http://dx.doi.org/10.62762/TETAI.2025.279350

RESEARCH ARTICLE

Graph-Driven Multimodal Feature Learning Framework for Apparent Personality Assessment

Kangsheng Wang[1,†], **Chengwei Ye**[2,†,*], **Huanzhen Zhang**[3,†], **Linuo Xu**[4] **and Shuyan Liu**[5]

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2] Homesite Group Inc, GA 30043, United States
[3] Payment Department, Chewy Inc, MA 02210, United States
[4] School of Information, Yunnan University of Finance and Economics, Yunnan 650000, China
[5] School of Information Science and Technology, Yunnan University, Yunnan 650000, China

Abstract

Predicting personality traits automatically has emerged as a challenging problem in computer vision. This paper introduces an innovative multimodal feature learning framework for personality analysis in short video clips. For visual processing, we construct a facial graph and design a Geo-based two-stream network incorporating an attention mechanism, leveraging both Graph Convolutional Networks (GCN) and Convolutional Neural Networks (CNN) to capture static facial expressions. Additionally, ResNet18 and VGGFace networks are employed to extract global scene and facial appearance features at the frame level. To capture dynamic temporal information, we integrate a BiGRU with a temporal attention module for extracting salient frame representations. To further enhance the model's robustness, we incorporate the VGGish CNN for audio-based features and XLM-Roberta for text-based features. Finally, a multimodal channel attention mechanism is introduced to integrate different modalities, and a Multi-Layer Perceptron (MLP) regression model is utilized to predict personality traits. Experimental results confirm that our proposed framework surpasses existing state-of-the-art approaches in performance.

Keywords: personality prediction, facial graph, graph convolutional network(GCN), convolutional neural network(CNN), attention mechanism, geometric and appearance features.

1 Introduction

In recent years, the study of personality analysis has gained significant interest across various fields, including computer vision, linguistics, and related disciplines [1]. Over the past few decades, researchers have developed different personality trait models

Academic Editor:
Aytuğ Onan

Submitted: 09 March 2025
Accepted: 11 April 2025
Published: 15 April 2025

Vol. 2, **No.** 2, 2025.
10.62762/TETAI.2025.279350

***Corresponding author:**
✉ Chengwei Ye
chengweiye18@gmail.com

† These authors contributed equally to this work

Citation
Wang, K., Ye, C., Zhang, H., Xu, L., & Liu, S. (2025). Graph-Driven Multimodal Feature Learning Framework for Apparent Personality Assessment. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 2(2), 57–67.

based on psychological scales, with the Big Five model being one of the most widely recognized. This model evaluates personality through five key dimensions: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). Personality traits play a crucial role in shaping individuals' decision-making processes and preferences, making personality prediction valuable for real-world applications such as job interviews [2], consumer behavior analysis, and beyond.

Leveraging multimodal features enhances the reliability of personality prediction by utilizing the complementary nature of different information sources. As a result, integrating multiple modalities has become a prominent approach in studying the Big Five personality traits. For example, Güçlütürk et al. [3] developed a deep residual network that effectively combines multiple modalities for personality trait recognition. Additionally, research has highlighted the predictive significance of facial features [1]. In particular, Ventura et al. [4] analyzed the role of specific Action Units (AUs) in facial feature assessment. To explore multimodal fusion, Subramaniam et al. [5] introduced two bi-modal end-to-end deep learning frameworks incorporating temporally structured audio and stochastic visual features. Furthermore, Kaya et al. [2] adapted a pre-trained Deep Convolutional Neural Network (DCNN) using an emotion change dataset to refine facial feature extraction. More recently, studies have investigated the link between visual contexts in short videos and personality traits. For instance, Suman et al. [6] utilized MTCNN and ResNet to derive facial and environmental attributes from visual inputs, while VGGish and an n-gram CNN model were employed for analyzing audio and text features. Similarly, Escalante et al. [7] proposed a multimodal deep learning approach incorporating audio, visual, and textual information for personality trait recognition [8, 9]. Despite these advances, most existing methods predominantly rely on CNN-based architectures, overlooking the potential of graph neural networks with visual attention to capture facial geometric structures. Moreover, while current models show promise in personality prediction, their accuracy still requires further enhancement.

