# GAME: Learning Multimodal Interactions via Graph Structures for Personality Trait Estimation

Kangsheng Wang , Yuhang Li , Chengwei Ye , Yufei Lin , Huanzhen Zhang , Bohan Hu * , Linuo Xu , Shuyan Liu

Posted Date: 21 May 2025

doi: 10.20944/preprints202505.1700.v1

Keywords: multimodal feature learning; graph neural networks; apparent personality analysis

 doi:10.20944/preprints202505.1700.v1



*Article*

# GAME: Learning Multimodal Interactions via Graph Structures for Personality Trait Estimation

**Kangsheng Wang [1,†], Yuhang Li [2,†], Chengwei Ye [3], Yufei Lin [3], Huanzhen Zhang [4], Bohan Hu [5,*], Linuo Xu [6] and Shuyan Liu [7]**

[1] University of Science and Technology Beijing
[2] Zhengzhou University
[3] Homesite Group Inc
[4] Chewy Inc
[5] Communication University of China
[6] Yunnan University of Finance and Economics
[7] Yunnan University
* Correspondence: hubohan17@gmail.com
† These authors contributed equally to this work.

**Abstract:** Apparent personality analysis from short videos poses significant challenges due to the complex interplay of visual, auditory, and textual cues. In this paper, we propose GAME, a Graph-Augmented Multimodal Encoder designed to robustly model and fuse multi-source features for automatic personality prediction. For the visual stream, we construct a facial graph and introduce a dual-branch Geo Two-Stream Network, which combines Graph Convolutional Networks (GCNs) and Convolutional Neural Networks (CNNs) with attention mechanisms to capture both structural and appearance-based facial cues. Complementing this, global context and identity features are extracted using pretrained ResNet18 and VGGFace backbones. To capture temporal dynamics, frame-level features are processed by a BiGRU enhanced with temporal attention modules. Meanwhile, audio representations are derived from the VGGish network, and linguistic semantics are captured via the XLM-Roberta transformer. To achieve effective multimodal integration, we propose a Channel Attention-based Fusion module, followed by a Multi-Layer Perceptron (MLP) regression head for predicting personality traits. Extensive experiments show that GAME consistently outperforms existing methods across multiple benchmarks, validating its effectiveness and generalizability.

**Keywords:** multimodal feature learning; Graph Neural Networks; apparent personality analysis

## 1. Introduction

Understanding and interpreting human personality through automatic computational methods has become a vital research frontier across computer vision, audio processing, and computational linguistics [1]. Among various models, the Big Five Personality Traits—namely Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—have emerged as a widely accepted psychological framework for trait assessment. Predicting such traits not only enhances personalized human-computer interaction but also has practical implications in domains such as recruitment [2], targeted advertising, and social robotics.

The ability to integrate heterogeneous cues from visual, auditory, and textual modalities has shown great promise in improving the reliability and interpretability of apparent personality analysis. Recent works [3] have demonstrated that fusing multiple modalities—such as facial expressions, voice characteristics, and spoken language—can significantly enhance prediction performance by capturing complementary information. Prior studies [1,4] have employed deep residual networks and CNN-based architectures for multimodal fusion, while others have investigated the temporal

 doi:10.20944/preprints202505.1700.v1



evolution of audio-visual signals using recurrent neural networks[15–17]. For instance, scene contexts [4], facial micro-expressions [5], and speech patterns have all been found to correlate with various personality indicators. However, most existing frameworks[18–22] rely heavily on convolutional structures and often neglect the spatial topology of facial features or the relational geometry embedded in visual data[6,7].

To address these limitations, we propose GAME (Graph-Augmented Multimodal Encoder), a novel end-to-end framework designed to learn rich, interpretable representations from multimodal inputs for apparent personality trait prediction. Unlike conventional models, GAME introduces a facial graph structure to explicitly model geometric dependencies across facial landmarks. A Geo Two-Stream Network is developed, combining Graph Convolutional Networks (GCNs) and CNNs under attention-based supervision to jointly capture structural and appearance features from facial images. Complementing this, we employ pretrained ResNet18 and VGGFace models to obtain global spatial representations, while a BiGRU with a temporal attention block captures salient temporal dynamics from video frames.

