The BLEMORE Workshop and Competition on Multimodal Blended Emotion Recognition

Tim Lachmann
tim.lachmann@su.se
Stockholm University
Sweden

Philipp Müller
philipp.mueller@dfki.de
German Research Center for Artificial Intelligence
Germany

Teimuraz Saghinadze
saghinadze.teimuraz@gtu.ge
Georgian Technical University
Georgia

Michal Balazia
michal.balazia@inria.fr
INRIA Sophia Antipolis
France

Alexandra Israelsson
alexandra.israelsson@psychology.su.se
Stockholm University
Sweden

Petri Laukka
petri.laukka@psyk.uu.se
Uppsala University
Sweden

ABSTRACT

Humans often do not only experience a single basic emotion at each moment in time, but frequently a blend of several emotions of variable intensity. Despite the importance of such blended emotions, most approaches to emotion recognition are designed to recognise single emotions only. To address this shortcoming, we present the BLEMORE workshop and competition on multimodal blended emotion recognition. BLEMORE introduces the first publicly available dataset of multimodal (video, audio) blended emotion expressions that includes information on the relative salience of each emotion contained in a blend. The dataset consists of 3,050 clips from 58 actors, performing 5 basic emotions and 10 different blends. Based on the BLEMORE dataset, we define two blended emotion prediction tasks: (1) predicting the presence of emotions in a given sample, and (2) predicting the relative salience of emotions contained in a blend. We conduct extensive baseline prediction experiments, evaluating the performance of several state-of-the-art video feature representations on these tasks. As such, the BLEMORE dataset is a valuable resource to facilitate research on emotion recognition systems that pay respect to the ubiquity of blended emotion expressions.

1 INTRODUCTION

An emotion rarely comes alone. Instead, humans often experience several emotions at the same time, which is called *blended emotions* (also called compound or mixed emotions) [24]. For example, when confronted with a loss that is found to be unjust, we may feel both sadness and anger at the same time. Blended emotions can also have varying levels of salience, as e.g. a surprise birthday party may lead to a large proportion of happiness but also a smaller proportion of surprise. Despite the relevance of blended emotions in our daily lives, research on emotion recognition systems has focused mainly on single emotions (e.g. [14]). This negligence can be attributed to (1) the limited number of suitable datasets and (2) a lack of awareness of the task of blended emotion recognition. The goal of BLEMORE is to bring emotion recognition research closer to the complexity of everyday human emotions [6], and inspire further development in recognition of blended multimodal emotion expressions. BLEMORE introduces a novel challenge on blended emotion recognition from multi-modal videos which for the first time addresses the challenging task of recognising the relative salience of emotions in a given blend. We invite two categories of submissions. First, papers addressing the BLEMORE challenge which follow a well-defined evaluation protocol. Second, we also invite submissions on the general topic of blended emotion recognition which may employ other datasets or problem formulations. By combining a competition with a workshop format, BLEMORE leads to concrete and measurable technical advances, but also provide a forum for reflection and exchange between researchers working on the still under-represented topic of blended emotion recognition.

2 RELATED WORK

Our work is related to existing datasets containing expressions of blended emotions, as well as methods for blended emotion recognition.

2.1 Datasets of Blended Emotions

The first datasets containing expressions of blended emotions were static image datasets [8, 9, 11, 18]. While CFEE and EmotionNet [8, 9] derived emotion labels from Action Unit detections, RAF-DB and iCV-MEFED [11, 18] are fully-human annotated. RAF-DB employs continuous labelling, meaning each compound emotion is represented as a continuous vector, with its components representing basic emotions. iCV-MEFED considers basic emotions and every pairing of them; hence, every compound emotion consists of dominant and complementary ones. Overall, both datasets recognize potential asymmetry between blends.

More relevant to our work are video datasets of blended emotion expressions. We present an overview over such datasets in Table 1.

C-EXPR-DB [15] consists of 400 videos sourced from YouTube, annotated frame by frame with 12 compound expressions along with other states. The annotations include valence-arousal (VA), action units (AU), speech, facial landmarks, bounding boxes, and facial attributes. Despite the relatively small number of videos, the dataset contains approximately 200,000 frames, equivalent to roughly 13 hours of footage. However, utilizing this data set to explore the interplay between single and compound expressions would be challenging, since it has no video segments with single emotion label.

MPIIEmo [23] is a video dataset of acted emotion expressions embedded in short narratives. Videos were annotated on a per-frame basis with dimensional and categorical emotion labels. The

Table 1: Overview over existing publicly available datasets on blended emotion recognition. With # Samples we refer to the number of individual video clips, even in the case of frame-wise annotations (such as in MPIIEmo).