To advance multimodal feature exploration and improve personality prediction accuracy, this paper presents a novel framework, MFLF-GSL (Multimodal Feature Learning Framework with Graph Structure Learning), designed for apparent personality analysis. This approach integrates visual, audio, and textual modalities to enhance predictive performance. For visual analysis, we propose a Geo two-stream method that captures both facial appearance and geometric attributes using a combination of Graph Convolutional Networks (GCN) and Convolutional Neural Networks (CNN) models with an attention mechanism. This method aims to model interactions between facial structure and appearance while emphasizing key facial regions relevant to personality traits. Specifically, GCN is responsible for extracting critical geometric relationships based on facial regions, whereas CNN captures local appearance-based features. Additionally, ResNet18 and VGGFace networks are utilized to derive spatial global facial appearance features from single images, contributing to personality inference. To capture temporal dependencies, a BiGRU network with a temporal attention module is employed to identify salient frame-level features over time.

To fully harness the complementary nature of multimodal data and enhance robustness, Log-Mel Spectrogram and VGGish CNN [10] are used for audio feature extraction, while the pre-trained XLM-RoBERTa [11] model is leveraged for text embeddings. Finally, a multimodal channel attention mechanism is introduced to integrate these features into a fused representation, which serves as input for an MLP model to predict the Big Five personality traits. The effectiveness of MFLF-GSL was validated through experiments on the ChaLearn First Impression-V2 dataset, demonstrating the proposed framework's ability to improve personality prediction performance.

The main contributions are summarized as follows:

- We introduce a multimodal feature learning framework that incorporates a graph structure learning network with an attention mechanism to extract spatial-temporal features for personality trait prediction.
- We develop an effective facial graph representation to extract both facial appearance and geometric features from static images. Additionally, we propose a novel graph structure learning network that captures personality-related representations by leveraging facial topology.
- A multimodal channel attention module is designed to efficiently integrate features from

multiple modalities. Furthermore, we introduce a temporal attention block module to emphasize crucial frame-level features in sequential data.

- Extensive experiments are conducted to validate the proposed framework, demonstrating its effectiveness in personality trait prediction.

2 Proposed approach

As depicted in Figure 1, our framework takes a short video of an individual as input and predicts their personality traits. It consists of four key stages: data preprocessing, feature extraction using modality-specific networks, feature fusion, and final regression via an MLP model. During data preprocessing, the input video is segmented into distinct streams corresponding to multi-focus visual input, audio, and text. For feature extraction, separate modality-specific networks are employed to learn personality-related features, which are subsequently integrated using a multimodal channel attention module. Finally, the fused feature representation is used to predict personality trait scores.

2.1 Feature extraction module

2.1.1 Visual modality

Data Pre-processing: In the preprocessing stage for the visual modality, the video data is first converted into an image sequence. A self-trained UltraFace model is then applied to detect and extract facial images. Additionally, the PFLD face keypoint detector is used to identify 113 facial keypoints.

Model Architecture: This paper introduces three primary visual feature extraction methods: local static facial appearance features, facial geometric features, and scene appearance features.

Local Facial Appearance and Geometric Features Based on CNN and GCN Models: We introduce a Geo two-stream network that leverages CNN and GCN models to extract both facial appearance and geometric features from static facial images, as depicted in Figure 2.

The Geo network comprises two parallel streams: one dedicated to facial appearance and the other to facial geometry. The facial appearance image is fed into a static facial appearance CNN stream, which extracts local facial appearance features. This stream consists of three convolutional modules, enhanced by an attention mechanism (SA module), which helps the CNN model focus on salient facial regions associated with personality traits.

The facial geometric GCN stream integrates both CNN and GCN models. The CNN model processes local image patches around facial keypoints to extract localized appearance features, while the GCN model is responsible for capturing geometric structural features based on keypoint relationships. Specifically, the GCN module operates on a graph representation, denoted as $G(R, E)$,

$$F_{geo} = G(R, E) \tag{1}$$

where R represents the relative coordinate sets of keypoints, and E defines the edge connections. The formulation is as follows:

$$H^{(l+1)} = \sigma\left(D^{-1/2}AD^{-1/2}H^{(l)}W^{(l)}\right) \tag{2}$$

where $H^{(l)}$ is the node feature matrix at layer l, A is the adjacency matrix, D is the degree matrix, $W^{(l)}$ is the trainable weight matrix, and σ represents the activation function.

