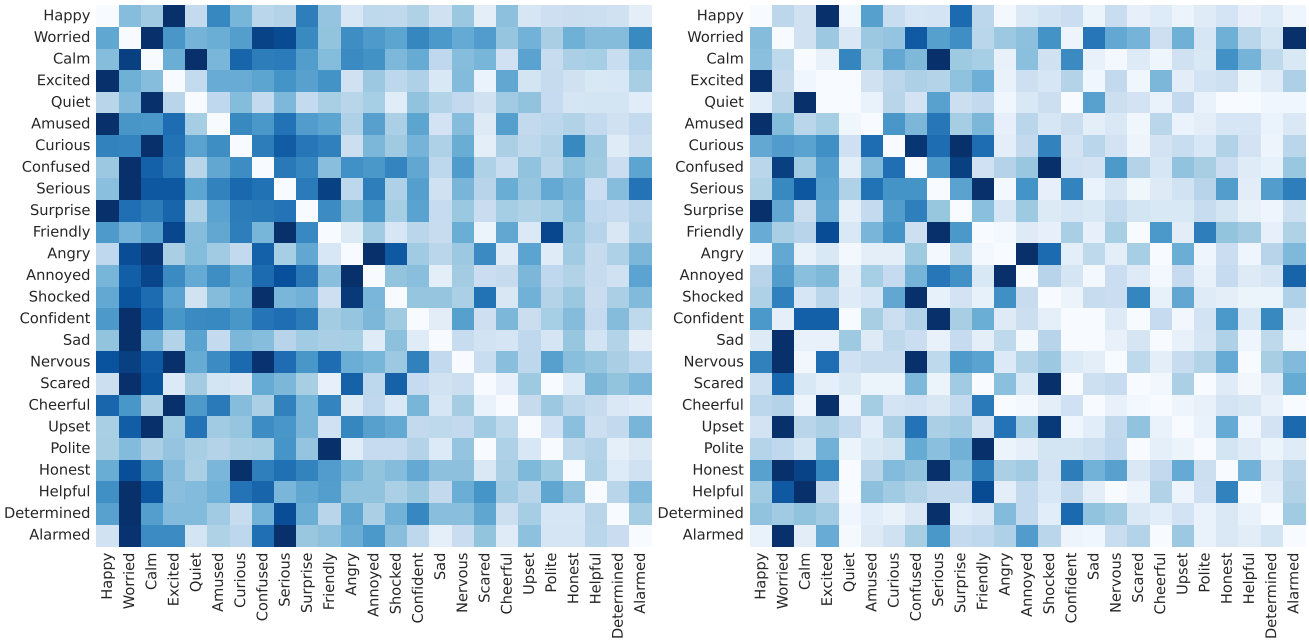

Figure 5. Normalized label co-occurrence matrices for the top-25 emotions associated with a *movie scene* (left) and *character-level emotions* (right).

Figure 6. Example face detections. The original face tracks do not work for dark scenes or profile faces, while our new detections and tracks are able to find them. Scene-036 from Forrest Gump, 1994.

person detections. If multiple faces are found within a person bounding box, the face with higher detection probability is selected. The resulting bounding boxes are tracked using the Kalman-filter based SORT (Simple Online and Realtime Tracking) algorithm [1]. Due to the mapping established between the face and person detections, the same track ID is shared between face and person tracks. For the rest of the discussion, we focus on face tracks.

Extending names from original to new face tracks. Since some of the newly generated tracks coincide with the original tracks from MovieGraphs, such tracks are assigned a name based on their IoU overlap score. In particular, for every detection in the original tracks, a corresponding new detection is mapped if the IoU score between the two is greater than a threshold (0.7 in our case). Thus, names from the original detections (or track), are mapped to the new track, and a majority vote of these names is used to decide the final name for a new track.

Face clustering and naming other tracks. Not all tracks are assigned a name through the above method due to missed detections in the original tracks. Thus, we perform clustering to increase the coverage. First, we extract good identity features from an InceptionResNetV1 [14] pretrained on the VGGFace2 [2] dataset. For clustering we use the C1C [8] algorithm which also uses track information for establishing must and cannot links between the face features. Individual face detections (features) are processed and clustered using C1C resulting in multiple partitions with varying number of clusters. We calculate the Silhouette score [13] for every partition and the one with highest score is selected as the representative partition. Now, based on the named tracks generated using the paragraph above, every cluster is assigned a probability corresponding to distinct names (via named detections) within the cluster. For clusters which do not have any named detection, equal probability is given to every name present in the scene. The cluster name-probabilities corresponding to the detections of unnamed tracks are extracted and the average of these soft scores is used to reflect the names for the newly discovered tracks. This way, we assign a name probability to new tracks and threshold it with 0.7 to select the final name for such new tracks.