Dataset	Participants	# Samples	# Single / Blended Samples	Single / Blended Classes	Modalities	Salience
C-EXPR-DB	-	400	-	0 + Other / 12	Visual, Audio	No
MPIIEmo	16	224	-	4 / 6	Visual, Audio	No
IMED	15	285	105 / 180	6 + Neutral / 12	Visual	No
CMED	-	1,050	385 / 665	3 / 4	Visual, ?	No
MD-MER	73	292	219 / 73	2 + Baseline / 1	Visual, ?	No
MAFW	-	8,996	4,938 / 4,058	10 + Neutral / 32	Visual, Audio	No
BlEmoRe	58	3,050	1,390 / 1,660	5 + Neutral / 10	Visual, Audio	Yes

total dataset comprises 224 videos, or 252k frames. Several categorical labels could be given to each video, but the authors did not provide an analysis of the relative frequency of blended versus single emotions.

IMED [19] consists of videos featuring 15 Indonesian subjects who were instructed to express 12 compound expressions, 6 basic expressions, and neutral expressions, totaling 285 videos. The recordings were later validated by experts.

CMED [33] is a composite dataset made up of 5 smaller datasets, with a total of 1,050 videos. The authors identified 12 compound expressions, 6 basic expressions, and neutral expressions using action unit (AU) coding. Due to the uneven distribution of classes, they further grouped the expressions into 7 categories: P (Positive), N (Negative), S (Surprised), PS, NS, PN, and NN.

MD-MER [30] is a multimodal data set comprising EEG, GSR, PPG, and frontal face videos from 73 participants. The participants were shown clips selected from a curated list of films from the Stanford film library. The recordings are categorized into three broad emotional states: positive, negative, and mixed, plus baseline.

MAFW [20] is one of the largest datasets available for compound facial expression recognition. It comprises 10,045 video clips, including audio and verbal scene descriptions, sourced from movies, TV shows, YouTube, and other media. Of these, 8,996 clips are used for the compound expression classification task: 4,938 represent a single expression (across 10 classes + neutral), while 4,058 represent compound expressions (across 32 classes). The data set is highly unbalanced; among the multiple-expression clips, 23 classes represent combinations of two expressions, while 9 classes consist of combinations of three expressions.

In contrast to previous datasets which are small (IMED, CMED, C-EXPR-DB) or highly unbalanced in the blended emotion class distributions (MAFW), we introduce BlEmoRe , the second largest dataset of blended emotion expressions which also features a highly balanced class distribution. As such, the BlEmoRe dataset will serve as a valuable resource for the development of future blended emotion recognition methods.

2.2 Methods for Blended Emotion Recognition

Video emotion recognition is a rapidly growing field. Multimodal models such as HumanOmni [32], UMBEnet [22], AVF-MAE++ [29], and VAEmo [5] mostly report better results than video-only models such as FineCliper [2], S4D [4], and MAE-DFER [26]. However, most of these methods were not evaluated on the blended emotion recognition task.

The MAFW-43 [20] dataset is the blended emotion recognition dataset that has been most frequently covered in the recent literature. However, the models AVF-MAE++ [29], T-MEP [31], HiC-MAE [27] and PTH-Net [17] fine-tuned on this dataset typically aim for optimal performance across a wide range of datasets, including the MAFW-11 class (non-blended emotion labels). Consequently, they often do not employ specific techniques to induce biases for blended emotions. One way to circumvent treating each blended emotion as a separate class, especially for datasets like RAF-DB [18] which have continuous labels, is to use only the base class in the cross-entropy loss. More recently, an interesting new objective function, called bi-center [7] loss, has been proposed. This method not only utilizes cross-entropy loss, but also introduces a loss term that anchors the features to their respective centers, corresponding to the base emotions. Overall, work on dedicated methods for blended emotion recognition remains limited.

3 THE BLEMORE CHALLENGE

3.1 Dataset

We introduce a new dataset, which features portrayals of blended emotions, recorded as part of a larger project on dynamic multimodal emotion expression wherein actors express a wide range of emotions through facial expressions, body movement and vocal sounds. This dataset was first introduced in [13], but the number of recordings included here is larger than in the previous study and is made publicly available for the first time as part of this challenge. Actors were instructed to express both single emotions (anger, disgust, fear, happiness, sadness, and neutral), and blended emotions consisting of all pairwise combinations of anger, disgust, fear, happiness, and sadness. All pairwise combinations were further conveyed with three different blend conditions:

- 50/50 = same amount of both emotions (e.g. 50/50 happiness-sadness (both happiness and sadness are expressed in equal proportions)
- 70/30 = the first emotion is more salient than the second emotion (e.g. 70/30 happiness-sadness conveys mainly happiness blended with a tinge of sadness)
- 30/70 = the second emotion is more salient than the first emotion (e.g. 30/70 happiness-sadness conveys mainly sadness blended with a tinge of happiness)

The actors were instructed that the emotions should be conveyed as convincingly as possible, as if interacting with another person (the camera), but without using overtly stereotypical expressions.