Here, F_{geo} represents the geometric feature representation of the facial keypoints, which is the output of $G(R, E)$. Specifically, $F_{geo} \in \mathbb{R}^{d_1 \times P}$ denotes that each keypoint is represented by a feature of dimension d_1 after being processed by the GCN module, where P corresponds to the total number of facial keypoints, set as $P = 113$. The same notation is consistently used in the following sections.

To enrich node features in the GCN stream, local appearance information from keypoints is incorporated as a supplementary descriptor. Specifically, for each keypoint in the static facial image, a local image patch is extracted, forming a set of local images:

$$V = \{V \in \mathbb{R}^{h \times w \times P}\}, \tag{3}$$

where h and w denote the height and width of each local patch, both set to 48.

To integrate local appearance features into the GCN stream, this paper employs P independent CNN modules, each dedicated to extracting localized features around a specific keypoint. These CNN modules share the same architecture but operate independently without parameter sharing. The local appearance representation extraction process is formulated as:

$$f_{local}^{(i)} = F^{(i)}(v_i), \quad i \in [0, P) \tag{4}$$

where $f_{local}^{(i)} \in \mathbb{R}^{(d_2 \times 1)}$ denotes the feature representation of the i-th local image, with a

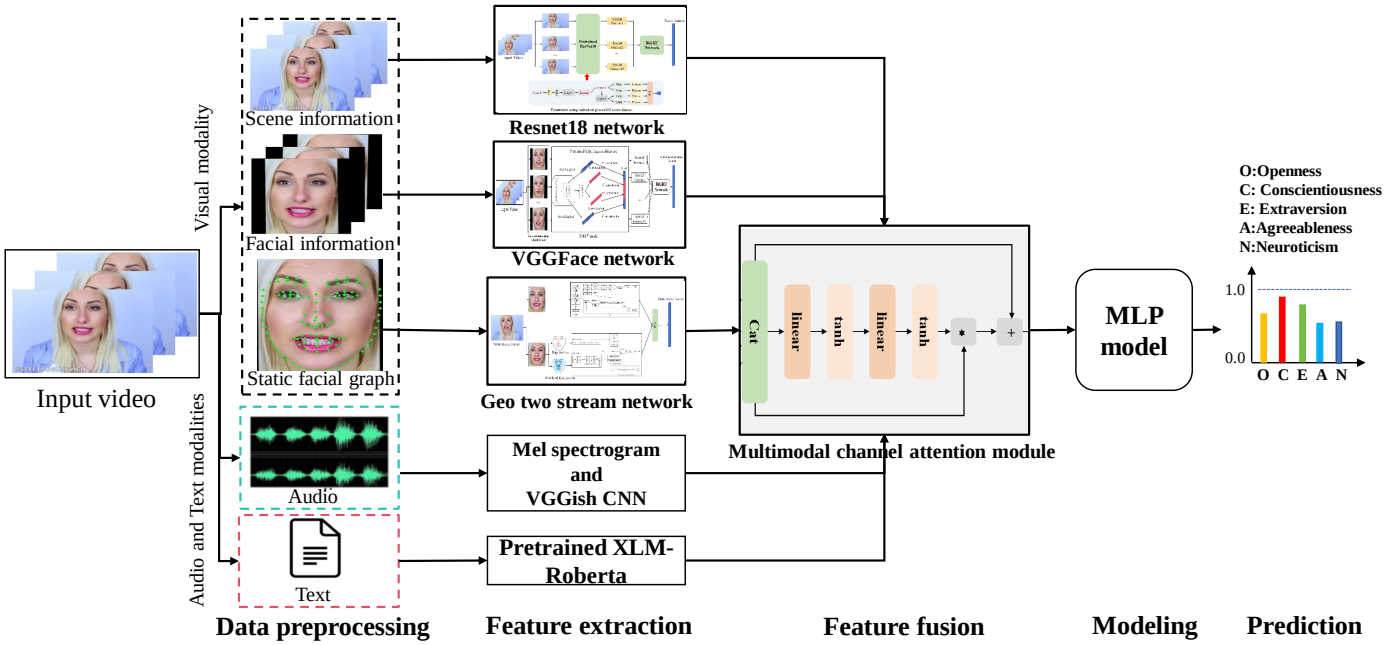


Figure 1. The proposed MFLF-GSL framework is designed for video-based apparent personality analysis. In this framework, a short input video is processed through three modalities: visual (image-based), audio (speech-based), and text (transcript-based). Each modality-specific feature is extracted using dedicated networks tailored to its data type. These learned features are then fused and passed through an MLP regressor to predict personality trait scores.

