This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Ensemble-based Prototype-Augmented Multimodal Fusion for Ambivalence/Hesitancy Recognition

Elena Ryumina
St. Petersburg Federal Research Center of the Russian Academy of Sciences
St. Petersburg, Russia
ryumina.e@iias.spb.su

Alexandr Axyonov
St. Petersburg Federal Research Center of the Russian Academy of Sciences
St. Petersburg, Russia
axyonov.a@iias.spb.su

Dmitry Sysoev
HSE University
St. Petersburg, Russia
dsysoev@hse.ru

Timur Abdulkadirov
HSE University
St. Petersburg, Russia
tnabdulkadirov@edu.hse.ru

Kirill Almetov
HSE University
St. Petersburg, Russia
koalmetov@edu.hse.ru

Yulia Morozova
HSE University
St. Petersburg, Russia
yuvmorozova_1@edu.hse.ru

Dmitry Ryumin
HSE University
St. Petersburg Federal Research Center of the Russian Academy of Sciences
St. Petersburg, Russia
daryumin@hse.ru

Abstract

Ambivalence/hesitancy recognition in unconstrained videos is a challenging problem due to the subtle, multimodal, and context-dependent nature of this behavioral state. In this paper, a multimodal approach for video-level ambivalence/hesitancy recognition is presented for the 10th ABAW Competition. The proposed approach integrates four complementary modalities: scene, face, audio, and text. Scene dynamics are captured with a VideoMAE-based model, facial information is encoded through emotional frame-level embeddings aggregated by statistical pooling, acoustic representations are extracted with Emotion-Wav2Vec2.0 and processed by a Mamba-based temporal encoder, and linguistic cues are modeled using fine-tuned transformer-based text models. The resulting unimodal embeddings are further combined using multimodal fusion models, including prototype-augmented variants. Experiments on the BAH corpus demonstrate clear gains of multimodal fusion over all unimodal baselines. The best unimodal configuration achieved an average MF1 of 70.02%, whereas the best multimodal fusion model reached 83.25%. The highest final test performance, 71.42%, was obtained by an ensemble of five prototype-augmented fusion models. The obtained results highlight the importance of complementary multimodal cues and robust fusion strategies for ambivalence/hesitancy recognition. The source code is publicly available[1].

1. Introduction

Affective computing aims to endow intelligent systems with the ability to perceive, model, and interpret human affect from signals such as facial behavior [42], speech [13], language [7], and body motion [26]. This capability is important for human-computer interaction, healthcare, education, and assistive technologies, where decisions often depend on subtle and context-dependent behavioral cues [34]. Within this area, the Ambivalence/Hesitancy (A/H) Video Recognition Challenge of the 10th Workshop and Competition on Affective & Behavior Analysis in-the-Wild (ABAW) focuses on a particularly difficult binary task: given a video, the goal is to predict whether it contains A/H or not at the video level [14].

A/H recognition is important because these states are strongly linked to decision uncertainty, resistance, and fluctuating motivation during behavior change. In digital behavioral health interventions, such signals can help identify

[1] https://github.com/LEYA-HSE/ABAW10-BAH

whether a person is ready to change, struggling with conflicting intentions, or at risk of disengagement [4]. Unlike basic emotions (such as happiness, surprise), A/H is subtle and often manifests through inconsistencies across modalities, for example, between what a person says, how they say it, and how they look while speaking. This makes the task inherently multimodal and particularly challenging [14, 34].

González-González et al. [14] established baselines using visual facial information, audio, and speech transcripts, and showed that text was the best unimodal cue, while multimodal fusion could yield additional gains but required specialized designs to capture cross-modal inconsistencies characteristic of A/H. Similarly, Savchenko and Savchenko [44] explored face, audio, and text modalities, with the best validation result achieved by combining textual and facial features, whereas Hallmen et al. [17] used text, vision, and audio, and their best final Behavioural Ambivalence/Hesitancy (BAH) submission was based on trimodal fusion. Therefore, multimodal modeling remains a promising direction for A/H recognition, especially when modality interactions are explicitly taken into account and fusion preserves modality-specific cues under uncertainty or partially inconsistent multimodal evidence [11, 53].

Beyond A/H-specific work, multimodal affective computing has highlighted the importance of modality-aware fusion. Prior work [16, 23, 24, 28, 29, 38–41, 54] suggests that integrating complementary visual, acoustic, and linguistic signals through dedicated fusion strategies can improve the robustness of affect-related prediction, while compact facial affect representations may serve as transferable features for downstream affective analysis.

This work proposes a multimodal approach that integrates audio, text, face, and scene information for video-level A/H recognition. First, a dedicated unimodal model is trained for each modality to learn compact modality-specific representations. The resulting modality embeddings are then projected into a shared latent space and fused by a Transformer-based multimodal module operating on modality tokens and augmented with a prototype-based classification objective. In this way, inter-modality dependencies are modeled directly at the video level, while the strengths of the specialized unimodal encoders are preserved for the final A/H prediction.

2. Related Work

González-González et al. [14] presented a broad benchmark suite including supervised unimodal, bimodal, and trimodal models based on facial video, audio, and speech transcripts, as well as zero-shot and personalization experiments. To study multimodal learning, they compared several fusion strategies, including concatenation-based fusion, co-attention, transformer-based fusion, and cross-attention fusion. Their results showed that text is the strongest unimodal cue, while effective A/H recognition requires specialized multimodal and temporal modeling to capture subtle inconsistencies across modalities. The official baseline reported for the current 10th ABAW challenge corresponds to a simpler zero-shot Multimodal Large Language Model (M-LLM) setup based on Video-LLaVA [30] with a simple prompt and vision modality only.

