Team RAS in 11th ABAW Competition: Multimodal Ambivalence Recognition Approach

Elena Ryumina[1], Maxim Markitantov[1], Alexandr Axyonov[1], Fedor Shchetinin[2], Timur Abdulkadirov[1], Dmitry Ryumin[1], and Alexey Karpov[1,3]

[1] St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia
{ryumina.e,markitantov.m,axyonov.a,ryumin.d,karpov}@iias.spb.su
timur.abdulkadirov4@gmail.com
[2] HSE University, St. Petersburg, Russia
fshchetinin@hse.ru
[3] ITMO University, St. Petersburg, Russia

Abstract. Automatic recognition of ambivalence and hesitancy is challenging because these states may be expressed through inconsistent linguistic, acoustic, facial, and contextual patterns, while top-performing systems often rely on computationally expensive ensembles. We present a single text-centered multimodal approach for video-level ambivalence and hesitancy recognition for the 11th Affective & Behavior Analysis in-the-Wild (ABAW) Challenge. The proposed approach combines linguistic, acoustic, facial, and scene features using text-centered multimodal fusion model. Text Residual Fusion treats text as the anchor modality and applies gated residual adjustments based on the other modalities. Experiments on the Behavioural Ambivalence/Hesitancy (BAH) corpus confirm that text is the strongest unimodal modality. The Text Residual Fusion model achieves an average Macro F1-score (MF1) of 75.14% across the Development and Public Test subsets. On the Private Test subset, it reaches an MF1 of 78.24%, outperforming the text model by 4.03%. These results demonstrate that complementary multimodal information can improve recognition performance without requiring a large model ensemble.

Keywords: Ambivalence recognition · Multimodal features· Text-centered multimodal fusion model.

1 Introduction

Ambivalence and hesitancy are complex affective states that may be expressed through inconsistent linguistic, acoustic, facial, and contextual information. Their automatic recognition is therefore relevant to affective computing, which studies human affect from face [25], audio [9], text [6], and bodily modalities [17], with applications in human–computer interaction, healthcare, education, and assistive systems [21]. The Ambivalence / Hesitancy (AH) Video Recognition task of the

11th Affective & Behavior Analysis in-the-Wild (ABAW) Challenge addresses this problem at the video-level using the Behavioural Ambivalence/Hesitancy (BAH) corpus [10]. In digital behavior-change scenarios, such states may reflect uncertainty, resistance, unstable motivation, or potential disengagement [4].

Previous studies have shown that transcripts provide the strongest unimodal signal for this task, whereas facial and acoustic information can further improve performance when fused effectively [10, 12, 26]. This finding motivates an asymmetric fusion strategy in which text serves as the primary semantic representation, while audio, face, and scene modalities provide complementary information. This design is consistent with recent studies on modality-aware and uncertainty-aware multimodal fusion [8], as well as with broader advances in multimodal affective computing [15, 16].

The 10th ABAW Challenge also highlighted a practical limitation of current top-performing systems [3,20,23]: the three highest-ranked solutions used ensembles or multiple prediction branches, and the winning method combined approximately twenty models [20]. Although such systems may achieve high benchmark performance, they require substantially greater computational resources, inference time, memory, and implementation effort. In this work, we propose a single text-centered multimodal approach that combines text, audio, face, and scene modalities without relying on a large ensemble. The proposed approach aims to preserve complementary multimodal information while offering a more compact and practically deployable solution for video-level AH recognition.

2 Related Work

2.1 Ambivalence and Hesitancy Recognition

The BAH benchmark introduced in [10] established the main experimental setting for video-level AH recognition. The authors evaluated unimodal and multimodal systems using face, audio, and text modalities. The examined fusion methods included feature concatenation, co-attention, Transformer-based fusion, and cross-attention. The results showed that text was the strongest individual modality. The official baseline also included a zero-shot Multimodal LargeLanguage Model (M-LLM) based on Video-LLaVA [18].

Submissions to the challenge confirmed this finding. Hallmen et al. [12] combined visual features extracted using a Vision Transformer (ViT) [5], acoustic features obtained using Wav2Vec2 [2], and textual features produced by Bidirectional Encoder Representations from Transformers (BERT) [7]. They used Long Short-Term Memorys (LSTMs) [13] and an Multilayer Perceptron (MLP) [22] for temporal modeling and multimodal fusion. Savchenko et al. [26] extracted facial, acoustic, and textual features with EmotiEffLib [27], Wav2Vec2 [2], and RoBERTa [19], respectively, and evaluated both early and late fusion strategies. Both studies identified text as the strongest unimodal source, while improvements from multimodal fusion depended on the effective preservation of complementary information [8].