D. Analyzing AP scores

Similar to Fig. 5 of the main paper, we present per-emotion scores for the top-10 emotions in the dataset in

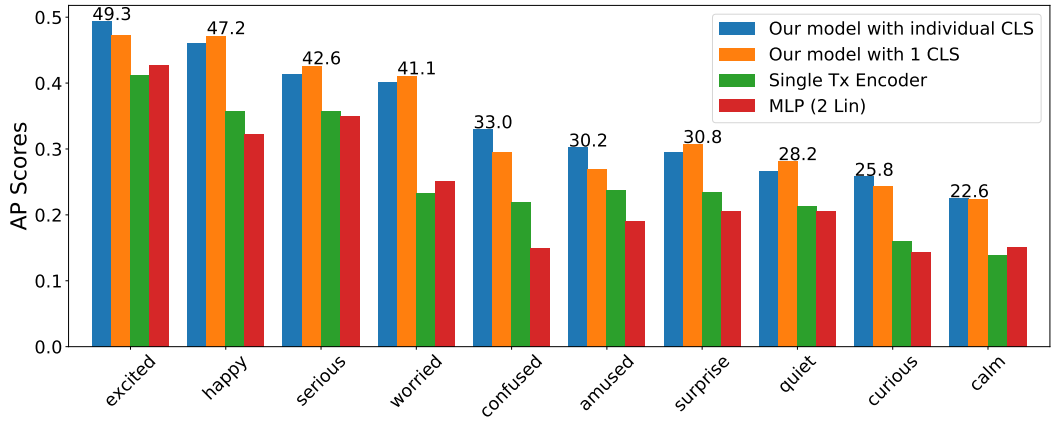

Figure 7. AP scores for the top-10 emotions label set sorted from high to low AP score for our model with individual CLS tokens.

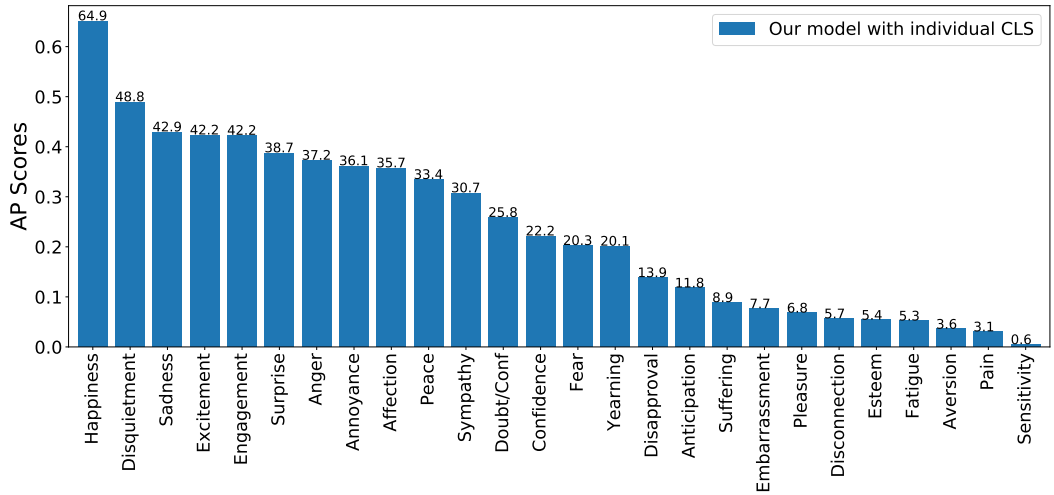

Figure 8. AP scores on the 26 grouped labels of the Emotic label set.

Fig. 7. We observe that our model with the individual classifier (CLS) tokens outperforms other approaches in 5 of 10 emotions.

In Fig. 8, we show the AP for each group of Emotic labels. We observe that challenging labels such as *pain*, *sensitivity*, perform much worse than others such as *happiness*, *sadness*, *anger*, *etc*.

E. Feature Ablation

We expand upon the feature ablation in Table 3 of the main paper to show the effect of additional feature combinations in Table 3. All the trends are similar, fine-tuning RoBERTa helps consistently, ResNet50 trained on FER appears to be a good representation for characters, and the MViT trained on Kinetics400 provides better results for both the label sets, while ResNet50 trained on Places365 is a close second.