Among the approaches proposed for the first public A/H challenge, Hallmen et al. [17] used a trimodal architecture combining text, vision, and audio. Their framework extracts visual features with a Vision Transformer (ViT) [5], audio representations with Wav2Vec 2.0 [2], and transcript embeddings with the Bidirectional Encoder Representations from Transformers (BERT) [9] text encoder, applies temporal modeling to the visual and audio streams with Long Short-Term Memorys (LSTMs) [19], and fuses the resulting modality-specific representations with an Multilayer Perceptron (MLP) [37]. Their analysis showed that text is the most informative modality. The best final BAH submission was obtained using trimodal fusion, which is consistent with recent studies [11, 27] indicating that preserving complementary modality-specific information during fusion improves multimodal prediction.

In contrast, Savchenko and Savchenko [44] used a lighter pipeline based on facial, acoustic, and textual features extracted with EmotiEffLib [45], Wav2Vec 2.0 [2], and RoBERTa-based [31] text representations. Audio and text features were aligned to the visual timeline by interpolation, and fusion was implemented either through early concatenation followed by a feed-forward classifier or through blending of unimodal predictions. In addition, a video-level Logistic Regression model over pooled text features was used as a filtering step. Their experiments showed that text was the strongest unimodal modality, while combining text and facial information yielded the best validation result.

The 10th ABAW submitted approaches mainly differ in the development of multimodal fusion strategies. Bekhouche et al. [3] used VideoMAE [48], HuBERT [20] and GoEmotions [6] for feature extraction from video, audio and test data, respectively. The authors also used modality-specific attention pooling and explicit pairwise conflict features. After that, they combined conflict-augmented early fusion with late text-guided fusion. Pereira et al. [33] also relied on SigLip2 [49] visual, HuBERT acoustic, and Foundation to Feature Large Language Model (F2LLM) [55] linguistic features, but added a fourth statistical modality and replaced a single fusion block with a committee of MLP, Random Forest, and Gradient Boosted Decision Trees (GBDT) classifiers over all modality combinations, combined by Particle Swarm Optimization (PSO)-based hard voting. Tang et al. [47] moved to the M-LLM setting: they split videos into short clips, fine-tuned Qwen3-Omni [52] on audiovisual prompt triplets, and aggregated

clip-level predictions with max pooling and majority voting across multiple fine-tuned models.

In summary, prior work shows that text is the strongest unimodal cue for A/H, while multimodal fusion remains beneficial, especially when it preserves complementary information across modalities and accounts for uncertainty in multimodal evidence. Unlike previous approaches, which mainly rely on face, audio, and text with relatively simple fusion schemes, our approach additionally incorporates scene information and uses a Transformer-based fusion module over modality-specific representations learned by dedicated unimodal models, further regularized by a prototype-based classification objective.

3. Proposed Approach

The overall pipeline of the proposed approach is shown in Figure 1. Recent multimodal approaches to A/H recognition rely on facial, acoustic, and textual features, and their results indicate that text is the strongest unimodal source of information. However, these methods also show that A/H cannot be reliably recognized from a single modality alone, as it often reflected in subtle cross-modal inconsistencies and temporally evolving behavioral patterns. In this work, we therefore extend the standard trimodal setup by incorporating scene-based visual information, which provides contextual nonverbal cues beyond the face, including posture, hand activity, and other visible manifestations of A/H-related behavior. To preserve complementary modality-specific information, we first learn dedicated unimodal features and then combine them using a Transformer-based fusion module. In addition, we consider a prototype-augmented variant in which the fused multimodal features is compared with learnable class-specific prototypes through an auxiliary training loss, while the final prediction is produced by the main classifier. The following subsections describe the main components of the proposed approach.

3.1. Scene-based Visual Model

To analyze behavioral dynamics and detect uncertainty, we employ the Video Masked Autoencoder (VideoMAE) architecture [48] based on the ViT [10] framework and pretrained on the Kinetics-400 corpus [21]. For each video, N frames are uniformly sampled, resized to 224×224, and normalized using ImageNet [8] statistics. The input video clip is processed using tubelet embedding, where it is partitioned into non-overlapping spatio-temporal patches of size $2 \times 16 \times 16$. These tubelets are projected into a $D = 768$ latent space and combined with learnable positional embeddings.

The resulting frames are processed by a Transformer encoder with spatio-temporal self-attention to model spatio-temporal dependencies. A compact scene embedding h_s is obtained by applying global average pooling to the output frames:

$$h_s = \frac{1}{N} \sum_{i=1}^{N} z_i, \tag{1}$$

where N is the number of output frames and z_i denotes the representation of the i-th frame.

3.2. Face-based Visual Model

For each video, N frames are uniformly sampled, and a YOLO-based face detector[2] is applied to each sampled frame. When multiple faces are detected, the largest bounding box is selected. If no face is detected, the full frame is used as a fallback crop.