2.2 Top Systems of 10th ABAW Challenge

The three highest-ranked solutions in the AH task of the 10th ABAW Challenge relied on model ensembles. The third-ranked system [23] combined scene, facial, acoustic, and textual features. Its final results on the Private Test subset was obtained using an ensemble of five prototype-augmented multimodal fusion models. The second-ranked ConflictAwareAH system [3] combined video, audio, and text modalities with pair-wise cross-modal conflict features and text-guided late fusion. Its final submission also used an ensemble of five models. The winning BROTHER system [20] employed a considerably larger ensemble of approximately twenty heterogeneous models and combined their predictions using a voting strategy.

Although these results demonstrate the effectiveness of model ensembles, the use of five to twenty models substantially increases computational cost, inference latency, memory requirements, and implementation complexity. These factors complicate the practical deployment and reproduction of such systems and motivate our focus on a single text-centered multimodal fusion model.

3 Proposed Approach

The pipeline of the proposed approach is shown in Figure 1. The details are described below.

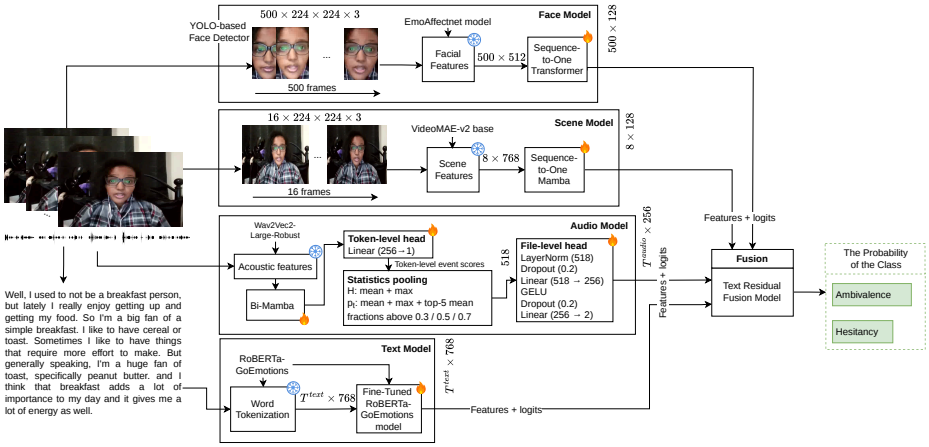

Fig. 1: Pipeline of the Text Residual Fusion model.

3.1 Text Model

Textual features are extracted using a RoBERTa-base model pre-trained on the GoEmotions dataset[4]. During training, the embedding layer and the first four Transformer encoder layers are frozen.

Let an input text sequence of length T^{text} be represented by a matrix of token embeddings:

$$X^{text} = [x_1^{text}, x_2^{text}, \ldots, x_{T^{text}}^{text}] \in \mathbb{R}^{T^{text} \times 768}. \tag{1}$$

Each token x_t^{text} is mapped to a contextual representation by the encoder layers, producing the hidden-state matrix:

$$H^{text} = \text{Encoder}(X^{text}) = [h_1^{text}, h_2^{text}, \ldots, h_{T^{text}}^{text}] \in \mathbb{R}^{T^{text} \times 768} \tag{2}$$

The pooled representation corresponding to the classification token is used for the final prediction. This sentence-level representation is passed to an MLP classification head consisting of a dropout layer, a Fully-Connected Layer (FCL), a tanh activation function, a second dropout layer, and a final FCL with two output units.

3.2 Audio Model

Before training the audio model, each video is converted into a waveform sampled at 16 kHz and processed as a full audio, without voice-activity filtering. The acoustic features are extracted from the tenth Transformer layer of a frozen Wav2Vec2 model [2][5]. The resulting audio sequence is represented as

$$X^{\text{audio}} = [x_1^{\text{audio}}, x_2^{\text{audio}}, \ldots, x_{T^{\text{audio}}}^{\text{audio}}] \in \mathbb{R}^{T^{\text{audio}} \times 1024}, \tag{3}$$

where x_t^{audio} denotes the acoustic embedding corresponding to the t-th temporal interval. The extracted features are first projected into a compact hidden space and then processed by a bi-directional Mamba [11] block, followed by dropout and a residual connection. The resulting temporal acoustic features are denoted by $H^{\text{audio}} \in \mathbb{R}^{T^{\text{audio}} \times 256}$.