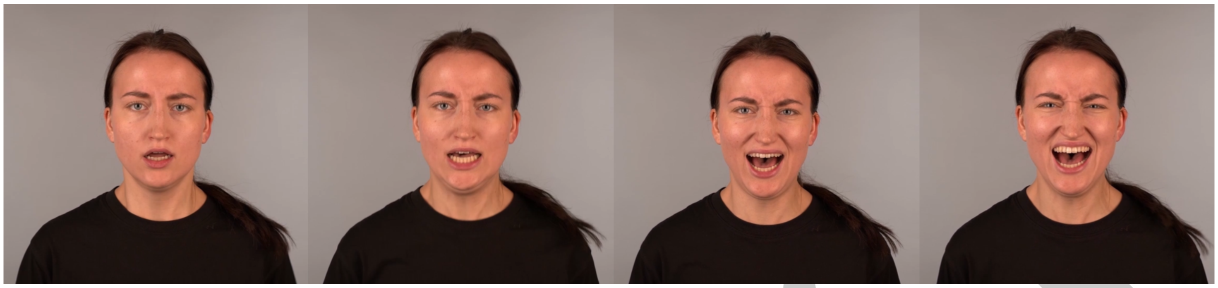


Figure 1: Examples of stills from the video recordings. The actor portrays a combination of anger and fear. Reproduced from [13] under CC BY 4.0.

It was also specified that they should try to express the emotion simultaneously through both the face/body and the voice. For the vocal expressions, they were free to choose any non-linguistic vocalization (e.g. cries, laughter, groans), but no words (including made up words) were allowed [16]. Example stills from an example portrayal of a blended emotion is shown in Figure 1.

Recordings were conducted in a room with studio lighting and dampened acoustics, using a high-quality camera and microphone. The audio level was calibrated relative to the loudest expected level and then kept constant during the recording session. The camera was placed in front of the actor at a distance of approximately 1.2 m, and the microphone was located 0.5 m above the actor and directed at the actor's chest (for details, see [13]).

For the current challenge we have selected recordings from 58 actors, for a total of 1390 recordings of single emotions, and 1660 recordings of blended emotions (see **Fig 3**). The duration of the recordings ranges from 1-30 seconds (see **Fig 2**).

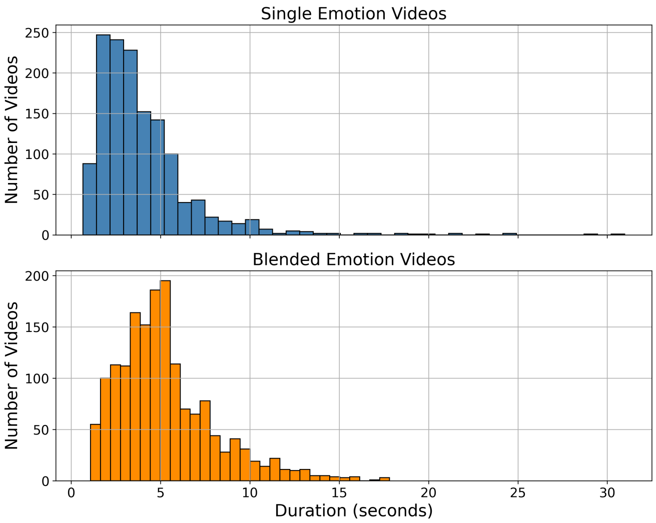


Figure 2: Distribution of video durations in the dataset for single and blended emotions.

The data were partitioned into a predefined training and test split with no actor overlap: 43 actors are included in the training set and 15 actors in the test set. The test set contains a fixed subset of actors selected to ensure balanced gender representation.

To enable consistent validation, we also provide five predefined cross-validation folds within the training set. Each fold contains a disjoint subset of actors and is approximately balanced in terms of gender and number of samples. This setup ensures that models are evaluated on their ability to generalize to previously unseen individuals.

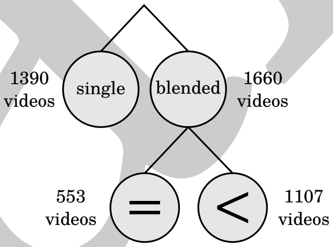


Figure 3: Structure of BLEMORE full dataset (train and test partition) which contains single emotions and blended emotion expressed with equal (=) and unequal (<) salience.

3.2 Evaluation Metrics

For this challenge, we employ two evaluation metrics: $ACC_{presence}$ and $ACC_{salience}$.

- $ACC_{presence}$ measures whether the correct label(s) are predicted without errors. A correct prediction must include all present emotions while avoiding false negatives (e.g., predicting only one emotion in a blend) and false positives (e.g., predicting emotions that are not part of the label).
- $ACC_{salience}$ extends $ACC_{presence}$ by considering the relative prominence of each emotion. It evaluates whether the predicted proportions reflect the correct ranking—whether the emotions are equally present or one is more dominant than the other. This metric only applies to blended emotions.

$ACC_{presence}$ is in effect a generic way to measure (blended) emotion recognition, and is easily comparable to models implemented on other datasets. $ACC_{salience}$ captures the unique feature of our dataset - emotion salience in different blend conditions.