The cropped face is resized to 224×224 and normalized with ImageNet [8] statistics, then passed to an EfficientNetB0 [46] extractor fine-tuned on the AffectNet+ corpus [12], hereafter referred to as EmotionEfficientNetB0. The extractor produces one emotional embedding vector per sampled frame. No extra embedding normalization is applied.

For each video, frame-level embeddings $\{z_i\}_{i=1}^{N}$ are aggregated using statistical pooling:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} z_i, \qquad \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (z_i - \mu)^2}. \tag{2}$$

The final face representation is formed by the concatenation $[\mu; \sigma]$. The resulting statistical representation is then used as input to an MLP, as shown in Figure 2.

3.3. Acoustic Model

For the audio modality, the audio track is extracted from each video and resampled to 16 kHz. A pre-trained Wav2Vec2.0 model [51][3], fine-tuned on the MSP-Podcast corpus [32], is used to extract acoustic emotion features. For simplicity, this encoder is referred to as Emotion-Wav2Vec2.0.

As a result, each audio segment is represented as a sequence of embeddings of size $T_a \times 1024$, where T_a denotes the number of temporal steps and 1024 is the feature dimension. The value of T_a depends on the duration of the input video.

The extracted acoustic representations are then processed by a sequential encoder to model temporal dependencies in the speech signal. In the main configuration, the extracted acoustic representations are processed by a Mamba encoder [15], followed by mean pooling over the

[2] https://github.com/lindevs/yolov8-face

[3] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

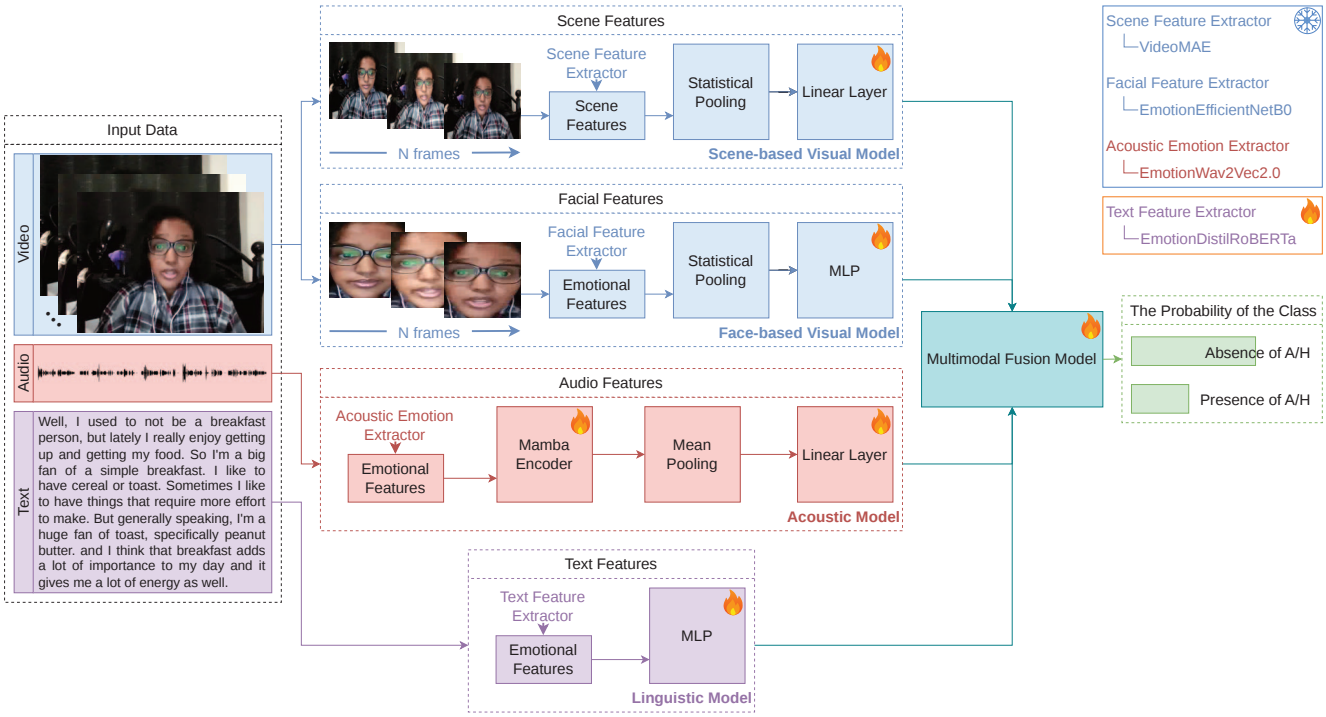

Figure 1. Pipeline of the proposed multimodal approach. The approach integrates four models, namely scene-based visual, face-based visual, acoustic, and linguistic. These models extract complementary modality-specific features from video, audio, and text. In the visual streams, N uniformly sampled frames are used as input. The extracted features are combined using the proposed multimodal fusion model to predict the probability of absence or presence of A/H. Snowflake icons indicate frozen pretrained models, whereas flame icons indicate trainable or fine-tuned modules.