For non-visual modalities, VGGish [8] and XLM-Roberta [9] are leveraged to extract informative features from audio and textual streams, respectively. To integrate these heterogeneous cues, we introduce a Multimodal Channel Attention Module, which dynamically reweights the contributions of each modality before passing the fused representation to a regression-based MLP for final prediction. Through comprehensive evaluations on the ChaLearn First Impression V2 dataset, GAME achieves superior performance compared to existing state-of-the-art methods, demonstrating its robustness and effectiveness.

Our main contributions are summarized as follows:

(1). We introduce GAME, a unified framework that combines graph structure learning and multimodal attention fusion to enhance apparent personality trait prediction.

(2). We design a novel facial graph structure and dual-stream visual encoder that captures both geometric topology and appearance cues critical to trait inference.

(3). We propose an effective attention-based temporal modeling module and a multimodal channel attention mechanism for dynamic feature integration.

(4). Extensive experiments validate the superiority of our method over previous approaches on benchmark datasets.

## 2. Proposed Approach

As depicted in Figure 1, our proposed framework GAME takes a short video of a subject as input and outputs a set of predicted personality trait scores. The overall pipeline consists of four key stages: data preprocessing, modality-specific feature encoding, attention-guided fusion, and trait regression.

Initially, the raw video is decomposed into three parallel data streams: one capturing visual information (including facial regions and global scene context), one extracting audio signals, and one representing textual content (e.g., transcripts). Each stream is then processed by a dedicated encoder tailored to its modality—facial graphs and image frames are fed into visual encoders, audio waveforms into a CNN-based feature extractor, and textual inputs into a pretrained language model.

Once the modality-specific features are obtained, a Multimodal Channel Attention Module adaptively reweights and merges them, allowing the model to emphasize more informative cues. The resulting unified representation is subsequently passed through a Multi-Layer Perceptron (MLP), which performs final personality trait regression across the Big Five dimensions.

 doi:10.20944/preprints202505.1700.v1



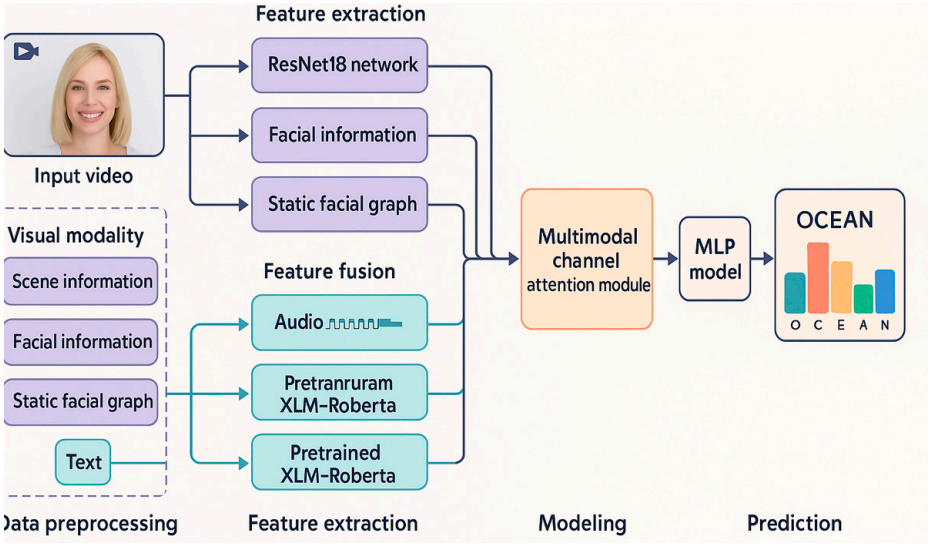


**Figure 1.** The framework of the proposed GAME framework.

*2.1. Feature Extraction Module*

**Visual Modality.**

*Data Pre-processing.* To process the visual stream, raw video clips are first decomposed into individual image frames. Each frame is analyzed using a custom-trained UltraFace detector to isolate facial regions. Following face localization, 113 facial landmarks are extracted using the PFLD keypoint detection network, enabling the construction of a dense facial topology.

*Visual Feature Encoding within GAME.* The visual module in GAME is structured around three complementary feature pathways: (1) local static facial appearance, (2) facial geometric structure, and (3) scene-level global context.