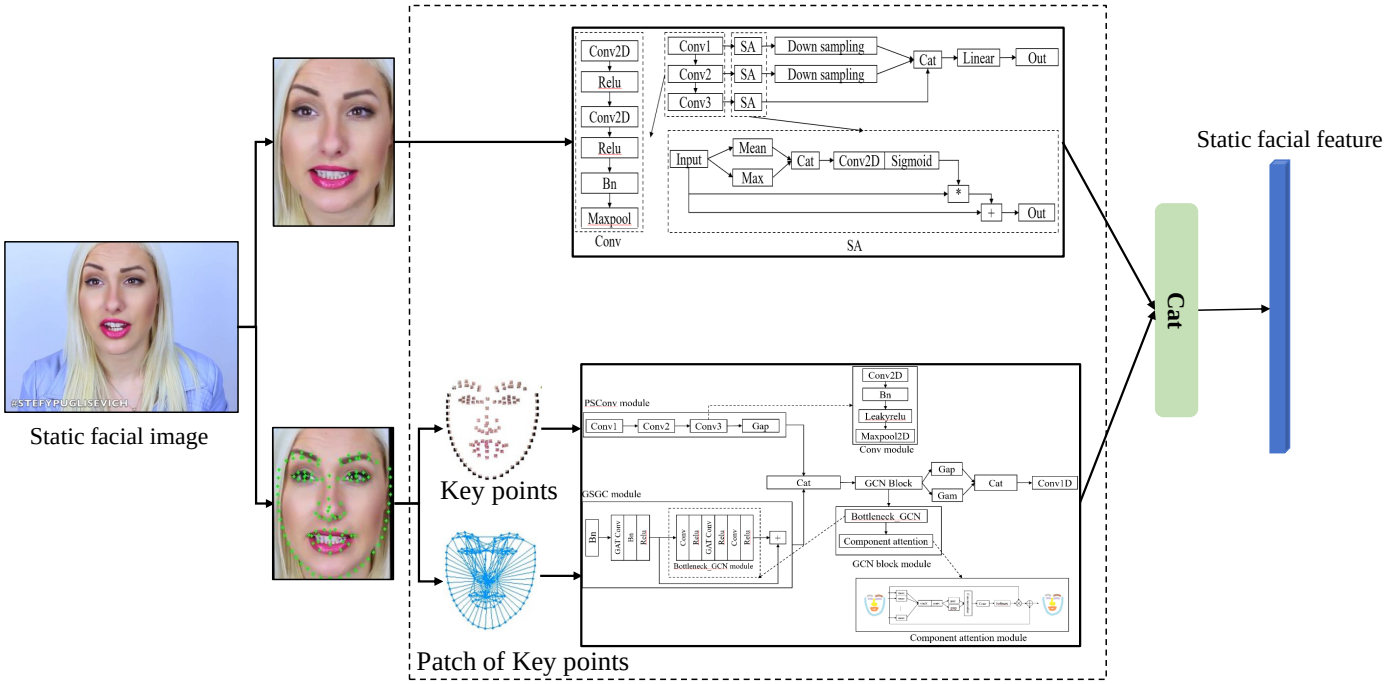


Figure 2. Local Static Facial Appearance and Geometric Features Based on CNN and GCN Architectures. Framework for extracting local static facial appearance and geometric features using CNN and GCN architectures. The input static facial image is processed along two branches: (1) a CNN-based network (top pathway) extracts texture features such as skin, wrinkles, and local appearance cues using convolutional and attention modules; (2) a GCN-based module (bottom pathway) processes facial keypoints and their geometric relationships. Facial landmarks are encoded as graphs to capture the spatial structure of facial components. Outputs from both branches are concatenated to form a comprehensive static facial feature representation.

channel dimension of d_2. The function $F^{(i)}$ represents the mapping function of the i-th CNN module, which extracts P local appearance features to construct a comprehensive representation of the local image F_{local},

formulated as:

$$F_{local} = \{f_{local}^0; f_{local}^1; ...; f_{local}^{P-1}\} \in \mathbb{R}^{d_2 \times P} \quad (5)$$

Finally, the extracted local appearance features F_{local} and geometric features F_{geo} corresponding to the detected keypoints are concatenated to form a unified representation, denoted as

$$F = [F_{geo}, F_{local}] \in \mathbb{R}^{(d_1+d_2) \times P} \quad (6)$$

This representation encapsulates the structural features of the facial image, where the overall feature dimension is $d_1 + d_2$.