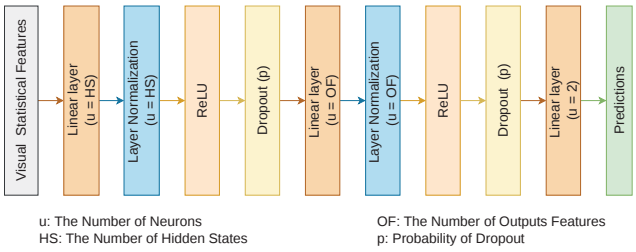

Figure 2. Architecture of a MLP for analysis of visual statistical feature. The network comprises three linear layers interleaved with layer normalization, ReLU activations, and dropout. The final layer produces the probability of absence or presence of A/H.

temporal dimension to obtain a compact acoustic embedding:

$$a = \frac{1}{T_a} \sum_{t=1}^{T_a} s_t, \tag{3}$$

where s_t denotes the representation at the t-th temporal step and a is the pooled acoustic embedding. The resulting embedding is then passed through a linear layer to produce the final prediction. In addition to the Mamba-based [15] variant, a Transformer-based [50] encoder was also evaluated as an alternative sequential architecture.

3.4. Linguistic Model

For the linguistic modality, all audio recordings are transcribed into text. These transcriptions were automatically extracted by the corpus authors [14].

Several text-based modeling strategies are considered. A classical Term Frequency-Inverse Document Frequency (TF-IDF) [43] representation is first used to encode the relative importance of words and phrases in each transcription. The resulting features are combined with conventional classifiers, including Logistic Regression and Gradient Boosting Machines (GBM) models such as LightGBM [22] and CatBoost [35].

Pre-trained transformer-based language models are also used to obtain contextualized text representations. EmotionDistilRoBERTa[4], allMiniLM[5], EmotionTextClassifier[6], and DeBERTa [18] are used to encode the transcriptions

[4] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/

[5] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[6] https://huggingface.co/michelleli99/emotion_text_classifier

into dense embeddings. Both the [CLS] token representation and mean pooling over token embeddings are explored, followed by a GBM-based classifier.

In the main configuration, EmotionDistilRoBERTa is directly fine-tuned for A/H recognition. Its output representation is then fed into an MLP-based classification head to produce the final prediction. In addition, other fine-tuned transformer variants, including EmotionTextClassifier, are evaluated for comparison.

3.5. Modality Fusion Model

A two-stage strategy is adopted. First, unimodal models for scene, face, audio, and text are trained independently on the target corpus. Then, for each video, one fixed-dimensional embedding per modality is extracted from the corresponding unimodal branch and used as input to the fusion model.

Let M be the number of modalities, $m \in \{1, \ldots, M\}$ the modality index, and $x_m \in \mathbb{R}^{d_m}$ the input embedding of modality m, where d_m denotes its original dimensionality. Each modality is projected into a shared latent space $\mathbb{R}^d$:

$$u_m = \phi_m(x_m), \qquad u_m \in \mathbb{R}^d, \tag{4}$$

where ϕ_m is a modality-specific projector composed of a linear layer, layer normalization, Gaussian Error Linear Unit (GELU), and dropout. The resulting modality tokens are stacked into a matrix:

$$U = [u_1; \ldots; u_M] \in \mathbb{R}^{M \times d}. \tag{5}$$

If one or more modality embeddings are unavailable for a given sample, a binary modality mask is provided to the fusion encoder so that the corresponding tokens are masked out during self-attention. A learnable modality embedding matrix $E_{\text{mod}} \in \mathbb{R}^{M \times d}$ is added to the token sequence:

$$Z^{(0)} = U + E_{\text{mod}}, \tag{6}$$

where $Z^{(0)}$ is the input token sequence to the Transformer encoder.

The token sequence is processed by a stack of Transformer encoder layers:

$$Z^{(l+1)} = T^{(l)}(Z^{(l)}), \qquad l = 0, \ldots, L-1, \tag{7}$$

where $T^{(l)}(\cdot)$ denotes the l-th Transformer encoder block and L is the number of layers. The final fused representation is obtained by masked mean pooling over the output modality tokens:

$$z_{\text{fused}} = \frac{\sum_{m=1}^{M} \mu_m z_m^{(L)}}{\sum_{m=1}^{M} \mu_m}, \tag{8}$$

where $z_m^{(L)}$ denotes the output representation of modality m at the last encoder layer and $\mu_m \in \{0, 1\}$ indicates whether modality m is available for the given sample.

For the prototype-augmented variant, the fused representation is compared with a set of learnable class-specific prototypes. Let $P_c = \{p_{c,k}\}_{k=1}^{K}$ be the prototype set for class c, where K is the number of prototypes per class. These prototypes are randomly initialized at the beginning of training and optimized jointly with the remaining model parameters. Therefore, they are not constructed from precomputed class centroids or batch-level statistics, but are learned end-to-end from the training data. This allows the prototypes to directly adapt to the target multimodal classification task. The similarity score for class c is computed as:

$$\hat{y}_c^{\text{proto}} = \log \sum_{k=1}^{K} \exp\left(\frac{\tilde{z}_{\text{fused}}^{\top} \tilde{p}_{c,k}}{\tau}\right), \tag{9}$$

where $\tilde{z}_{\text{fused}}$ and $\tilde{p}_{c,k}$ denote the ℓ_2-normalized fused representation and class prototypes, respectively, and τ is a temperature parameter. In the implemented model, the prototype head does not directly produce the final multimodal prediction; instead, it contributes an auxiliary loss term during training, while the final output logits are produced by the main linear classifier. This design allows the prototype head to provide an additional supervision signal that helps structure the fused representation in the shared latent space during training, while keeping the final prediction defined by the main classifier.