We provide all training with ground truth labels to participants. Test data are shared without ground truth information. For a fair comparison, participants submit their prediction for evaluation on our servers. In addition, we encourage submissions to provide information about how various blended emotions are expressed, e.g. by providing lists of predictor importance, to increase the explainability of their systems.

4 METHOD

In this section, we describe the baseline architecture used for the BlEmoRe challenge. Our approach extracts video features using pre-trained encoders and applies a simple feed-forward neural network for multi-label prediction.. We detail the video encoding methods, feature aggregation procedure, and model training pipeline below.

4.1 Video Encoding Methods

We extracted features using five pre-trained models, encompassing both frame-based and spatiotemporal encoders:

- **OpenFace 2.0** [1]: OpenFace extracts frame-level facial behavior features, including 17 action unit (AU) intensities, 6 head pose parameters, and 8 gaze features, for a total of 31 features per frame. OpenFace operates at the frame level and captures fine-grained facial dynamics.
- **CLIP** [25]: We use the vision encoder from CLIP to obtain frame-level embeddings. CLIP encodes each frame independently and was trained using contrastive learning on a large corpus of image-text pairs.
- **ImageBind** [10]: Similar to CLIP, ImageBind provides frame-level embeddings by aligning visual information with multiple sensory modalities. We use its vision encoder to extract per-frame features.
- **VideoMAE V2** [28]: VideoMAEv2 operates on spatiotemporal tube tokens and reconstructs masked video cubes. We use the ViT-B/16 variant to extract spatiotemporal features. VideoMAEv2 encodes temporal dynamics by operating on 3D spatiotemporal patches (cubes).
- **Video Swin Transformer** [21]: This hierarchical spatiotemporal encoder extends the Swin Transformer to video by applying shifted 3D window attention. It extracts features from local spatiotemporal neighbourhoods, capturing both spatial and temporal context.

Thus, OpenFace, CLIP, and ImageBind provide frame-level representations, while VideoMAEv2 and Video Swin Transformer yield spatiotemporal representations based on patches or cubes.

4.2 Audio Encoding Methods

We use two large self-supervised audio encoders trained on raw 16kHz waveform speech to obtain frame-level representations at 20ms resolution, each producing 1024-dimensional embeddings:

- **HuBERT** [12]: A model trained on 60k hours of speech using masked prediction of k-means cluster labels. We use the Large variant (LL-60k).
- **WavLM** [3]: A model trained on 94k hours of speech with a combined masked prediction and denoising objective. We use the Large configuration.

4.3 Feature Aggregation and Subsampling

We employed two different strategies to obtain video-level features:

- **Aggregation**: For frame-level encoders (OpenFace, CLIP, ImageBind, HuBERT, WavLM), and as one variant for spatiotemporal encoders (VideoMAEv2, Video Swin Transformer), we computed seven statistical measures per feature dimension: mean, standard deviation, 10th, 25th, 50th (median), 75th, and 90th percentiles. The aggregated statistics were concatenated into a fixed-size feature vector.
- **Subsampling**: For spatiotemporal encoders (VideoMAEv2, Video Swin Transformer), videos are processed as sequences of fixed-length clips, each consisting of 16 frames. Each clip is passed through the encoder to obtain an embedding summarizing the spatiotemporal information within that segment. We refer to these embeddings as *subsamples*. Under the subsampling strategy, instead of aggregating the subsample embeddings into a single video-level vector, we treat each subsample independently during training.

Prior to training, all features were standardized to zero mean and unit variance using the statistics computed on the training set.

4.4 Multi-Modal Fusion

To assess the potential of combining visual and auditory information, we implemented a simple **early fusion** approach. Specifically, we concatenated aggregated features from pairs of video and audio encoders before feeding them into the classifier.

Rather than exhaustively exploring all encoder combinations, we selected the two top-performing video encoders (VideoMAEv2 and ImageBind) and paired each with both audio encoders (HuBERT and WavLM).

4.5 Label Encoding

Each video was annotated with either a single or blended emotion along with relative salience information.

- **Single-emotion recordings** were encoded as one-hot vectors with a value of 1 at the corresponding emotion index and 0 elsewhere.
- **Blended emotions** were represented as soft probability distributions over two emotions, where the label values correspond to the salience proportions normalized to sum to 1 (e.g., 70% happiness and 30% sadness are encoded as [0, 0, 0, 0.7, 0.3, 0]).

The resulting target vector for each video has six dimensions, one for each emotion category (anger, disgust, fear, happiness, sadness, and neutral). Each dimension can either have a hard assignment in the case of single-emotion recordings, or soft assignments for blended emotions, except in the case of neutral which never occurs within blends.

4.6 Classification Models

We trained simple feedforward neural networks with three different configurations:

- **Linear**: A single linear layer mapping input features directly to emotion probabilities.

- **MLP-256**: A network with one hidden layer containing 256 units, followed by ReLU activation.
- **MLP-512**: A network with one hidden layer containing 512 units, followed by ReLU activation.

All models produced a six-dimensional output corresponding to the six emotion categories. A softmax activation was applied to the output logits to obtain a probability distribution over the classes.