Global spatial-temporal scene feature based on Resnet18: To enhance the extraction of spatial-temporal scene features in short videos, as depicted in Figure 3, we utilize the ResNet18 model pre-trained on the Place365 dataset to directly derive scene features from individual video frames. Specifically, each frame is processed sequentially through the pre-trained ResNet18, generating a series of spatial-temporal scene representations. These extracted features are then passed through a BiGRU integrated with a temporal attention-block module, as shown in Figure 4, which selectively emphasizes key temporal scene frames.

Global facial appearance feature based on VGGFace: The study employs VGGFace [12] as a feature extractor to process each frame of facial images. Given that VGGFace is originally developed for facial recognition, it may not be directly suited for predicting the Big Five personality traits. To address this, the model undergoes fine-tuning using facial data from the Big Five dataset. Once optimal fine-tuned parameters are obtained, they are integrated into the VGGFace feature extraction process. Features from individual frames are then extracted using the DAN+ method. Finally, temporal dynamics are captured by leveraging a BiGRU with a temporal attention block, as shown in Figure 4, enabling the extraction of global facial appearance features from video sequences.

2.1.2 *Acoustic modality*

Data Pre-processing: The audio signal is converted into a Log-Mel Spectrogram representation. This transformation is carried out using Python's resample and soundfile libraries, which handle audio data input and output. Variations in speech patterns and intonation can serve as indicators of an individual's Big Five personality traits.

Model Architecture: In this study, we utilize the pre-trained VGGish CNN [10] as the audio feature extractor. The VGGish model processes each audio clip, approximately $\frac{15}{25}$ seconds in length, to generate 128-dimensional feature vectors that capture high-level semantics and meaning. For a video lasting around 15 seconds, this results in a feature matrix of size [25×128]. To capture the temporal dynamics, we employ a BiGRU to extract the final audio features.

2.1.3 *Textual modality*

Data Pre-processing: To process the text and break it down into semantically meaningful units, this model splits the text into binary characters. These characters are then mapped to a predefined dictionary, producing an index sequence that represents the foundational features of the sentence. This sequence serves as the input for subsequent feature extraction.

Model Architecture: The pre-trained XLM-Roberta model [11] is used to extract text features, comprising 12 hidden layers and 12 self-attention heads, which produce an output of 768-dimensional feature vectors.

2.2 Multimodal fusion

In this paper, we introduce a multimodal channel attention module. The five extracted feature vectors are first concatenated into a single feature vector, which is then fed into the module. The module evaluates the contribution of each branch and reduces information redundancy caused by the diversity of features. To prevent information loss, a residual structure is integrated into the module. As illustrated in Figure 5, the module includes two fully connected layers that compute the attention weight α for each dimension in the multimodal representation F. The attention weight α can be calculated as follows:

$$\alpha = tanh(W_2 tanh(W_1 F + b) + c \quad (7)$$

where W_1, W_2, b and c represent the weight matrices and bias of the two fully connected layers, respectively. $tanh$ is used to confine the attention weight within the interval $[-1, 1]$, and then the obtained attention vector α to perform element multiplication with each dimension of the multimodal vector F. The formula is as follows:

$$F^{'} = F \cdot \alpha + F \quad (8)$$

where $F^{'}$is the output of the module, which is fed into MLP to predict the five major personality traits.