Because AH behavior may occur only during short intervals of a recording, frame-level annotations are used as auxiliary temporal supervision to help the model localize such events. These annotations are converted into token-level soft targets. For the set $\mathcal{F}_t$ of video frames temporally aligned with the t-th audio token, the target is defined as

$$\widetilde{y}_t = \frac{1}{|\mathcal{F}_t|} \sum_{j \in \mathcal{F}_t} y_j. \tag{4}$$

[4] https://huggingface.co/SamLowe/roberta-base-go_emotions
[5] https://huggingface.co/facebook/wav2vec2-large-robust

A token-level head consisting of a single FCL produces a token-level logit z_t and the corresponding probability $p_t = \sigma(z_t)$. For file-level classification, masked mean and max pooling over H^{audio} are concatenated with the mean, maximum, top-K mean, and threshold statistics of the token-level probabilities, producing a 518-dimensional descriptor. This descriptor is processed by an MLP consisting of two FCLs. Layer Normalization (LN) and dropout are applied before the first layer, while a GELU activation function and dropout are applied between the two FCLs. The final FCL contains two output units. The audio model is trained using a combination of file-level and token-level supervision:

$$\mathcal{L} = \mathcal{L}_{\text{file-level}} + 0.5\mathcal{L}_{\text{token-level}} + 0.02\mathcal{L}_{\text{smooth}} + 0.1\mathcal{L}_{\text{event}}. \tag{5}$$

where $\mathcal{L}_{\text{file-level}}$ denotes the main cross-entropy loss for the file-level AH label. The auxiliary term $\mathcal{L}_{\text{token-level}}$ is a token-level binary cross-entropy loss whose soft targets are obtained from the 24 fps frame-level annotations. The smoothness loss penalizes isolated peaks in the token-level probabilities, while the event loss applies binary cross-entropy to the maximum token logit, encouraging the model to detect short AH intervals within long recordings. The auxiliary-loss weights were selected based on performance of the Development subset.

3.3 Face Model

Face regions are detected using a YOLO-based face detector[6]. The detected regions are resized to 224×224 pixels and processed by the frame-level facial emotion encoder Emo-AffectNet [24], which produces a 512-dimensional feature vector for each face. A video containing T^{face} retained face regions is represented as

$$X^{\text{face}} = [x_1^{\text{face}}, x_2^{\text{face}}, \ldots, x_{T^{\text{face}}}^{\text{face}}] \in \mathbb{R}^{T^{\text{face}} \times 512}, \tag{6}$$

where x_t^{face} denotes the frame-level embedding. A maximum of $T^{\text{face}} = 500$ frames is used. Longer sequences are uniformly subsampled, while shorter sequences are remained unchanged.

The temporal dynamics of facial AH expressions are modeled with a Transformer-based encoder [28]. The Emo-AffectNet embeddings are projected into a 128-dimensional hidden space, combined with positional encodings, and processed by a single Transformer layer with eight attention heads. The temporally pooled representation is then passed to an MLP classification head consisting of a FCL, LN, a GELU activation function, dropout, and a final FCL with two units. The output of the penultimate FCL is used as the facial classification features, and the output of the final layer represents the facial logits.

To regularize the model, we apply flow matching [14] in both the temporal space and the logit spaces. Let Z denote the temporal features produced by the Transformer-based encoder from the input sequence X^{face}. For feature-space flow matching, a noise tensor Z_0 is interpolated with Z using $t \sim \mathcal{U}(0, 1)$:

$$Z_t = (1-t)Z_0 + tZ, \mathcal{L}_{\text{FM}}^{\text{feat}} = \|f_\theta(Z_t, Z, t) - (Z - Z_0)\|_2^2. \tag{7}$$

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

For logit-space flow matching, the initial logits $\hat{y}_0$ are interpolated with the one-hot target vector q:

$$y_t = (1-t)\hat{y}_0 + tq, \mathcal{L}_{\mathrm{FM}}^{\mathrm{logit}} = \|g_\psi(y_t, z, t) - (q - \hat{y}_0)\|_2^2 . \tag{8}$$

The final training loss is defined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{FM}} \left(\mathcal{L}_{\mathrm{FM}}^{\mathrm{feat}} + \mathcal{L}_{\mathrm{FM}}^{\mathrm{logit}} \right), \tag{9}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the class-weighted cross-entropy loss for two-class classification. Flow matching is performed using four integration steps, with $\lambda_{\mathrm{FM}} = 0.1$.