F. Adapting SoTA Methods for our Task

The MovieGraphs dataset has not been used directly to predict emotions at a scene or character level. Related to using labels from MovieGraphs, Affect2MM [12] extracts scene-level emotion timelines for the entire movie, but relies on one emotion per scene. This is quite different from our vision of a multi-label setting where the scene and each character can present multiple emotions.

For a fair comparison to previous work, we chose models that have attained SoTA in image, video and multimodal emotion recognition. We share details on how these methods are adapted to make them suitable for our task.

EmotionNet [17] is a recent SoTA for emotion recognition from web images. It uses a joint embedding training approach which uses emotional keywords associated with a given image and aligns its learned text embedding (pretrained on massive text data) with image embedding extracted from a standard feature backbone (ResNet50). To adapt EmotionNet for our task, we used word2vec [11] for

	Video			Character			Dialog		Metrics (mAP)			
									Top-10		Top-25	
	MViT K400	R50 P365	R152 INet	R50 FER	VGG-M FER	IRv1 VGG-F	RB FT	RB PT	Scene	Char	Scene	Char
1	-	✓	-	-	-	✓	-	✓	25.07±0.12	15.48±0.15	16.41±0.24	8.31±0.17
2	-	-	✓	-	-	✓	-	✓	25.85±0.24	15.63±0.21	16.45±0.09	8.31±0.09
3	-	-	✓	-	✓	-	-	✓	29.20±0.22	19.88±0.27	18.93±0.38	10.16±0.17
4	✓	-	-	-	-	✓	-	✓	29.27±0.08	18.07±0.22	18.35±0.09	0.09±0.08
5	-	✓	-	-	✓	-	-	✓	29.30±0.21	19.73±0.17	19.05±0.19	10.31±0.00
6	✓	-	-	-	✓	-	-	✓	29.34±0.08	20.50±0.04	19.07±0.19	10.34±0.17
7	-	✓	-	-	-	✓	✓	-	29.34±0.17	19.49±0.03	20.73±0.08	10.75±0.02
8	-	-	✓	-	-	✓	✓	-	29.47±0.14	19.29±0.10	20.74±0.11	10.79±0.07
9	-	✓	-	✓	-	-	-	✓	29.69±0.38	20.25±0.14	20.16±0.29	11.06±0.12
10	-	-	✓	✓	-	-	-	✓	30.19±0.38	20.27±0.26	19.83±0.07	11.06±0.16
11	✓	-	-	✓	-	-	-	✓	31.39±0.34	21.18±0.18	20.88±0.28	11.46±0.08
12	✓	-	-	-	✓	-	✓	-	31.50±0.36	21.60±0.09	21.49±0.30	11.64±0.20
13	-	-	✓	-	✓	-	✓	-	31.96±0.20	21.81±0.37	21.28±0.25	11.58±0.26
14	✓	-	-	-	-	✓	✓	-	32.23±0.07	21.45±0.07	22.10±0.11	11.63±0.06
15	-	✓	-	-	✓	-	✓	-	32.42±0.26	22.32±0.27	21.45±0.17	11.62±0.05
16	-	-	✓	✓	-	-	✓	-	33.44±0.33	22.89±0.24	22.75±0.18	12.52±0.12
17	-	✓	-	✓	-	-	✓	-	33.46±0.21	22.98±0.16	22.69±0.22	12.48±0.20
18	✓	-	-	✓	-	-	✓	-	**34.22**±0.18	**24.35**±0.23	**23.86**±0.10	**13.36**±0.11

Table 3. Extended feature ablations. The different feature backbones are (MViT, K400): MViT pretrained on Kinetics400, (R50, P365): ResNet50 on Places365, (R152, INet): ResNet152 on ImageNet, (R50, FER): ResNet50 on Facial Expression Recognition (FER), (VGG-M, FER): VGG-M on FER, (IRv1, VGG-F): InceptionResNet-v1 trained on VGG-Face dataset, (RB, FT): pretrained RoBERTa finetuned for emotion recognition and (RB, PT): pretrained RoBERTa. Best numbers in bold, close second in italics.