Accordingly, the overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{proto}} \mathcal{L}_{\text{proto}} \tag{10}$$

where $\mathcal{L}_{\text{cls}}$ is the main classification loss computed from the output logits of the linear classifier, and $\mathcal{L}_{\text{proto}}$ is the auxiliary classification loss computed from the prototype-based logits.

4. Experiments

4.1. Research Corpus

The BAH corpus is the core research corpus for the 10th ABAW A/H challenge. It was collected to support multimodal recognition of ambivalence and hesitancy in videos recorded in a realistic digital behavior change scenario [14]. Participants answered a predefined set of questions designed to elicit neutral, positive, negative, willing, resistant, ambivalent, and hesitant responses. The data were collected through an online platform with an avatar-guided interaction setup in order to mimic real-world personalized behavioral interventions [14].

The full corpus contains 1,427 videos from 300 participants, totaling 10.60 hours of recordings. It includes video-level and frame-level expert annotations, timestamps of A/H segments, cropped and aligned faces, speech transcripts with timestamps, and participant metadata [14].

Since ambivalence and hesitancy are difficult to separate reliably in practice, the task is formulated as a binary classification problem indicating the presence or absence of A/H [14]. For the 10th ABAW A/H challenge, the corpus is partitioned participant-wise into training, validation, public test, and private test subsets. Video-level prediction is evaluated using Macro F1-score (MF1) as the main metric.

4.2. Experimental Setup

During the development of the unimodal models, several alternative configurations were evaluated. Scene modeling was performed using frame sequences with $N = 16$ resized to 224×224, trained for 15 epochs with AdamW, a Learning Rate (LR) of $2e$-5, weight decay of $1e$-2, batch size 4, cosine annealing, and Label Smoothing (LS) of 0.1. For the face-based visual modality, both statistical features with an MLP and raw embeddings extracted by ViT [10] and Contrastive Language-Image Pretraining (CLIP) [36] in combination with Transformer [50] and Mamba [15] encoders were considered. The best configuration was based on statistical features with an MLP. A grid search over the number of frames, hidden states, output features, LR, and optimizer selected a setup with $N = 30$ frames, 16 hidden states, 256 output features, a LR of $1e$-3, and AdamW.

For the acoustic modality, embeddings from different EmotionWav2Vec2.0 layers and different temporal encoders, including Mamba and Transformer, were evaluated. The best result was obtained with layer 10 and Mamba. In all primary runs, AdamW was used. The selected acoustic setup employed hidden size 256, feed-forward size 512, dropout 0.1, mean pooling, Mamba state size 8, convolution kernel size 4, and expansion factor 2. In the linguistic modality, TF-IDF features were evaluated with vocabulary sizes from 100 to 10,000 and n-grams from unigrams to trigrams, with hyperparameter tuning for Logistic Regression and GBM models. Transformer-based text models were also fine-tuned using partially frozen backbones and jointly trained MLP heads. The number of hidden layers varied from 1 to 3, the hidden size from 64 to 128, and dropout from 0 to 0.3. AdamW and Stochastic Gradient Descent (SGD) were explored, the LR was searched in the range from $1e$-5 to 0.1, the batch size was set to 16, and training lasted from 3 to 20 epochs with early stopping.

Multimodal fusion was performed using embeddings extracted from the selected scene, face, audio, and text models. Both non-prototype and prototype-augmented variants were evaluated, and the final system was based on the prototype-augmented fusion model operating on embedding-level inputs. To reduce sensitivity to random initialization, model selection relied on a stability-oriented hyperparameter search with Optuna [1] using five fixed random seeds: 42, 2025, 7777, 12345, and 31415. Each candidate configuration was trained and evaluated five times, and the final score was computed as the average MF1 across these runs. The selected fusion model used a shared latent dimensionality of 128, 6 Transformer encoder layers, 4 attention heads, a feed-forward expansion factor of 6, no `[CLS]` token, and dropout of 0.45. The prototype head used 16 learnable prototypes per class with temperature $\tau = 0.3$. Training was performed with RMSprop, a LR of $9.44e$-5, weight decay of $5.55e$-4, LS of 0.02, gradient clipping of 0.5, and a cosine learning-rate scheduler. The prototype loss weight was set to $\lambda_{\text{proto}} = 0.2$. Final predictions were obtained by averaging the class probabilities of the 5 seed-specific models. All models were trained on an NVIDIA A100 Tensor Core Graphics Processing Unit (GPU) with 80 GB of memory [25].

4.3. Experimental Results

The experimental results are presented in Table 1. Among the unimodal models, the best average MF1 was obtained by the fine-tuned EmotionDistilRoBERTa model (70.02%), followed closely by the fine-tuned EmotionTextClassifier (70.00%) and the acoustic model based on EmotionWav2Vec2.0 and Mamba (69.03%). The TF-IDF-based text models also showed competitive results, reaching 68.03% with Logistic Regression and 68.79% with CatBoost. In contrast, the face- and scene-based models achieved lower average scores of 62.67% and 61.96%, respectively.

All multimodal fusion models outperformed the unimodal baselines. Among the single fusion models, the best average result was achieved by the prototype-augmented fusion model based on the selected scene, face, audio, and text modalities (Model ID 12), with 83.25%, while the corresponding fusion model without the prototype head (Model ID 11) reached 82.66%. These results indicate that both multimodal integration and prototype-based classification are beneficial under the development and public test settings.