3.4 Scene Model

Scene information is extracted from the full video frames rather than from cropped face regions. We use a VideoMAE-v2-base visual encoder [29] as the scene feature extractor. Sixteen frames are uniformly sampled from the entire video to preserve information about the background, body posture, and interaction context. For each video, the encoder produces a sequence of embeddings:

$$X^{\mathrm{scene}} = [x_1^{\mathrm{scene}}, x_2^{\mathrm{scene}}, \ldots, x_{T^{\mathrm{scene}}}^{\mathrm{scene}}] \in \mathbb{R}^{T^{\mathrm{scene}} \times 768}, \tag{10}$$

where x_t^{scene} denotes the visual representation of the t-th sampled temporal position.

The extracted sequence of scene features is projected into a latent space and processed by a Mamba-based encoder [11]. In the main configuration, the temporal encoder has a hidden dimension of 128 and consists of two Mamba layers with a state dimension of 64, a one-dimensional convolution kernel of size 3, and an expansion factor of 1. Given the projected sequence $H^{\mathrm{scene}} = [h_1^{\mathrm{scene}}, h_2^{\mathrm{scene}}, \ldots, h_{T^{\mathrm{scene}}}^{\mathrm{scene}}]$, each Mamba layer applies LN, sequence mixing, dropout, and a residual connection:

$$H^{\mathrm{scene}(l+1)} = H^{\mathrm{scene}(l)} + \mathrm{Dropout}\left(\mathrm{Mamba}\left(\mathrm{LN}(H^{\mathrm{scene}(l)})\right)\right). \tag{11}$$

The output sequence is aggregated using masked mean pooling to obtain a single scene representation, which is passed to the same MLP classification head as that used in the face model. The output of the penultimate FCL is used as the scene classification features, while the temporal hidden states is used as the temporal scene features.

3.5 Multimodal Fusion Model

Let $\mathcal{M} \subseteq \{\mathrm{audio}, \mathrm{text}, \mathrm{face}, \mathrm{scene}\}$ denote the set of active modalities. For each modality m, we use a vector representation and, when available, the corresponding unimodal logits ℓ_m. Temporal representations are converted into compact vectors using summary statistics. The best configuration uses:

$$s_m = [\mu_m, \sigma_m, \mu_m^{\Delta}, \sigma_m^{\Delta}], \tag{12}$$

where μ_m and σ_m denote the mean and standard deviation of the temporal features, while μ_m^{Δ} and σ_m^{Δ} denote the mean and standard deviation of their first-order temporal differences. The input to the fusion model is denoted by u_m. For example, when the unimodal logits are concatenated with the temporal statistics, $u_m = [s_m; \ell_m]$.

The **Text Residual Fusion** model uses text as the anchor modality, because it provides the strongest unimodal performance in our experiments. After modality-specific projections, the textual representation $b = h_{\text{text}}$ is adjusted using gated residuals from the remaining modalities. For each non-text modality m, the gate is computed from the concatenated textual and modality-specific representations:

$$g_m = \sigma\left(\text{MLP}[b; h_m]\right), z = \text{LN}\left(b + \sum_{m \neq \text{text}} g_m \, d_m(h_m)\right), \tag{13}$$

where $d_m(\cdot)$ is a residual MLP, g_m is a learned gate, and $\odot$ denotes element-wise multiplication. The gate controls the contribution of modality m to the textual representation. The fused vector z is then passed to a two-layer MLP classifier. This design preserves the text as the primary source of information while allowing the audio, face, and scene modalities to provide input-dependent residual adjustments. The pipeline of the proposed model is shown in Figure 2.

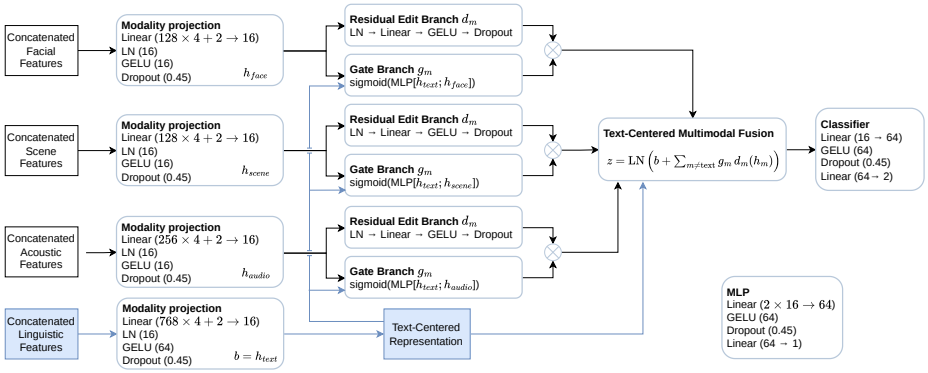

Fig. 2: Pipeline of the Text Residual Fusion model.