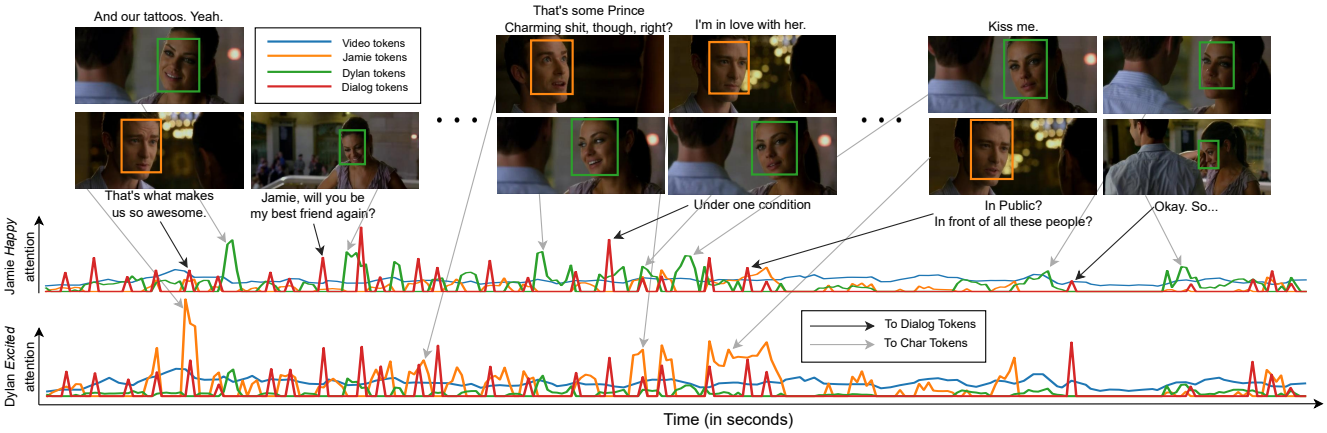

Figure 9. A scene from the movie *Friends with Benefits* with self-attention scores for multiple modalities for two character-level predictions: *Jamie is happy* and *Dylan is excited*. From the figure we can infer that the *happy* classifier token attends to the *Jamie* character tokens with spikes observed when she smiles or laughs, while *Dylan's excited* classifier token attends primarily to the dialog utterance tokens. We can see this as very few face snaps indicate that *Dylan* is excited, in fact, *Dylan's* face is not even visible often. However, dialog utterances like *That's what makes us so awesome*, *Hey, I miss you*, and *Jamie, will you be my best friend again?* are extremely useful for the model to infer the emotions.

extracting text embeddings and ResNet50 for frames. Since we use a video as input, the frame features are max-pooled to generate a single representation. We use the proposed embedding loss and provide the emotion labels as the keywords for joint embedding training. This learned ResNet50 is finetuned for multilabel emotion recognition where the individual frame features are max-pooled before passing to the logits layer.

CAER (Context Aware Emotion Recognition) [10] is a deep Convolutional Network which consists of two stream encoding networks to separately extract the facial and con-

text features which are fused using an adaptive fusion network. Detections from our extended face tracks are used as inputs for the face encoding stream and the full video frame with masked faces was used as input to context encoding stream. Since CAER is designed to extract emotions from images we adapt it to videos by applying max-pooling over the fused features from both the streams to generate a single representation for a video. This adapted model is trained to predict multiple scene-level emotions.

M2FNet [4] is a transformer based model originally developed for Emotion Recognition in Conversations (ERC) and features a fusion-attention mechanism to modulate the attention given to each utterance considering the audio and visual features. As this model is designed for utterance emotion recognition we apply a max-pooling operation over the final outputs of fusion attention module to generate a feature representation for all the utterances in a video. Since this model provides two strategies to consider visual features: one with the video frame and another that combines multiple faces in a frame, we use them to predict either scene- or character-level emotions separately.

AttendAffectNet [15] proposes two multi-modal self-attention based approaches for predicting emotions from movie clips. We adapted the proposed Feature AttendAffectNet model in our work. It leverages the transformer encoder block where every input token represents a different modality. These modality feature vectors are generated by average pooling over respective features. Following the proposed mechanism, a classification head was attached at the end of the model for predicting multi-label emotions. We adopt the same backbone representations, MViT [5] pre-trained on Kinetics400 [3] and ResNet50 pretrained on FER13 [6], for their work to extract scene and face features respectively.

SoTA results. Reflecting Tables 4 and 5 in the main paper, we present the Table 4 and Table 5 and also include standard deviation over 3 runs.

G. Additional Qualitative Analysis

Fig. 9 shows another example (similar to Fig. 6 from the main paper) where we visualize the emotions for two characters *Jamie* and *Dylan*. We see that our model looks at relevant video frames, dialog utterances, and character representations while making the predictions. The scene described above is of a *proposal*, where the protagonist, *Dylan*, clears out some misunderstanding and proposes to the female lead character, *Jamie*, in between an ongoing flash mob (*scene*). As mentioned, in the Fig. 9 caption, both the characters develop emotion: *happy* and *excited*. From the facial expressions as well as from dialog utterances, it is apparent enough for the readers to predict emotions, but from model's point-of-view culminating all these signals and making sense of them, that too for complex human emotions, is a great job.