To model facial appearance and structure at a fine-grained level, we design a Geo Two-Stream Network, which integrates convolutional and graph-based processing. As shown in Figure 2, this architecture contains two parallel branches. One branch takes cropped facial images as input and passes them through a CNN pipeline composed of three convolutional layers, enhanced with a Saliency Attention (SA) module. This module guides the network to focus on personality-relevant facial regions by dynamically emphasizing high-salience areas.

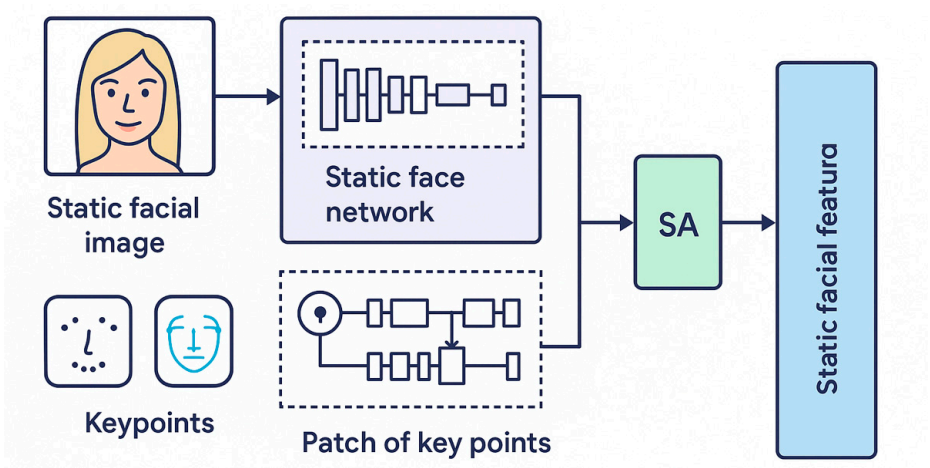


**Figure 2.** Local static facial appearance and geometric features.

The second branch addresses geometric reasoning. It first employs a CNN to extract visual features from small image patches centered on each landmark. These features, together with the spatial configuration of landmarks, are passed to a Graph Convolutional Network (GCN). The GCN

 doi:10.20944/preprints202505.1700.v1



operates on a facial graph defined as G(R, E), where R denotes the set of landmark coordinates and E represents the adjacency structure encoding their spatial relationships. This allows the model to capture not only localized appearance cues but also the underlying facial structure, which is essential for apparent personality trait estimation.

$$Fgeo = G(R, E)$$

Within the facial geometric stream of the proposed GAME framework, the output of the graph-based module is denoted as $F_{geo}$, which represents the geometric feature embedding derived from the facial landmark graph $G(R, E)$. Here, $F_{geo} \in R^{d_1 \times P}$, where $d_1$ is the dimensionality of each node's feature vector post-GCN processing, and $P = 113$ corresponds to the total number of detected facial landmarks. This notation is consistently used throughout subsequent descriptions.

To enhance the representational richness of each graph node, we supplement the geometric topology with localized appearance cues. Specifically, for every landmark, a square patch centered around its coordinates is extracted from the static facial image, resulting in a set of local regions $\mathcal{V} = \{V_i \in R^{h \times w}\}_{i=1}^{P}$, where each patch has a fixed size of h = 48 and w = 48. These image patches serve as localized visual context for the corresponding keypoints.

To encode these patches, we construct a dedicated CNN module for each landmark, yielding P individual CNN encoders. Each of these modules processes a specific keypoint's local patch and extracts a high-level representation of its visual appearance. The resulting local appearance features are subsequently aligned with the graph nodes and integrated into the GCN pipeline, thereby allowing the model to reason jointly over both structural and visual information associated with each facial keypoint.

To obtain expressive representations of each facial keypoint, we utilize a set of dedicated CNN modules to encode the corresponding local appearance patches. For the i-th keypoint, the extracted local feature is denoted as:

$$f_{\text{local}}^{(i)} = \mathcal{F}^{(i)}(v_i), \quad i \in [0, P)$$

where $f_{\text{local}}^{(i)} \in R^{d_2 \times 1}$ represents the appearance embedding produced by the i-th CNN module $\mathcal{F}^{(i)}$, and v_i is the local patch centered on the i-th landmark. The value $d_2$ denotes the feature channel size for each local image.