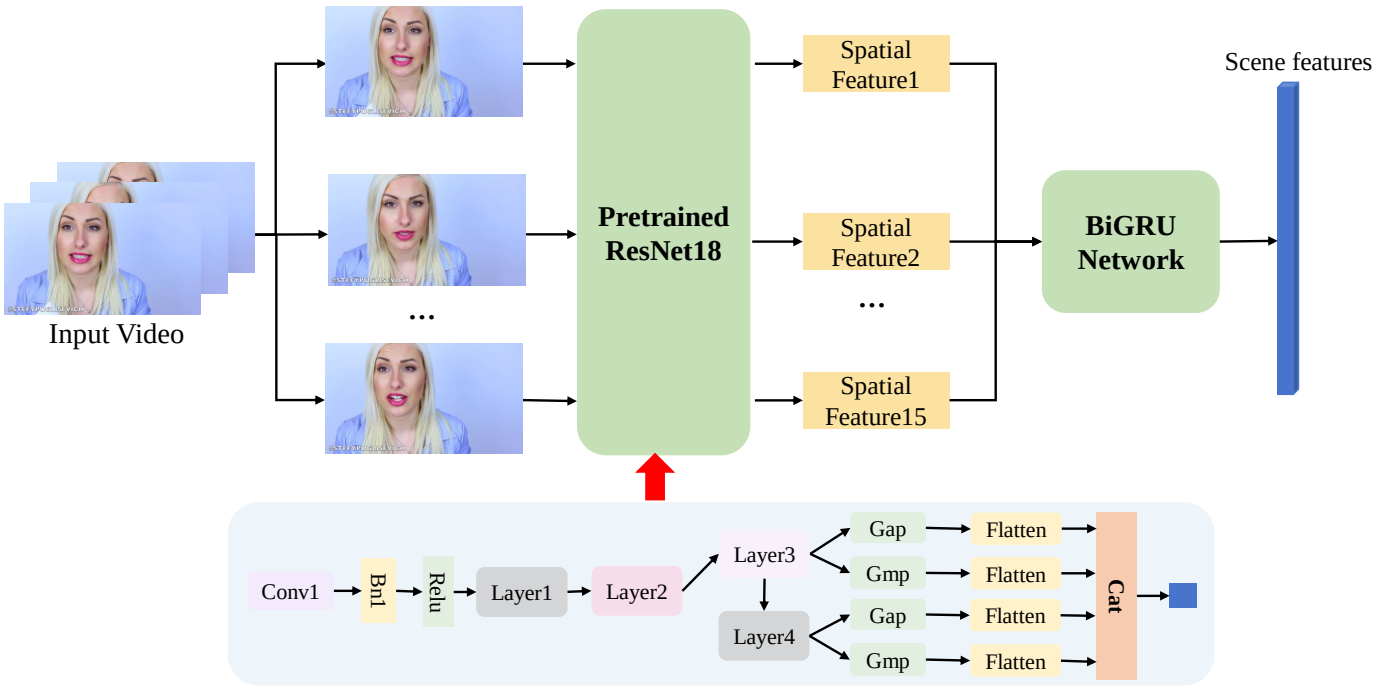


Figure 3. Global spatial-temporal scene feature based on Resnet18. Architecture for extracting global spatial-temporal scene features from an input video using a ResNet18 and BiGRU pipeline. The video is divided into a sequence of 15 frames, each passed through a pretrained ResNet18 network (trained on the Places365 scene dataset) to extract spatial features. These frame-level spatial features are then fed into a BiGRU network to model temporal dependencies and generate a unified scene-level representation. The bottom subfigure shows the internal structure of ResNet18, including convolutional and pooling layers, and the use of global average and max pooling (GAP, GMP) for multi-level feature flattening and concatenation.

