A. Training Details

We expand and provide a more detailed look into the feature extraction for each of the datasets in Section A.1. To ensure reproducability of our work, we provide extensive details into all training hyperparameters for all the datasets.

A.1. Feature Extraction

A.1.1 SENDv1 Dataset

We used the facial features, audio features and the text embeddings as input for SENDv1. We used the extracted features for the three modalities as explained by Ong et al. [51]. To summarize, for audio features they used openSMILE v2.3.0 [19] to extract the extended GeMAPS (eGeMAPS) set of 88 parameters for every 0.5-second window. For text features, they provide third-party commissioned professional transcripts for the videos. The transcript was then aligned (every 5 seconds) and a 300-dimensional GloVe word embeddings [55] was used. For the facial features they provide 20 action points [18] extracted using the Emotient software by iMotions [1] for each frame (30 per second).

A.1.2 LIRIS-ACCEDE Dataset

Like mentioned in Table 1, we used the facial features, audio features, scene descriptors and visual aesthetic features. While we used the already available features for audio and visual aesthetics, we extract the facial features and scene descriptors ourselves. The audio features provided were extracted using the openSMILE toolbox [2], which compute a $1,582$ dimensional feature vector. For the visual aesthetics, the authors provide the following: Auto Color Correlogram, Color and Edge Directivity Descriptor, Color Layout, Edge Histogram, Fuzzy Color and Texture Histogram, Gabor, Joint descriptor joining CEDD and FCTH in one histogram, Scalable Color, Tamura, and Local Binary Patterns extracted using the LIRE [3] library. We extracted the face features ourselves using Bulat et al. [9]. These result in 68 action units with the 3D coordinates. For the scene descriptors we Xiao et al.'s [76] 4096 dimensional intermediate representation.

A.1.3 MovieGraphs Dataset

For MovieGraphs dataset as summarized in Table 1, we use all the features except the audio as the audios were not provided in the dataset. We now explain below how we retrieve the features. We extracted the face features ourselves using Bulat et al. [9]. These result in 68 action units with the 3D coordinates. For the transcript, we used the 300-dimensional GloVe word embeddings [55] to obtain the feature representation. For visual aesthetic features, we extracted various features for color, edges, boxes and segments using Peng et al. [54]. For the scene and situation descriptors we used the provided text in the dataset and used the 300-dimensional GloVe word embeddings again to make them into feature representations.

[1] https://imotions.com/emotient/
[2] http://audeering.com/technology/opensmile/
[3] http://www.lire-project.net/

Table 6: **MovieGraphs Labels Generation:** We list the attributes of Moviegraphs used for all clips for grouping them into 26 discrete emotion labels.

Class Id	Emotion Labels	Attribute Labels available in MovieGraphs
0	Affection	loving, friendly
1	Anger	anger, furious, resentful, outraged, vengeful
2	Annoyance	annoy, frustrated, irritated, agitated, bitter, insensitive, exasperated, displeased
3	Anticipation	optimistic, hopeful, imaginative, eager
4	Aversion	disgusted, horrified, hateful
5	Confident	confident, proud, stubborn, defiant, independent, convincing
6	Disapproval	disapproving, hostile, unfriendly, mean, disrespectful, mocking, condescending, cunning, manipulative, nasty, deceitful, conceited, sleazy, greedy, rebellious, petty
7	Disconnection	indifferent, bored, distracted, distant, uninterested, self-centered, lonely, cynical, restrained, unimpressed, dismissive
8	Disquietment	worried, nervous, tense, anxious, afraid, alarmed, suspicious, uncomfortable, hesitant, reluctant, insecure, stressed, unsatisfied, solemn, submissive
9	Doubt/Conf	confused, skeptical, indecisive
10	Embarrassment	embarrassed, ashamed, humiliated
11	Engagement	curious, serious, intrigued, persistent, interested, attentive, fascinated
12	Esteem	respectful, grateful
13	Excitement	excited, enthusiastic, energetic, playful, impatient, panicky, impulsive, hasty
14	Fatigue	tire, sleepy, drowsy
15	Fear	scared, fearful, timid, terrified
16	Happiness	cheerful, delighted, happy, amused, laughing, thrilled, smiling, pleased, overwhelmed, ecstatic, exuberant
17	Pain	pain
18	Peace	content, relieved, relaxed, calm, quiet, satisfied, reserved, carefree
19	Pleasure	funny, attracted, aroused, hedonistic, pleasant, flattered, entertaining, mesmerized
20	Sadness	sad, melancholy, upset, disappointed, discouraged, grumpy, crying, regretful, grief-stricken, depressed, heartbroken, remorseful, hopeless, pensive, miserable
21	Sensitivity	apologetic, nostalgic
22	Suffering	offended, hurt, insulted, ignorant, disturbed, abusive, offensive
23	Surprise	surprise, surprised, shocked, amazed, startled, astonished, speechless, disbelieving, incredulous
24	Sympathy	kind, compassionate, supportive, sympathetic, encouraging, thoughtful, understanding, generous, concerned, dependable, caring, forgiving, reassuring, gentle
25	Yearning	jealous, determined, aggressive, desperate, focused, dedicated, diligent
26	None	-

A.2. Training Hyperparameters

Table 7 lists all the necessary dataset-specific training hyperparameters used in the proposed Affect2MM model.

	Dataset		
Hyperparameters	SENDv1	MovieGraphs	LIRIS-ACCEDE
Dropout Ratio	0.5	0.5	0.5
Optimizer	Adam	Adam	Adam
Embedding Dimension (Facial Expression)	32	204	204
Embedding Dimension (Visual Aesthetics)	N/A	41	317
Embedding Dimension (Audio)	88	300	1584
Embedding Dimension (Action/Situation)	N/A	300	N/A
Embedding Dimension (Scene)	N/A	300	4096
Embedding Dimension (Textual)	300	300	N/A
Hidden Dimension (Linear Layers)	512	1024	512
Hidden Dimension (cLSTM Encoder)	512	1024	512
Hidden Dimension (LSTM Decoder)	512	1024	512
Number of hidden layers	1	1	1
Epochs	10	10	20
Batch Size	1	1	1
Learning Rate (Affect2MM model)	1e-4	1e-4	1e-4
Learning Rate (Multivariate VAR)	0.001	0.001	0.001
Activation Function of Linear layers	LeakyReLU	LeakyReLU	LeakyReLU
Dimension of FCN Layers	[(512 x 4), (4 x 1)]	[(1024 x 4), (4 x 27)]	[(512 x 4), (4 x 2)]

Table 7: **Hyperparameters Details:** Training hyperparameters for SENDv1, MovieGraph and LIRIS-ACCEDE dataset.

A.3. MovieGraphs Dataset Labels Generation

We provide the attribute values we used as "emotional keywords" to group into the 26 emotion labels that were then used for training MovieGraphs Dataset in table 6.

B. Codes

To enable reproducability and further research in this domain we release codes for Affect2MM at https://github.com/affect2mm/emotion-timeseries.