*Global Spatial-Temporal Scene Representation with ResNet18.* To effectively capture the evolving scene context throughout video sequences, we adopt a ResNet18 model pretrained on the Places365 dataset as the backbone for scene-level feature extraction (see Figure 3). Each video frame is independently passed through the pretrained ResNet18, yielding a sequence of frame-wise scene embeddings. These embeddings encapsulate global spatial information about the environment surrounding the subject. To further model temporal dependencies and highlight frames with strong contextual cues, the resulting scene feature sequence is processed by a Bidirectional GRU (BiGRU) augmented with a temporal attention block (Figure 4). This module enables the model to selectively emphasize informative moments in the video that are more indicative of personality-related traits.

*Global Facial Appearance Encoding via Fine-tuned VGGFace.* For holistic representation of facial appearance across time, we incorporate the VGGFace [10] model as a deep feature extractor. While VGGFace is originally trained for face recognition, it may not be optimally aligned with the goal of personality trait estimation. To address this, we fine-tune the model on facial frames derived from the Big Five personality dataset, allowing it to adapt to trait-relevant facial cues. Each frame is processed using the fine-tuned VGGFace encoder, enhanced with the DAN+ feature aggregation technique to obtain refined per-frame appearance descriptors. These frame-level features are then passed through a BiGRU combined with a temporal attention mechanism, which dynamically weights frames according to their personality-related saliency. The resulting output is a global facial appearance feature that captures both spatial details and temporal dynamics of the subject's face.

 doi:10.20944/preprints202505.1700.v1



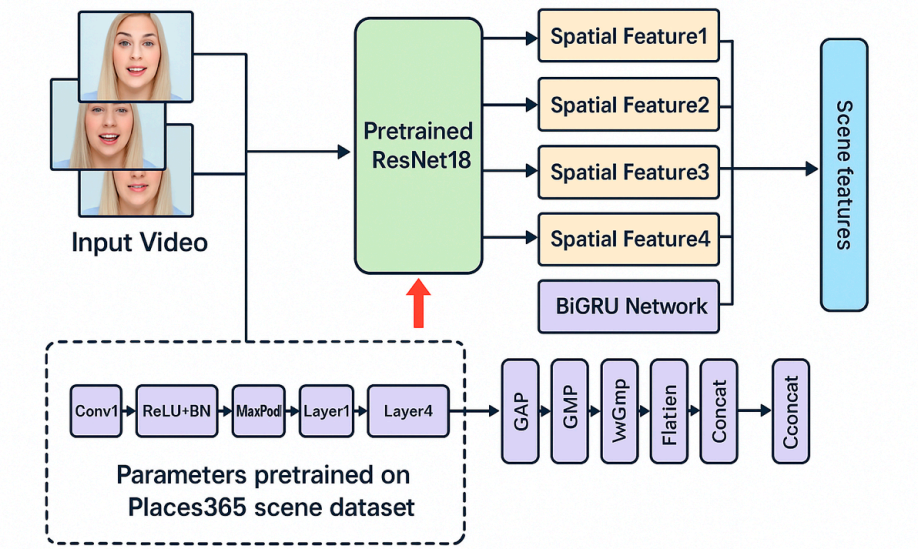


**Figure 3.** Global spatial-temporal scene feature based on Resnet18.

**Textual Modality.**

*Text Preprocessing.* To enable effective semantic modeling of spoken language content, the input transcript is first segmented into binary-level character tokens. Each character is mapped to a predefined vocabulary, forming an index-based representation of the sentence. This low-level encoding serves as the foundation for subsequent high-dimensional feature learning.

*Text Encoder Architecture.* We leverage the XLM-Roberta model—a multilingual transformer-based encoder pretrained on large-scale textual corpora—for sentence-level embedding extraction. The model architecture comprises 12 transformer layers and 12 self-attention heads, producing a fixed-length output vector of 768 dimensions for each sentence. These embeddings capture deep semantic and syntactic cues, which are integrated into the multimodal personality prediction pipeline.