4 Experiments

4.1 Research Corpus

The BAH corpus is the benchmark dataset for the AH task of the 11th ABAW Challenge. It was introduced for the multimodal AH recognition in realistic

Table 1: Experimental results (MF1, %) obtained by various configurations of the proposed approach. FM refers to flow matching. LS to label smoothing. TS to temporal smoothing.

ID	Modality	Features	Temporal model	Regularization	Devel subset	Public Test subset	Average Devel/Public Test	Private Test subset
1	Text	RoBERTa-GoEmotions (freeze layers=4)	–	–	**72.55**	**72.10**	**72.33**	74.21
2	Audio	Wav2Vec2 [2]	Bi-Mamba	TS	70.19	69.28	69.74	–
3	Face	EmoAffectNet [24]	Transformer	FM	64.07	61.27	62.67	–
4	Scene	VideoMAE-v2 base [29]	MambaSSM	LS	61.07	61.18	61.12	–
5	Text, Audio, Face, Scene	IDs 1, 2, 3 and 4	Text Residual Fusion	LS	76.13	74.14	75.14	78.24

digital behavior change scenarios [10]. Participants responded to a predefined set of questions intended to elicit neutral, positive, negative, willing, resistant, ambivalent, and hesitant responses during online interactions guided by virtual avatar [10].

The corpus contains 1427 videos from 300 participants, with a total duration of 10.60 hours. It includes video-level and frame-level annotations, temporal boundaries of AH episodes, aligned face regions, timestamped speech transcripts, and participant metadata [10]. Following the challenge protocol, AH are treated jointly as a binary classification task. The dataset is divided at the participant level into the Train, Development, Public Test, and Private Test subsets. Performance is evaluated at the video-level using Macro F1-score (MF1) [10].

4.2 Experimental Results

For all unimodal and multimodal experiments, we used Optuna [1] to optimize the training procedure and model architecture. The final configurations (e.g., learning rate, dropout, hidden dimensions, batch size, and other parameters) were selected on the Development subset and evaluated on the Public Test subset.

As shown in Table 1, text is the strongest unimodal modality, followed by audio. Both fusion models outperform all unimodal configurations on the Development and Public Test subsets. Text Residual Fusion achieves the highest Development MF1 of 76.13%, the highest average MF1 across the Development and the Public Test subsets of 75.14%, and the best Private Test MF1 of 78.24%. On the Private Test subset, it outperforms the text model by 4.03%.

5 Conclusion

This work presented a single text-centered multimodal approach that combines textual, acoustic, facial, and, scene features for video-level AH recognition. The proposed Text Residual Fusion model uses text as the anchor modality and adds complementary information through gated residual adjustments. The proposed fusion model achieves the best performance on the Private Test subset and the most consistent results across the evaluation subsets.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. pp. 2623–2631 (2019). https://doi.org/10.1145/3292500.3330701
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: NeurIPS. vol. 33, pp. 12449–12460 (2020). https://doi.org/10.48550/arXiv.2006.11477
3. Bekhouche, S.E., Telli, H., Benlamoudi, A., Herrouz, S.E., Taleb-Ahmed, A., Hadid, A.: Conflict-aware multimodal fusion for ambivalence and hesitancy recognition. arXiv (2026). https://doi.org/10.48550/arXiv.2603.15818
4. Bijkerk, L.E., Spigt, M., Oenema, A., Geschwind, N.: Engagement with mental health and health behavior change interventions: An integrative review of key concepts. J. Context. Behav. Sci. **32**, 100748 (2024). https://doi.org/10.1016/j.jcbs.2024.100748
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021). https://doi.org/10.1109/ICCV48922.2021.00951
6. Deng, J., Ren, F.: A survey of textual emotion recognition and its challenges. IEEE Trans. Affect. Comput. **14**(1), 49–67 (2023). https://doi.org/10.1109/TAFFC.2021.3053275
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. vol. 1, pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423
8. Fang, Y., Huang, W., Wan, G., Su, K., Ye, M.: EMOE: Modality-specific enhanced dynamic emotion experts. In: CVPR. pp. 14314–14324 (2025). https://doi.org/10.1109/CVPR52734.2025.01335
9. George, S.M., Ilyas, P.M.: A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. Neurocomputing **568**, 127015 (2024). https://doi.org/10.1016/j.neucom.2023.127015
10. González-González, M., Belharbi, S., Zeeshan, M.O., Sharafi, M., Aslam, M.H., Pedersoli, M., Lameiras Koerich, A., Bacon, S.L., Granger, E.: BAH dataset for ambivalence/hesitancy recognition in videos for digital behavioural change. In: ICLR (2026). https://doi.org/10.48550/arXiv.2505.19328
11. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv (2023). https://doi.org/10.48550/arXiv.2312.00752
12. Hallmen, T., Kampa, R.N., Deuser, F., Oswald, N., André, E.: Semantic matters: Multimodal features for affective analysis. In: CVPRW. pp. 5761–5770 (2025). https://doi.org/10.1109/CVPRW67362.2025.00570
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
14. Jha, O.G., Bamniya, M., Borthakur, A.: Discriminative flow matching via local generative predictors. arXiv (2026). https://doi.org/10.48550/arXiv.2603.13928
15. Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Kotsia, I., Granger, E., Pedersoli, M., Bacon, S., Baird, A., Gagne, C., et al.: Advancements in affective and behavior analysis: The 8th ABAW workshop and competition. In: CVPRW. pp. 5572–5583 (2025). https://doi.org/10.1109/CVPRW67362.2025.00554