The models were trained to minimize the Kullback–Leibler divergence between the predicted softmax distribution and the ground-truth label distribution, reflecting either a hard one-hot vector for single emotions or a soft distribution for blended emotions.

4.7 Training and Post-processing

Training was done using the Adam optimizer with a batch size of 32 for aggregation-based features and 512 for subsampled features. The learning rate was set to 5×10^{-6} with weight decay of 1×10^{-3}. Models were trained for a fixed number of epochs: 200 epochs for aggregation-based features and 300 epochs for subsampled features.

To convert model outputs into discrete emotion presence and salience predictions, we applied a post-processing step involving thresholding. A grid search was conducted on the validation folds to find the optimal thresholds:

- **Presence threshold** (α): Determines which emotions are predicted as present.
- **Salience threshold** (β): Determines whether the blend is classified as equal (50/50) or dominant/subdominant (70/30 or 30/70).

The optimal values of α and β were selected based on the validation runs and subsequently applied to the held-out test set for final evaluation.

For the subsampling approach, predictions were first generated at the subsample level. The final video-level prediction was obtained by averaging the logits across all subsamples before applying the softmax activation and thresholding steps.

5 RESULTS

We evaluated the baseline models using five-fold cross-validation on the training set and report performance in terms of presence accuracy (ACC_{presence}) and salience accuracy (ACC_{salience}). Model selection was based on a composite validation score, computed as the average of presence and salience accuracy.

We evaluated three classifier configurations: **Linear**, **MLP-256**, and **MLP-512**. Across modalities and settings, the **MLP-512** configuration usually yielded the best validation performance. The results are reported for this configuration.

To contextualize the model performance, we computed trivial baselines based on class distributions. For the single-emotion baseline, always predicting the most frequent single emotion resulted in ACC_{presence} of 7.82% on the training set and 7.41% on the test set, with ACC_{salience} of 0% by definition. For the blend baseline, always predicting the most frequent blend yielded ACC_{presence} of 5.70% and 5.89%, and ACC_{salience} of 3.54% and 3.59% on the training and test sets, respectively.

5.1 Validation Results

Table 2 summarizes the cross-validated performance (mean ± std) for each encoder, combinations of encoders, and method. Aggregation-based features generally outperformed subsampled features. Among unimodal encoders, ImageBind yielded the highest composite score (ACC_{presence}=0.290, ACC_{salience}=0.130), WavLM had the best performance overall among the audio encoders (ACC_{presence}=0.265, ACC_{salience}=0.121). Multimodal combinations consistently outperformed unimodal variants, with ImageBind + WavLM achieving the best composite score (ACC_{presence}=0.345, ACC_{salience}=0.170).

For subsampled features, the highest scores were achieved by VideoMAEv2, with ACC_{presence}=0.260, ACC_{salience}=0.124.

Table 2: Validation results (mean ± std) over folds for aggregation-based and subsampled features using the MLP-512 model. Multimodal results are included at the bottom.

Encoder	Method	ACC_{presence}	ACC_{salience}
CLIP	Aggregation	0.266 ± 0.021	0.105 ± 0.012
ImageBind	Aggregation	0.290 ± 0.028	0.130 ± 0.008
OpenFace	Aggregation	0.228 ± 0.014	0.119 ± 0.014
VideoMAEv2	Aggregation	0.273 ± 0.025	0.106 ± 0.014
VideoSwin	Aggregation	0.225 ± 0.026	0.089 ± 0.033
HuBERT	Aggregation	0.243 ± 0.023	0.104 ± 0.024
WavLM	Aggregation	0.265 ± 0.027	0.121 ± 0.012
VideoMAEv2	Subsampling	0.260 ± 0.030	0.124 ± 0.027
VideoSwin	Subsampling	0.210 ± 0.024	0.103 ± 0.020
ImageBind + WavLM	Aggregation	0.345 ± 0.035	0.170 ± 0.055
ImageBind + HuBERT	Aggregation	0.339 ± 0.023	0.158 ± 0.053
VideoMAEv2 + WavLM	Aggregation	0.343 ± 0.022	0.140 ± 0.028
VideoMAEv2 + HuBERT	Aggregation	0.332 ± 0.016	0.138 ± 0.012
trivial baseline (single emotion)		0.078	0.000
trivial baseline (blend)		0.057	0.035