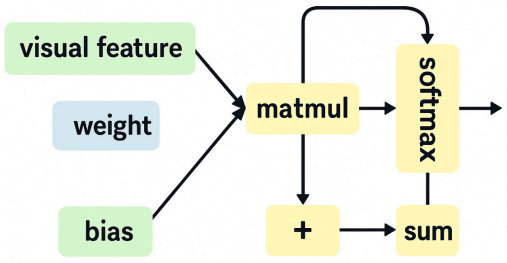


**Figure 4.** Temporal attention block module.

*2.2. Multimodal Fusion*

To effectively integrate the heterogeneous features extracted from different modalities, we design a Multimodal Channel Attention Module. This module aims to adaptively evaluate the relevance of each feature stream while reducing redundancy caused by modal diversity.

As illustrated in Figure 5, the feature representations obtained from five different branches (e.g., visual, audio, and text) are first concatenated into a single composite feature vector F. This vector is then passed through two stacked fully connected layers with non-linear activation to compute an attention weight vector \alpha, formulated as:

$$\alpha = \tanh(W_2 \cdot \tanh(W_1 F + b) + c)$$

 doi:10.20944/preprints202505.1700.v1



where $W_1$ and $W_2$ are the weight matrices of the fully connected layers, and $b$, $c$ denote their respective biases. The use of the \tanh activation constrains the attention values within the interval [-1, 1], enabling fine-grained modulation of each dimension in the multimodal representation.

To preserve original information while applying attention, a residual connection is introduced. The final attended feature $F'$ is obtained by:

$$F' = F \cdot \alpha + F$$

This residual-enhanced output $F'$ is subsequently fed into a downstream regression head for trait prediction.

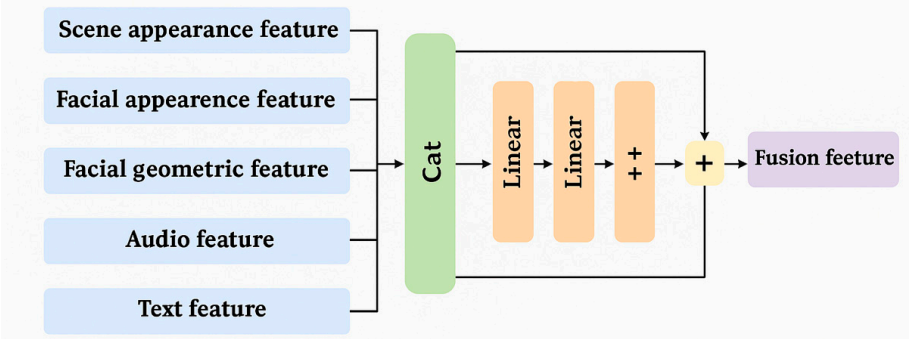


**Figure 5.** Multimodal channel attention module.

*2.3. Model Training*

For the final stage of the pipeline, a three-layer Multi-Layer Perceptron (MLP) is employed to estimate scores corresponding to the Big Five personality traits. The first two layers use ReLU activation and are followed by dropout layers to prevent overfitting. The output layer utilizes the sigmoid function to ensure that the predicted scores fall within the normalized range [0, 1], suitable for trait regression.

*2.4. Loss Function Design*

To ensure robust and stable optimization, we formulate the training objective as a composite multi-task loss, defined as:

$$\mathcal{L}\text{Total} = \mathcal{L}\text{RMSE} + \mathcal{L}\log c\,osh + \mathcal{L}\text{B}$$

Here, $\mathcal{L}\text{RMSE}$ penalizes large prediction errors via Root Mean Square Error, $\mathcal{L}\log c\,osh$ provides a smooth approximation of absolute error, $\mathcal{L}\text{B}$ is used to enhance robustness to outliers by focusing on small-magnitude errors.

## 3. Experiment

*3.1. Setup*

All experiments are implemented using the PyTorch framework and executed on a system equipped with an NVIDIA GeForce RTX 3090 GPU. For optimization, the Geo Two-Stream Network is trained using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.1, while other modality-specific components are optimized using the Adam optimizer. These components are initialized with a learning rate of 0.0001, and the regularization configuration includes a weight decay of 1e-4. Momentum for SGD follows PyTorch's default settings.