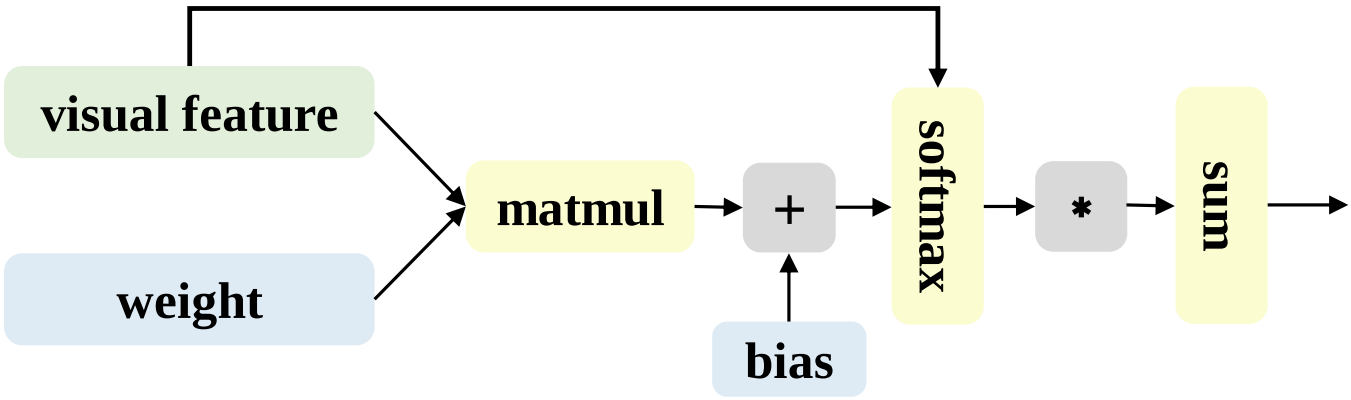


Figure 4. Temporal attention block module. Structure of the temporal attention block module. The visual feature is first linearly projected via a matrix multiplication with a learnable weight, followed by the addition of a bias term. The resulting scores are normalized using a softmax function to obtain attention weights, which are then applied to the original visual feature via element-wise multiplication. The weighted features are finally aggregated using a summation operation, enabling the model to focus on temporally relevant information across frames.

2.3 Model training

In this paper, a three-layer perceptron (MLP) is employed as the model for Big Five personality prediction. The activation of the first two linear layers uses the ReLU activation function, with a dropout layer added following each activation to prevent overfitting. The final layer employs a sigmoid activation function to ensure the predicted output is within the range of $[0, 1]$.

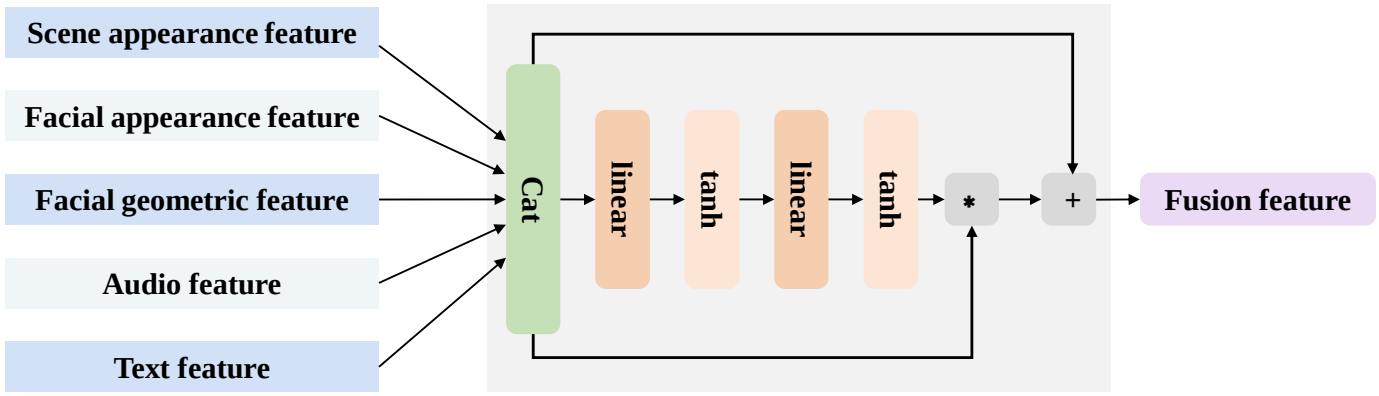


Figure 5. Multimodal channel attention module. Multimodal channel attention module for feature fusion. Inputs from five different modalities—scene appearance, facial appearance, facial geometric structure, audio, and text—are concatenated into a unified feature vector. This concatenated representation is passed through a series of linear and tanh activation layers to compute attention weights. These weights are applied to the original concatenated features through element-wise multiplication. A residual connection adds the weighted features back to the original input, producing the final fused feature vector that captures cross-modal dependencies and highlights informative channels across modalities.

2.4 Loss function

We jointly optimize the entire framework using a multi-task loss function, which can be defined as follows:

$$Loss_{Total} = Loss_{RMSE} + Loss_{logCosh} + Loss_{Bell} \quad (9)$$

where $Loss_{RMSE}$ is the RMSE loss, $Loss_{logCosh}$ is log-Cosh loss, and $Loss_{Bell}$ is the Bell Loss.