5.2 Test Set Results

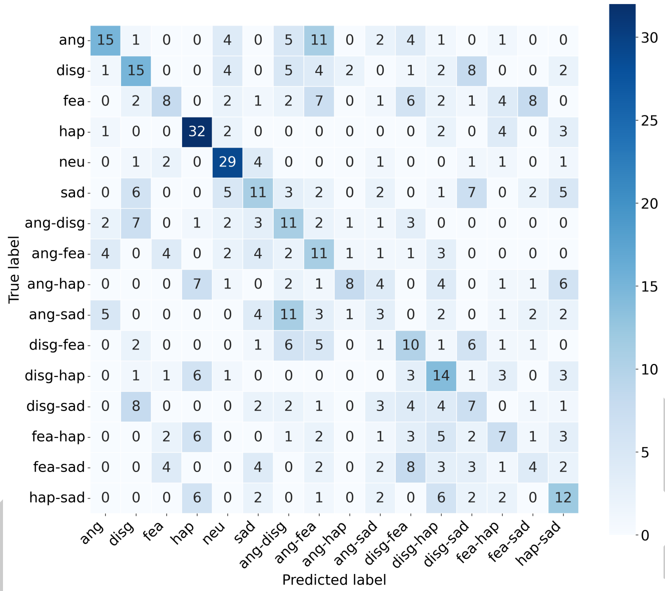
Table 3 summarizes test performance for the best models. The top unimodal model was WavLM (ACC_{presence} = 0.311, ACC_{salience} = 0.084). Among visual models, VideoMAEv2 reached the highest ACC_{presence} = 0.293, but lower ACC_{salience} = 0.054. Multimodal combinations showed clear benefits, VideoMAEv2 + HuBERT yielded the best results with ACC_{presence} = 0.332 and ACC_{salience} = 0.114.

5.3 Confusion Matrix

To better understand prediction patterns on the test set, we visualized a confusion matrix based on the post-processed outputs of the best-performing model (VideoMAEv2 + HuBERT). Each sample was assigned a single composite label corresponding to its ground-truth emotion configuration (e.g., *happiness-sadness*, *neutral*). This allows for a qualitative inspection of how well the model distinguishes between single and blended emotions. The confusion matrix is shown in Figure 4.

Table 3: Test set results for aggregation-based and subsampled features using the MLP-512 model. Multimodal results are included at the bottom.

Encoder	Method	ACC_{presence}	ACC_{salience}
CLIP	Aggregation	0.258	0.096
ImageBind	Aggregation	0.261	0.087
OpenFace	Aggregation	0.226	0.081
VideoMAEv2	Aggregation	0.293	0.054
VideoSwin	Aggregation	0.214	0.093
HuBERT	Aggregation	0.274	0.120
WavLM	Aggregation	0.311	0.084
VideoMAEv2	Subsampled	0.231	0.081
VideoSwin	Subsampled	0.197	0.087
VideoMAEv2 + HuBERT	Aggregation	0.332	0.114
VideoMAEv2 + WavLM	Aggregation	0.332	0.102
ImageBind + HuBERT	Aggregation	0.298	0.084
ImageBind + WavLM	Aggregation	0.327	0.114
trivial baseline (single emotion)		0.074	0.000
trivial baseline (blend)		0.059	0.036