To evaluate the performance of our proposed GAME framework, we conduct experiments on the ChaLearn First Impression-V2 (CFI-V2) dataset [14], which is widely used for apparent personality and interview analysis tasks. This dataset contains a total of 10,000 short video clips, which are partitioned into 6,000 for training, 2,000 for validation, and 2,000 for testing.

 doi:10.20944/preprints202505.1700.v1



For quantitative evaluation, we adopt the mean accuracy metric across the Big Five personality traits. This metric is defined as:

$$\text{Mean Accuracy} = 1.0 - \frac{1}{5N} \sum_{j=1}^{5} \sum_{i=1}^{N} \left| p_{ij} - \hat{p}_{ij} \right|$$

where $p_{ij}$ denotes the ground truth score of the j-th trait for the i-th video sample, $\hat{p}_{ij}$ is the corresponding predicted value, and $N$ is the total number of samples. This evaluation criterion reflects the average absolute prediction error normalized across all five traits.

### *3.2. Comparison with the SOTA*

To validate the effectiveness of the proposed GAME framework, we conduct a series of performance comparisons against several state-of-the-art methods. Tables 1 and 2 summarize the results on the ChaLearn First Impression-V2 (CFI-V2) dataset for both the validation and test sets.

As observed from the results, our method consistently outperforms existing approaches across all five personality traits. Specifically, GAME achieves an average prediction accuracy of 91.85% on the validation set and 91.68% on the test set. These results demonstrate a clear improvement over previous best-performing models, highlighting the effectiveness of our graph-augmented multimodal design in capturing nuanced personality indicators from short video data.

**Table 1.** Comparison with state-of-the-art methods on the first impressions dataset.

| Methods | Modalities | Open. | Cons. | Extr. | Agre. | Neur. | ACC(mean) |
|---|---|---|---|---|---|---|---|
| Wei et al. [1] | Visual and audio | 0.9120 | 0.9170 | 0.9130 | 0.9130 | 0.9100 | 0.9130 |
| Kaya et al. [2] | Visual and audio | 0.9169 | 0.9166 | 0.9206 | 0.9161 | 0.9149 | 0.9170 |
| Güçlütürk et al. [3] | Visual, audio and text | 0.9110 | 0.9150 | 0.9110 | 0.9110 | 0.9100 | 0.9116 |
| Bekhouche et al. [11] | Visual | 0.9138 | 0.9166 | 0.9175 | 0.9166 | 0.9130 | 0.9155 |
| Subramaniam et al. [5] | Visual and audio | 0.9131 | 0.9136 | 0.9145 | 0.9157 | 0.9098 | 0.9133 |
| Gurpınar et al. [12] | Visual and audio | 0.9140 | 0.9140 | 0.9190 | 0.9140 | 0.9120 | 0.9150 |
| Suman et al. [6] | Visual, audio and text | - | - | - | - | - | 0.9146 |
| Our method | Visual, audio and text | 0.9179 | 0.9215 | 0.9191 | 0.9187 | 0.9152 | **0.9185** |

**Table 2.** Comparison with state-of-the-art methods on the first impressions dataset.

| Methods | Modalities | Open. | Cons. | Extr. | Agre. | Neur. | ACC(mean) |
|---|---|---|---|---|---|---|---|
| Subramaniam et al. [5] | Visual and audio | 0.9117 | 0.9119 | 0.9150 | 0.9119 | 0.9099 | 0.9121 |
| Wei et al. [1] | Visual and audio | 0.9120 | 0.9170 | 0.9130 | 0.9130 | 0.9100 | 0.9130 |
| Zhang et al. [13] | Visual and audio | 0.9123 | 0.9166 | 0.9133 | 0.9126 | 0.9100 | 0.9130 |
| Güçlütürk et al. [3] | Audio, visual, text | 0.9111 | 0.9152 | 0.9112 | 0.9112 | 0.9104 | 0.9118 |
| Bekhouche et al. [11] | Visual | 0.9101 | 0.9138 | 0.9155 | 0.9103 | 0.9083 | 0.9116 |
| Ventura et al. [4] | Visual and audio | 0.9100 | 0.9140 | 0.9150 | 0.9120 | 0.9070 | 0.9116 |
| Suman et al [6] | Visual, audio and text | 0.9111 | 0.9192 | 0.9173 | 0.9132 | 0.9103 | 0.9143 |
| Our method | Visual, audio and text | 0.9169 | 0.9206 | 0.9189 | 0.9139 | 0.9136 | **0.9168** |