The final test results show a different trend. Among the submitted four-modality systems, the best performance was obtained by the ensemble of five prototype-augmented fusion models (Model ID 14), which achieved 71.42%. The ensemble of five fusion models without the prototype head (Model ID 13) reached 70.17%, while the single fusion models achieved 68.32% (Model ID 11) and 65.21% (Model ID 12). Thus, although the prototype-augmented single model achieved the highest average result, ensembling was essential for the strongest final test performance and improved robustness on the private evaluation split.

The ablation study of the best single model (Model ID 11) further confirms the benefits of combining modalities. The strongest two-modality result was obtained by combining scene and text features (Model ID 19), with an average MF1 of 80.39%. This result suggests that scene-based nonverbal features are highly informative for A/H recog-

ID	Model Configuration			BAH sub-corpus			
	Modality	Features	Classifier	Devel. (MF1, %)	Test (MF1, %)	Average (MF1, %)	Private test (MF1, %)
1	Face	EmotionEfficientNetB0 + Statistical Features	MLP	65.29	60.05	62.67	–
2	Scene	VideoMAE	Linear Layer	61.71	62.21	61.96	–
3	Audio	EmotionWav2Vec2.0 + Mamba	Linear Layer	67.20	70.87	69.03	–
4	Text	TF-IDF	Logistic Regression	68.30	67.75	68.03	–
5	Text	TF-IDF	CatBoost	65.56	72.02	68.79	–
6	Text	Fine-tuned EmotionTextClassifier	MLP	69.28	70.72	70.00	–
7	Text	Fine-tuned EmotionDistilRoBERTa	MLP	68.54	71.49	70.02	–
8	Models IDs 1, 2, 3 and 4	MFM	Linear Layer	80.79	77.03	78.91	–
9	Models IDs 1, 2, 3 and 5	MFM	Linear Layer	77.91	78.54	78.22	–
10	Models IDs 1, 2, 3 and 6	MFM	Linear Layer	78.35	77.03	77.69	–
11	Models IDs 1, 2, 3 and 7	MFM	Linear Layer	85.38	79.94	82.66	68.32
12	Models IDs 1, 2, 3 and 7	MFM with Prototype Head	Linear Layer	83.79	82.72	83.25	65.21
13	Models IDs 1, 2, 3 and 7	Ensemble of Five MFMs	Linear Layer	81.94	80.64	81.29	70.17
14	Models IDs 1, 2, 3 and 7	Ensemble of Five MFMs with Prototype Head	Linear Layer	83.00	80.77	81.89	71.42
			Ablation Study				
15	Models IDs 1 and 2	MFM	Linear Layer	78.07	77.09	77.58	–
16	Models IDs 1 and 3	MFM	Linear Layer	63.36	71.44	67.40	–
17	Models IDs 1 and 7	MFM	Linear Layer	65.29	61.19	63.24	–
18	Models IDs 2 and 3	MFM	Linear Layer	77.37	77.66	77.51	–
19	Models IDs 2 and 7	MFM	Linear Layer	81.77	79.00	80.39	–
20	Models IDs 3 and 7	MFM	Linear Layer	67.05	70.99	69.02	–
21	Models IDs 1, 2 and 3	MFM	Linear Layer	76.10	79.15	77.62	–
22	Models IDs 1, 2 and 7	MFM	Linear Layer	79.89	77.65	78.77	–
23	Models IDs 1, 3 and 7	MFM	Linear Layer	68.08	70.41	69.25	–
24	Models IDs 2, 3 and 7	MFM	Linear Layer	79.89	77.63	78.76	–

Table 1. Experimental results on the BAH corpus for video-level A/H recognition. MFM refers to Multimodal Fusion Model.

nition and complement verbal information from text at a similarly important level. This indicates that both visible behavioral manifestations and linguistic content are critical for A/H recognition. Among the three-modality settings, the best performance was achieved by two different models with combining modalities such as: (1) face, scene, and text (Model ID 22), with 78.77% and (2) scene, audio, and text (Model ID 24), with 78.76%. Overall, the best results were obtained with four-modality fusion.

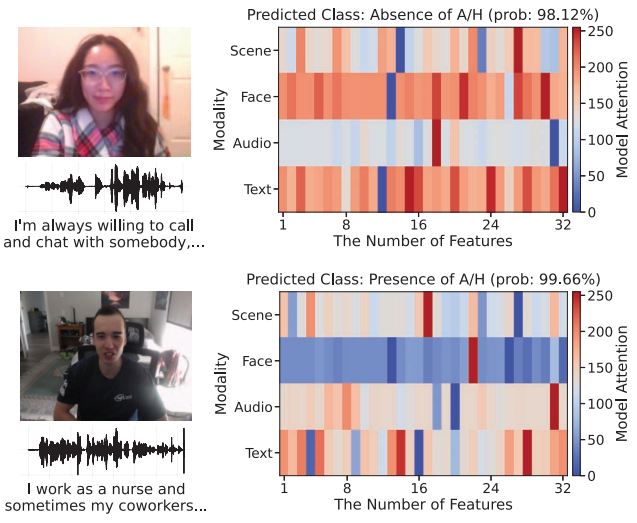

Figure 3. Visualization of the attention of the best single model (Model ID 11).