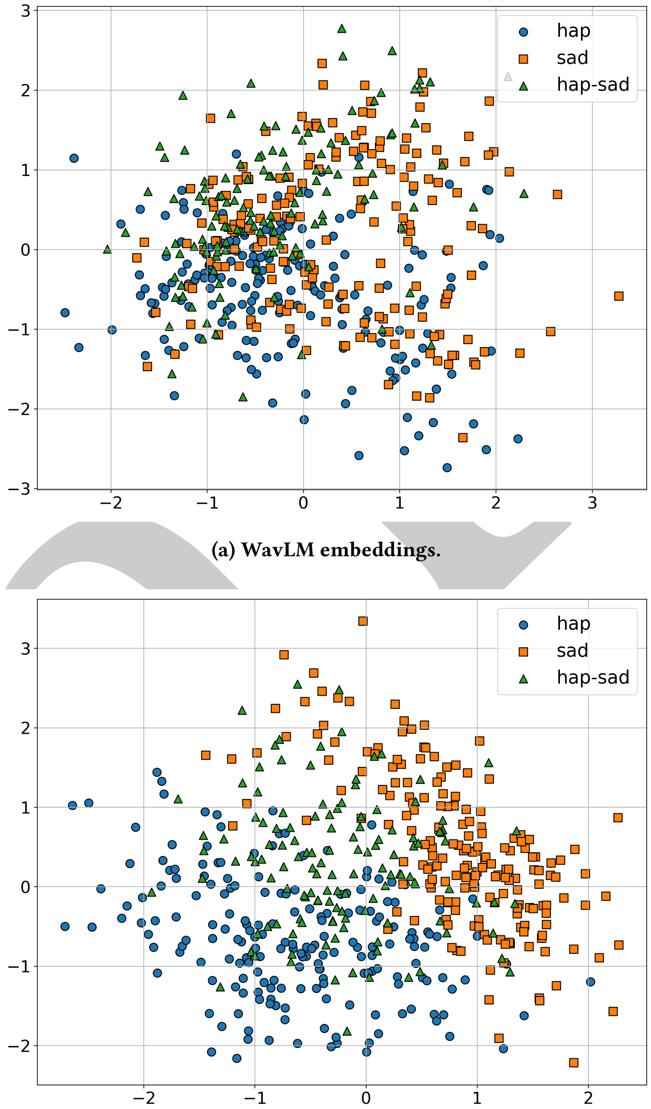


	ang	disg	fea	hap	neu	sad	ang-disg	ang-fea	ang-hap	ang-sad	disg-fea	disg-hap	disg-sad	fea-hap	fea-sad	hap-sad
ang	15	1	0	0	4	0	5	11	0	2	4	1	0	1	0	0
disg	1	15	0	0	4	0	5	4	2	0	1	2	8	0	0	2
fea	0	2	8	0	2	1	2	7	0	1	6	2	1	4	8	0
hap	1	0	0	32	2	0	0	0	0	0	0	2	0	4	0	3
neu	0	1	2	0	29	4	0	0	0	1	0	0	1	1	0	1
sad	0	6	0	0	5	11	3	2	0	2	0	1	7	0	2	5
ang-disg	2	7	0	1	2	3	11	2	1	1	3	0	0	0	0	0
ang-fea	4	0	4	0	2	4	2	11	1	1	1	3	0	0	0	0
ang-hap	0	0	0	7	1	0	2	1	8	4	0	4	0	1	1	6
ang-sad	5	0	0	0	0	4	11	3	1	3	0	2	0	1	2	2
disg-fea	0	2	0	0	0	1	6	5	0	1	10	1	6	1	1	0
disg-hap	0	1	1	6	1	0	0	0	0	0	3	14	1	3	0	3
disg-sad	0	8	0	0	0	2	2	1	0	3	4	4	7	0	1	1
fea-hap	0	0	2	6	0	0	1	2	0	1	3	5	2	7	1	3
fea-sad	0	0	4	0	0	4	0	2	0	2	8	3	3	1	4	2
hap-sad	0	0	0	6	0	2	0	1	0	2	0	6	2	2	0	12

Figure 4: Confusion matrix for the test set using the best overall model (VideoMAEv2 + HuBERT, Aggregation).

5.4 Feature Visualization

To qualitatively explore the structure of the encoded representations, we applied PCA (Principal Component Analysis) to the aggregated embeddings obtained from the WavLM and VideoMAEv2 encoders. For visualization purposes only, we applied a per-actor normalization strategy that standardizes features within each individual actor. This reduces actor-specific biases and prevents clustering based on identity, which otherwise dominates the projection space. The resulting 2D plots are shown in Figure 5 for samples labeled with *happy*, *sad*, or their blends. The clusters show distinct patterns corresponding to single and blended emotion portrayals.



(a) WavLM embeddings.

(b) VideoMAEv2 embeddings.

Figure 5: 2D PCA projection of embeddings for the happy and sad emotions.

5.5 Discussion

Overall, aggregation-based features consistently outperformed subsampled features across both visual and audio modalities. This suggests that pooling frame-level or segment-level features into global summary statistics provides more robust representations for downstream prediction.

Multimodal fusion provided a clear performance boost over unimodal features, particularly on the validation set where multimodal fusions clearly outperformed every unimodal encoding both in terms of ACC_{presence} and ACC_{salience}. This suggests that audio and visual modalities carry complementary information that can be leveraged even with a simple early fusion strategy.

There is a general drop in model performance on the test set even though these models are trained on a larger amount of data (the entire training set). Furthermore, the ranking of different models is not always congruent between validation and test evaluation.

The drop in performance can at least in part be explained by the fact that we do not supervise when to stop training by using any held out validation set in this phase.

Another important factor to consider is that while the models are trained to minimize Kullback-Leibler divergence, which encourages the predicted distribution to match the soft target labels, the post-processing applies discrete presence and salience thresholds to translate the distribution into presence/non-presence and equal/unequal salience of emotions. In the validation phase, these thresholds can be tuned dynamically on the predicted distribution. On the test set, there is no such accommodation, which introduces a source of randomness into the final test performance. If the predicted distributions at the end of training align well with the threshold values tuned on the validation set, performance improves; however, small drifts can lead to drops in presence or salience accuracy. This is evident in the test performance of the VideoMAEv2 Aggregation model. The ACC_{presence} is high while the ACC_{salience} is close to trivial performance. Likely, a high presence threshold allowed the model to minimize the impact of false positives, yielding high ACC_{presence}, while also predicting fewer blended emotions, which in turn yielded a lower ACC_{salience}. These issues highlight the challenges of blended emotion recognition, which in our implementation is modelled as a multi-label, soft classification task. However, this problem can be approached in several alternative ways, such as ranking-based formulations, regression on salience proportions, or multi-task setups, each with different trade-offs.

REFERENCES

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[2] Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. 2024. Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2301–2310.

[3] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (Oct. 2022), 1505–1518.

[4] Yin Chen, Jia Li, Yu Zhang, Zhenzhen Hu, Shiguang Shan, Meng Wang, and Richang Hong. 2025. Static for Dynamic: Towards a Deeper Understanding of Dynamic Facial Expressions Using Static Expression Data. arXiv:2409.06154 [cs.CV]

[5] Hao Cheng, Zhiwei Zhao, Yichao He, Zhenzhen Hu, Jia Li, Meng Wang, and Richang Hong. 2025. VAEmo: Efficient Representation Learning for Visual-Audio Emotion with Knowledge Injection. *arXiv preprint arXiv:2505.02331* (2025).