## 4. Discussion

### *4.1. Effectiveness of Feature Fusion Strategies*

Given that not all modalities contribute equally to the prediction of the Big Five personality traits, we explore the impact of different fusion strategies on model performance. To address the issue of modal imbalance, we introduce a Channel Attention-based Fusion Module (Channel Attn), which adaptively reweights feature contributions across modalities.

As shown in Table 3, the Channel Attn module consistently outperforms traditional fusion approaches, confirming its ability to emphasize informative modalities while suppressing less relevant ones. These results indicate that our attention-guided fusion mechanism not only enhances

 doi:10.20944/preprints202505.1700.v1



the representational capacity of the integrated feature space but also improves the overall robustness and accuracy of the personality prediction framework.

**Table 3.** Performance of different feature fusion strategy.

| Dataset | Fusion Type | Method | ACC(mean) |
|---|---|---|---|
| CFI | Early | Concatenation | 0.9180 |
| | | Modality Attn | 0.9180 |
| | | TFN fusion | 0.9182 |
| | | AMBF Attn | 0.9151 |
| | | Channel Attn(us) | **0.9185** |
| | Hybrid | CentralNet | 0.9162 |

*4.2. Impact of the Temporal Attention Block*

To evaluate the contribution of temporal modeling, we conduct experiments to assess the effectiveness of the temporal attention block module within our framework. Table 4 reports the comparative results of the model with and without this module.

The results clearly demonstrate that integrating the temporal attention mechanism leads to a noticeable improvement in prediction accuracy. By dynamically emphasizing key frames across the video sequence, the attention block enables the model to focus on temporally salient moments that are more indicative of personality traits. This confirms the importance of fine-grained temporal weighting in capturing behaviorally rich cues from video data. Sample Heading (Third Level). Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

**Table 4.** Comparison of network with temporal attention-block.

| Temporal attention block module | ACC(mean) |
|---|---|
| - | 0.9178 |
| ✓ | **0.9185** |

## 5. Conclusion

In this paper, we proposed GAME, a novel Graph-Augmented Multimodal Encoder designed for apparent personality trait prediction from short video data. The framework introduces a graph-based attention network that models static facial geometry, enhancing the discriminative power of visual features. To capture rich multimodal cues, pretrained backbones are employed to extract global scene context, facial appearance, audio signals, and text representations.

To better exploit temporal dependencies, we incorporate a Bi-GRU/LSTM module equipped with a temporal attention block, which allows the model to highlight frames containing behaviorally relevant information. All modality-specific features are then fused via a channel-wise attention mechanism, ensuring adaptive weighting across modalities before feeding into an MLP regression head.

Extensive experiments on the ChaLearn First Impression V2 dataset demonstrate the effectiveness of the proposed framework, achieving competitive performance and outperforming

 doi:10.20944/preprints202505.1700.v1



several state-of-the-art baselines. This work highlights the potential of integrating graph structures and attention mechanisms in multimodal personality analysis.