Figure 3 presents two examples of the proposed multi-modal model predictions (Model ID 11) together with the corresponding modality-wise attention patterns. The first example corresponds to the absence of A/H, where the model assigns a high confidence score of 98.12% and relies primarily on facial and textual features. The second example corresponds to the presence of A/H, where the model is also highly confident (99.66%) but shifts its attention toward audio and text modalities. These examples show that the contribution of each modality varies across samples and depends on the affective content of the input. Importantly, unlike unimodal models, the proposed multimodal model can adaptively combine complementary evidence from several modalities and reweight their importance for each sample, which is especially important for A/H recognition, where relevant cues may appear inconsistently across face, speech, language, and scene context.

Table 2. Performance comparison (MF1) of the SOTA approaches on the private test subset. A, V, T and S are audio, video, text and scene modalities, respectively.

Approach	Modalities	MF1
Pereira et al. [33]	AVT	72.66
Bekhouche et al. [3]	AVT	71.51
Ours	AVST	71.42
Tang et al. [47]	M-LLM	67.48
Baseline	M-LLM	34.28

Table 2 demonstrates that the proposed approach achieves a competitive performance on the private test set,

with an MF1 of 71.42%. This result is slightly lower than the best-performing approach by Pereira et al. [33], which achieved 72.66%, but it is close to the performance of Bekhouche et al. [3], this approach also achieved 71.51%.

5. Conclusions

This paper presented a multimodal approach for video-level A/H recognition on the BAH corpus. The proposed approach combined scene, face, audio, and text modalities and consistently outperformed unimodal baselines. Among the unimodal models, the best average MF1 of 70.02% was achieved by the fine-tuned EmotionDistilRoBERTa model, confirming the strong contribution of the linguistic modality. At the multimodal level, the best average result of 83.25% was obtained by the prototype-augmented fusion model, while the best final test performance of 71.43% was achieved by the ensemble of five prototype-augmented fusion models.

The ablation study showed that scene and text provide the strongest complementary signal among modality pairs, and that extending the fusion scheme to all four modalities yields the most effective overall solution. The comparison between single fusion models and their ensembles further indicates that robust model aggregation is important for generalization on the private test split. Overall, these results demonstrate that prototype-augmented multimodal fusion is an effective and robust strategy for A/H recognition in unconstrained videos.

6. Acknowledgements

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4E0002 and the agreement with HSE University № 139-15-2025-009, with partial support from the computational resources of HPC facilities at HSE University.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *KDD*, pages 2623–2631, 2019. 6

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, pages 12449–12460, 2020. 2

[3] Salah Eddine Bekhouche, Hichem Telli, Azeddine Benlamoudi, Salah Eddine Herrouz, Abdelmalik Taleb-Ahmed, and Abdenour Hadid. Conflict-aware multimodal fusion for ambivalence and hesitancy recognition. *arXiv preprint arXiv:2603.15818*, 2026. 2, 7, 8

[4] Laura E Bijkerk, Mark Spigt, Anke Oenema, and Nicole Geschwind. Engagement with mental health and health behavior change interventions: an integrative review of key concepts. *J. Context. Behav. Sci.*, 32:1–14, 2024. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[6] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, 2020. 2

[7] Jiawen Deng and Fuji Ren. A survey of textual emotion recognition and its challenges. *IEEE Trans. Affect. Comput.*, 14(1):49–67, 2023. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, pages 1–22, 2021. 3, 6

[11] Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *CVPR*, pages 14314–14324, 2025. 2

[12] Ali Pourramezan Fard, Mohammad Mehdi Hosseini, Timothy D Sweeny, and Mohammad H Mahoor. Affectnet+: A database for enhancing facial expression recognition with soft-labels. *IEEE Trans. Affect. Comput.*, 17(1):784–800, 2026. 3

[13] Swapna Mol George and P Muhamed Ilyas. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 568: 1–23, 2024. 1

[14] Manuela González-González, Soufiane Belharbi, Muhammad Osama Zeeshan, Masoumeh Sharafi, Muhammad Haseeb Aslam, Marco Pedersoli, Alessandro Lameiras Koerich, Simon L Bacon, and Eric Granger. Bah dataset for ambivalence/hesitancy recognition in videos for digital behavioural change. In *ICLR*, pages 1–17, 2026. 1, 2, 4, 5, 6

[15] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *CoLM*, pages 1–16, 2024. 3, 4, 6

[16] Xiang Guo, Linyu Dong, Wenyi Zhu, Yuzhong Zhang, and Yang Liu. Light-rank: a multimodal neural network for recognizing the top personality traits in asynchronous video interviews. *Signal Image Video Process.*, 20(2):1–13, 2026. 2

[17] Tobias Hallmen, Robin-Nico Kampa, Fabian Deuser, Norbert Oswald, and Elisabeth André. Semantic matters: Mul-

timodal features for affective analysis. In *CVPRW*, pages 5724–5733, 2025. 2

[18] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*, pages 1–21, 2021. 4

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 2

[20] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 2

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, pages 1–22, 2017. 3

[22] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *NeurIPS*, pages 1–9, 2017. 4

[23] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Eric Granger, Marco Pedersoli, Simon Bacon, Alice Baird, Chris Gagne, et al. Advancements in affective and behavior analysis: The 8th abaw workshop and competition. In *CVPRW*, pages 5572–5583, 2025. 2