User study on understanding expressiveness. We asked 2 people to look at about 30 random clips that have positive labels for *angry*, *scared*, *cheerful* and independently mark yes when the emotion was apparent in the video (V), dialog (D), and character/face (C), similar to a multi-label setup. Note, our model's attention scores suggest that *cheerful* is an expressive emotion (character tokens are helpful), while *scared* and *angry* can rely on dialog and video context.

Below, we present the fraction of times each modality was picked by the users. For *angry*, the annotators favored V: 62%, D: 80%, and C: 59%, due to several neutral-faced instances with harsh dialog and violent actions. *Scared*, V: 56%, D: 48%, C: 62%, was sometimes expressed through screaming or crying, with no modality standing out strongly. Finally, *cheerful*, V: 41%, D: 64%, C: 79%, was observed most prominently on character faces and through dialog. This analysis aligns with our observations in Figure 7 of the main paper that the expressiveness scores are for faces and applicable to our particular dataset.

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple Online and Realtime Tracking. In *International Conference on Image Processing (ICIP)*, 2016. 4

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2018. 4

[3] João Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 7

[4] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 1, 7, 8

[5] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale Vision Transformers. In *International Conference on Computer Vision (ICCV)*, 2021. 7

[6] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing (ICONIPS)*, 2013. 7

[7] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. MovieNet: A Holistic Dataset for Movie Understanding. In *European Conference on Computer Vision (ECCV)*, 2020. 3

Method	Top 10		Top 25		Emotic	
	Val	Test	Val	Test	Val	Test
Random	16.87±0.23	13.84±0.20	9.73±0.10	7.57±0.08	11.47±0.11	11.36±0.09
CAER [10]	18.35±0.10	15.38±0.13	11.84±0.07	9.49±0.08	13.91±0.06	12.68±0.02
ENet [17]	19.14±0.10	16.14±0.05	11.22±0.06	9.08±0.08	13.55±0.06	12.64±0.03
AANet [15]	21.55±0.18	17.55±0.16	12.55±0.15	10.20±0.13	14.71±0.19	13.37±0.20
M2Fnet [4]	24.55±0.39	19.10±0.06	16.02±0.14	13.05±0.31	18.27±0.16	16.76±0.20
EmoTx	**34.22**±0.18	**29.35**±0.18	**23.86**±0.10	**19.47**±0.10	**23.67**±0.03	**21.40**±0.03

Table 4. Comparison against SoTA for scene-level predictions. *AANet* denotes AttendAffectNet, while *ENet* refers to EmotionNet.

Method	Top 10		Top 25		Emotic	
	Val	Test	Val	Test	Val	Test
Random	12.49±0.15	11.37±0.14	5.84±0.05	5.36±0.05	6.40±0.05	6.32±0.05
AANet [15]	17.43±0.28	16.04±0.19	8.64±0.19	7.20±0.15	8.53±0.17	7.75±0.11
M2Fnet [4]	20.82±0.28	19.01±0.45	10.67±0.38	9.71±0.34	11.30±0.35	9.92±0.02
EmoTx (Ours)	**24.35**±0.23	**22.31**±0.11	**13.36**±0.11	**11.71**±0.05	**12.29**±0.08	**11.76**±0.10

Table 5. Comparison against SoTA for character-level predictions. *AANet* denotes AttendAffectNet.

[8] Kalogeiton, Vicky, and Zisserman, Andrew. Constrained video face clustering using 1nn relations. In *British Machine Vision Conference (BMVC)*, 2020. 4

[9] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[10] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware Emotion Recognition Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 6, 8

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *International Conference on Neural Information Processing Systems (ICONIPS)*, 2013. 5

[12] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. Affect2MM: Affective Analysis of Multimedia Content Using Emotion Causality. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 5

[13] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. 4

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4

[15] Ha Thi Phuong Thao, BT Balamurali, Dorien Herremans, and Gemma Roig. AttendAffectNet: Self-Attention based Networks for Predicting Affective Responses from Movies. In *International Conference on Pattern Recognition (ICPR)*, 2021. 1, 7, 8

[16] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. MovieGraphs: Towards Understanding Human-Centric Situations from Videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[17] Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, and Dimitris Samaras. Learning Visual Emotion Representations From Web Data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 5, 8

[18] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, pages 1499–1503, 2016. 3