[6] Alan Cowen, Disa Sauter, Jessica Tracy, and Dacher Keltner. 2019. Mapping the Passions: Toward a High-Dimensional Taxonomy of Emotional Experience and Expression. *Psychological Science in the Public Interest* 20 (07 2019), 69–90.

[7] Rongkang Dong and Kin-Man Lam. 2024. Bi-center loss for compound facial expression recognition. *IEEE Signal Processing Letters* 31 (2024), 641–645.

[8] Shichuan Du and Aleix M Martinez. 2015. Compound facial expressions of emotion: from basic research to clinical applications. *Dialogues in clinical neuroscience* 17, 4 (2015), 443–455.

[9] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. 2016. 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*. 5562–5570.

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[11] Jianzhu Guo, Zhen Lei, Jun Wan, Egils Avots, Noushin Hajarolasvadi, Boris Knyazev, Artem Kuharenko, Julio C Silveira Jacques Junior, Xavier Baró, Hasan Demirel, et al. 2018. Dominant and complementary emotion recognition from still images of faces. *IEEE Access* 6 (2018), 26391–26403.

[12] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv:2106.07447 [cs.CL]

[13] Alexandra Israelsson, Anja Seiger, and Petri Laukka. 2023. Blended Emotions can be Accurately Recognized from Dynamic Facial and Vocal Expressions. *Journal of Nonverbal Behavior* 47 (05 2023), 1–18.

[14] Smith K. Khare, Victoria Blanes-Vidal, Esmaeil S. Nadimi, and U. Rajendra Acharya. 2024. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion* 102 (2024), 102019.

[15] Dimitrios Kollias. 2023. Multi-Label Compound Expression Recognition: C-EXPR Database & Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5589–5598.

[16] Petri Laukka, Hillary Elfenbein, Nela Söder, Henrik Nordström, Jean Althoff, Wanda Chui, Frederick Iraki, Thomas Rockstuhl, and Nutankumar Thingujam. 2013. Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in psychology* 4 (07 2013), 353.

[17] Min Li, Xiaoqin Zhang, Tangfei Liao, Sheng Lin, and Guobao Xiao. 2024. PTH-Net: Dynamic Facial Expression Recognition without Face Detection and Alignment. *IEEE Transactions on Image Processing* (2024).

[18] Shan Li, Weihong Deng, and JunPing Du. 2017. 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*. 2852–2861.

[19] Dewi Yanti Liliana, T. Basaruddin, and Imelda Ika Dian Oriza. 2018. The Indonesian Mixed Emotion Dataset (IMED): A Facial Expression Dataset for Mixed Emotion Recognition. In *Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality* (Nagoya, Japan) *(AIVR 2018)*. Association for Computing Machinery, New York, NY, USA, 56–60.

[20] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022. *MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild.* ACM, New York, NY, USA.

[21] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video Swin Transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3192–3201.

[22] Xinji Mai, Junxiong Lin, Haoran Wang, Zeng Tao, Yan Wang, Shaoqi Yan, Xuan Tong, Jiawen Yu, Boyang Wang, Ziheng Zhou, et al. 2024. All rivers run into the sea: Unified modality brain-inspired emotional central mechanism. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 632–641.

[23] Philipp M Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling. 2015. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 663–669.

[24] Vincent Oh and Eddie Tong. 2022. Specificity in the Study of Mixed Emotions: A Theoretical Framework. *Personality and Social Psychology Review* 26 (04 2022), 108886832210833.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR.

[26] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6110–6121.

[27] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2024. Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *Information Fusion* 108 (2024), 102382.

[28] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14549–14560.

[29] Xuecheng Wu, Heli Sun, Yifan Wang, Jiayu Nie, Jie Zhang, Yabing Wang, Junxiao Xue, and Liang He. 2025. AVF-MAE++: Scaling Affective Video Facial Masked Autoencoders via Efficient Audio-Visual Self-Supervised Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 9142–9153.

[30] Pei Yang, Niqi Liu, Xinge Liu, Yezhi Shu, Wenqi Ji, Ziqi Ren, Jenny Sheng, Minjing Yu, Ran Yi, Dan Zhang, and Yong-Jin Liu. 2024. A Multimodal Dataset for Mixed

Emotion Recognition. *Scientific Data* 11 (08 2024).

[31] Xiaoqin Zhang, Min Li, Sheng Lin, Hang Xu, and Guobao Xiao. 2023. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 5 (2023), 3192–3203.

[32] Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Weixuan chen, Xihan Wei, and Liefeng Bo. 2025. HumanOmni: A Large Vision-Speech Language Model for Human-Centric Video Understanding.
arXiv:2501.15111 [cs.CV]

[33] Yue Zhao and Jiancheng Xu. 2020. Compound Micro-Expression Recognition System. In *2020 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*. 728–733.