16. Kollias, D., Zafeiriou, S., Kotsia, I., Slabaugh, G., Senadeera, D.C., Zheng, J., Yadav, K.K.K., Shao, C., Hu, G.: From emotions to violence: Multimodal fine-grained behavior analysis at the 9th ABAW. In: ICCVW. pp. 1–12 (2025). https://doi.org/10.1109/ICCVW69036.2025.00006
17. Leong, S.C., Tang, Y.M., Lai, C.H., Lee, C.K.M.: Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. Comput. Sci. Rev. **48**, 100545 (2023). https://doi.org/10.1016/j.cosrev.2023.100545
18. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-LLaVA: Learning united visual representation by alignment before projection. In: Empir. Methods Nat. Lang. Process. pp. 5971–5984. Association for Computational Linguistics (2024). https://doi.org/10.48550/arXiv.2311.10122
19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv (2019). https://doi.org/10.48550/arXiv.1907.11692
20. Pereira, A., Barros, P., Fernandes, B.: BROTHER: Behavioral recognition optimized through heterogeneous ensemble regularization for ambivalence and hesitancy. In: CVPRW. pp. 5362–5369 (2026). https://doi.org/10.48550/arXiv.2603.14361
21. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017). https://doi.org/10.1016/j.inffus.2017.02.003
22. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986). https://doi.org/10.1038/323533a0
23. Ryumina, E., Axyonov, A., Sysoev, D., Abdulkadirov, T., Almetov, K., Morozova, Y., Ryumin, D.: Ensemble-based prototype-augmented multimodal fusion for ambivalence/hesitancy recognition. In: CVPRW. pp. 5409–5418 (2026). https://doi.org/10.48550/arXiv.2603.12848
24. Ryumina, E., Dresvyanskiy, D., Karpov, A.: In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. Neurocomputing **514**, 435–450 (2022). https://doi.org/10.1016/j.neucom.2022.10.013
25. Sajjad, M., Ullah, F.U.M., Ullah, M., Christodoulou, G., Cheikh, F.A., Hijji, M., Muhammad, K., Rodrigues, J.J.P.C.: A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines. Alex. Eng. J. **68**, 817–840 (2023). https://doi.org/10.1016/j.aej.2023.01.017
26. Savchenko, A., Savchenko, L.: Leveraging lightweight facial models and textual modality in audio-visual emotional understanding in-the-wild. In: CVPRW. pp. 5824–5834 (2025). https://doi.org/10.1109/CVPRW67362.2025.00577
27. Savchenko, A.V., Sidorova, A.P.: EmotiEffNet and temporal convolutional networks in video-based facial expression recognition and action unit detection. In: CVPRW. pp. 4849–4859 (2024). https://doi.org/10.1109/CVPRW63382.2024.00488
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. vol. 30, pp. 5998–6008 (2017). https://doi.org/10.5555/3295222.3295349
29. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE V2: Scaling video masked autoencoders with dual masking. In: CVPR. pp. 14549–14560 (2023). https://doi.org/10.1109/CVPR52729.2023.01398