[24] Dimitrios Kollias, Stefanos Zafeiriou, Irene Kotsia, Greg Slabaugh, Damith Chamalke Senadeera, Jianian Zheng, Kaushal Kumar Keshlal Yadav, Chunchang Shao, and Guanyu Hu. From emotions to violence: Multimodal fine-grained behavior analysis at the 9th abaw. In *ICCVW*, pages 1–12, 2025. 2

[25] Pavel Kostenetskiy, Roman Chulkevich, and Viachesvlav Kozyrev. Hpc resources of the higher school of economics. *Journal of Physics: Conference Series*, 1740(1):012050, 2021. 6

[26] Sze Chit Leong, Yuk Ming Tang, Chung Hin Lai, and Carman Ka Man Lee. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *Comput. Sci. Rev.*, 48:1–13, 2023. 1

[27] Chenxin Li, Yifan Liu, Panwang Pan, Hengyu Liu, Xinyu Liu, Wuyang Li, Cheng Wang, Weihao Yu, Yiyang Lin, and Yixuan Yuan. Infobridge: Balanced multimodal integration through conditional dependency modeling. In *ICCV*, pages 393–404, 2025. 2

[28] Yongwei Li, Wei Gao, and Jianwu Li. Multimodal emotion recognition via unified granularity contrastive learning and similar negative discrimination. *Pattern Recognit.*, pages 1–9, 2026. 2

[29] Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–18, 2026. 2

[30] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024. 2

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[32] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. Affect. Comput.*, 10(4):471–483, 2017. 3

[33] Alexandre Pereira, Bruno Fernandes, and Pablo Barros. Brother: Behavioral recognition optimized through heterogeneous ensemble regularization for ambivalence and hesitancy. *arXiv preprint arXiv:2603.14361*, 2026. 2, 7, 8

[34] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion*, 37:98–125, 2017. 1, 2

[35] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *NeurIPS*, pages 1–11, 2018. 4

[36] 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. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 6

[37] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. 2

[38] Elena Ryumina, Alexandr Axyonov, Timur Abdulkadirov, Darya Koryakovskaya, and Dmitry Ryumin. Cross-lingual bimodal emotion recognition with llm-based label smoothing. *Big Data Cogn. Comput.*, 9(11):1–38, 2025. 2

[39] Elena Ryumina, Maxim Markitantov, Alexandr Axyonov, Dmitry Ryumin, Mikhail Dolgushin, and Alexey Karpov. Zero-shot multimodal compound expression recognition approach using off-the-shelf large visual-language models. In *ICCVW*, pages 71–79, 2025.

[40] Elena Ryumina, Dmitry Ryumin, Alexandr Axyonov, Denis Ivanko, and Alexey Karpov. Multi-corpus emotion recognition method based on cross-modal gated attention fusion. *Pattern Recognit. Lett.*, 190:192–200, 2025.

[41] Elena Ryumina, Alexandr Axyonov, Darya Koryakovskaya, Timur Abdulkadirov, Angelina Egorova, Sergey Fedchin, Alexander Zaburdaev, and Dmitry Ryumin. Ssl-mepr: A semi-supervised multi-task cross-domain learning framework for multimodal emotion and personality recognition. *Mach. Learn. Knowl. Extr.*, 8(3):1–41, 2026. 2

[42] Muhammad Sajjad, Fath U Min Ullah, Mohib Ullah, Georgia Christodoulou, Faouzi Alaya Cheikh, Mohammad Hijji, Khan Muhammad, and Joel JPC Rodrigues. A comprehensive survey on deep facial expression recognition: chal-

lenges, applications, and future guidelines. *Alex. Eng. J.*, 68: 817–840, 2023. 1

[43] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975. 4

[44] Andrey Savchenko and Lyudmila Savchenko. Leveraging lightweight facial models and textual modality in audio-visual emotional understanding in-the-wild. In *CVPRW*, pages 5787–5797, 2025. 2

[45] Andrey V Savchenko and Anna P Sidorova. Emotieffnet and temporal convolutional networks in video-based facial expression recognition and action unit detection. In *CVPRW*, pages 4849–4859, 2024. 2

[46] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 3

[47] Liang Tang, Hongda Li, Jiayu Zhang, Long Chen, Shuxian Li, Siqi Pei, Tiaonan Duan, and Yuhao Cheng. Nuanced emotion recognition based on a segment-based mllm framework leveraging qwen3-omni for ah detection, 2026. 2, 7

[48] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, pages 10078–10093, 2022. 2, 3

[49] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 1–11, 2017. 4, 6

[51] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 10745–10759, 2023. 3

[52] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 2

[53] Qu Yang, Qinghongya Shi, Tongxin Wang, and Mang Ye. Uncertain multimodal intention and emotion understanding in the wild. In *CVPR*, pages 24700–24709, 2025. 2

[54] Bengong Yu, Jun Wang, Chenyue Li, Zhonghao Xi, Xianxian Zhao, and Yue Li. Graph-prototype distillation with prototype-guided contrastive training for multimodal emotion recognition in conversations. *Knowl.-Based Syst.*, pages 1–15, 2026. 2

[55] Ziyin Zhang, Zihan Liao, Hang Yu, Peng Di, and Rui Wang. F2llm technical report: Matching sota embedding performance with 6 million open-source data. *arXiv preprint arXiv:2510.02294*, 2025. 2