## References


1. Wei, X.-S., Zhang, C.-L., Zhang, H., Wu, J.: Deep bimodal regression of apparent personality traits from short video sequences. IEEE Trans. Affective Comput. **9**(3), 303–315 (2017)
2. Kaya, H., Gurpinar, F., Salah, A.A.: Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs. In: *Proc. CVPR Workshops*, pp. 1–9. IEEE, Honolulu (2017)
3. Güçlütürk, Y., Güçlü, U., Baró, X., Escalante, H.J., Guyon, I., Escalera, S., van Gerven, M.A.J., van Lier, R.: Multimodal first impression analysis with deep residual networks. IEEE Trans. Affective Comput. **9**(3), 316–329 (2017)
4. Ventura, C., Masip, D., Lapedriza, A.: Interpreting CNN models for apparent personality trait regression. In: *Proc. CVPR Workshops*, pp. 55–63. IEEE, Honolulu (2017)
5. Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., Mittal, A.: Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In: Leibe, B. et al. (eds.) *ECCV 2016 Workshops*, Part III, LNCS, vol. 9915, pp. 337–348. Springer, Cham (2016)
6. Suman, C., Saha, S., Gupta, A., Pandey, S.K., Bhattacharyya, P.: A multi-modal personality prediction system. Knowl.-Based Syst. **236**, 107715 (2022)
7. Escalante, H.J., Kaya, H., Salah, A.A., Escalera, S., Güçlütürk, Y., Güçlü, U., Baró, X., Guyon, I., Jacques Jr., J.C.S., Madadi, M., et al.: Modeling, recognizing, and explaining apparent personality from videos. IEEE Trans. Affective Comput. **13**(2), 894–911 (2020)
8. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: CNN architectures for large-scale audio classification. In: *Proc. ICASSP*, pp. 131–135. IEEE, New Orleans (2017)
9. Qu, S., Yang, Y., Que, Q.: Emotion classification for Spanish with XLM-Roberta and TextCNN. In: *IberLEF@SEPLN*, pp. 94–100 (2021)
10. Parkhi, O., Vedaldi, A., Zisserman, A.: Deep face recognition. In: *Proc. BMVC*, pp. 1–12. BMVA Press, Swansea (2015)
11. Bekhouche, S.E., Dornaika, F., Ouafi, A., Taleb-Ahmed, A.: Personality traits and job candidate screening via analyzing facial videos. In: *Proc. CVPR Workshops*, pp. 10–13. IEEE, Honolulu (2017)
12. Gurpinar, F., Kaya, H., Salah, A.A.: Multimodal fusion of audio, scene, and face features for first impression estimation. In: *Proc. ICPR*, pp. 43–48. IEEE, Cancún (2016)
13. Zhang, C.-L., Zhang, H., Wei, X.-S., Wu, J.: Deep bimodal regression for apparent personality analysis. In: Leibe, B. et al. (eds.) *ECCV 2016*, LNCS, vol. 9905, pp. 311–324. Springer, Cham (2016)
14. Escalante, H.J., Guyon, I., Escalera, S., Jacques Jr., J.C.S., Madadi, M., Baró, X., Ayache, S., Viegas, E., Güçlütürk, Y., Güçlü, U., et al.: Design of an explainable machine learning challenge for video interviews. In: *Proc. IJCNN*, pp. 3688–3695. IEEE, Anchorage (2017)
15. Qi, X., Zhang, Z., Zheng, H., et al.: MedConv: Convolutions Beat Transformers on Long-Tailed Bone Density Prediction. *arXiv preprint* arXiv:2502.00631 (2025)
16. Wang, K., Zhang, X., Guo, Z., et al.: CSCE: Boosting LLM Reasoning by Simultaneous Enhancing of Causal Significance and Consistency. *arXiv preprint* arXiv:2409.17174 (2024)
17. Liu, S., Wang, K.: Comprehensive Review: Advancing Cognitive Computing through Theory of Mind Integration and Deep Learning in Artificial Intelligence. In: *Proc. 8th Int. Conf. on Computer Science and Application Engineering*, pp. 31–35 (2024)
18. Zhang, X., Wang, K., Hu, T., et al.: Enhancing Autonomous Driving through Dual-Process Learning with Behavior and Reflection Integration. In: *ICASSP 2025 – IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, Seoul (2025)
19. Zou, B., Guo, Z., Qin, W., et al.: Synergistic Spotting and Recognition of Micro-Expression via Temporal State Transition. In: *ICASSP 2025 – IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, Seoul (2025)

 doi:10.20944/preprints202505.1700.v1



20. Hu, T., Zhang, X., Ma, H., et al.: Autonomous Driving System Based on Dual Process Theory and Deliberate Practice Theory. *Manuscript* (2025)
21. Zhang, X., Wang, K., Hu, T., et al.: Efficient Knowledge Transfer in Multi-Task Learning through Task-Adaptive Low-Rank Representation. *arXiv preprint* arXiv:2505.00009 (2025)
22. Wang, K., Ye, C., Zhang, H., et al.: Graph-Driven Multimodal Feature Learning Framework for Apparent Personality Assessment. *arXiv preprint* arXiv:2504.11515 (2025)