3 Experiments

3.1 Experimental Setting

Implementation details: Our experiments are performed on a system equipped with an NVIDIA GeForce RTX 3090 GPU, utilizing PyTorch for model training. For the optimization process, the Geo two-stream network is trained using Stochastic Gradient Descent (SGD), while the other modalities are optimized with the Adam optimizer. The Geo two-stream network starts with an initial learning rate of 0.1, whereas the remaining models are initialized with a learning rate of 0.0001. Both weight decay and momentum parameters are set to 0.0001 and the default value.

Table 1. Summary of key hyperparameter settings.

Model	Optimizer	Learning Rate	Batch Size
Geo Two-Stream	SGD	0.1	32
Facial Stream	Adam	1e-4	32
Audio/Text Module	Adam	1e-4	32
Scene Branch (ResNet18)	Adam	1e-4	32

Dataset analysis: The CFI dataset annotations exhibit certain imbalances in the distribution of personality trait scores. For example, traits such as *Agreeableness* and *Conscientiousness* are skewed toward higher values, while others like *Neuroticism* show a more uniform distribution. To mitigate potential biases during training, all trait scores are normalized using z-score normalization. Moreover, we adopt regression-based evaluation metrics to reflect the continuous nature of the ground truth, and ensure that the training, validation, and test splits maintain similar statistical distributions across all five traits.

Datasets and evaluation criteria: To evaluate the performance of the proposed approach, we use the ChaLearn First Impression-V2 (CFI) dataset [20], which is publicly available and focuses on personality and interview prediction. The dataset is split into three subsets: the training set consisting of 6,000 videos, the validation set with 2,000 videos, and the test set, also containing 2,000 videos. Notably, the dataset's 10,000 videos are annotated with continuous ground-truth values by workers on Amazon Mechanical Turk.

For evaluation, the mean accuracy is used as the main quantitative metric across all the Big Five personality traits.

$$Mean\ Accuracy = 1.0 - \frac{1}{5N}\sum_{j=1}^{5}\sum_{i=1}^{N}|p_{ij} - \hat{p}_{ij}| \quad (10)$$

In the above equation, p_{ij} is the ground truth for the ith sample and jth trait, and $\hat{p}_{ij}$ is the predicted value

Table 2. Comparison with state-of-the-art methods on the first impressions dataset (Validation set).

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. [13]	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. [14]	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 3. Comparison with state-of-the-art methods on the first impressions dataset (Testing set).

Methods	Modalities	Open.	Cons.	Extr.	Agre.	Neur.	ACC(mean)
Güçlütürk et al. [15]	Visual and audio	0.9110	0.9140	0.9110	0.9100	0.9090	0.9110
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
Kaya et al. [2]	Visual and audio	0.9170	0.9198	0.9213	0.9137	0.9146	0.9173
Zhang et al. [16]	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. [13]	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
Vo et al. [17]	Visual, audio and text	0.8864	0.8772	0.8816	0.8958	0.8814	0.8845
Principi et al. [18]	Visual and audio	0.9167	0.9224	0.9159	0.9149	0.9134	0.9167
Escalante et al. [7]	Visual, audio and text	-	-	-	-	-	0.9161
Suman et al. [6]	Visual, audio and text	0.9111	0.9192	0.9173	0.9132	0.9103	0.9143
Zhao et al. [19]	Visual, audio and text	-	-	-	-	-	0.9167
Our method	Visual, audio and text	0.9169	0.9206	0.9189	0.9139	0.9136	**0.9168**

for that sample. N represents the total number of data samples present in the dataset.

Hyperparameter tuning: Key hyperparameters, such as learning rate, batch size, and weight decay, were tuned using grid search based on the validation set performance. For the Geo two-stream network, we tested learning rates in {0.01, 0.05, 0.1}, and selected 0.1 as it yielded the best convergence. Other models used Adam optimizer with learning rates tested in {1e-5, 5e-5, 1e-4}, where 1e-4 was found optimal. Batch size was fixed to 32 for all models to balance performance and memory consumption. A summary of key hyperparameter settings is provided in Table 1.

3.2 Comparison with the state-of-the-art

In this section, we compare the proposed framework with other advanced methods. Tables 2 and 3 display the performance comparison of the proposed framework with state-of-the-art methods on the CFI dataset for the validation and testing sets, respectively. From the experimental results, it can be seen that our framework achieved average Big Five personality prediction accuracy of 91.85% and 91.68% on the validation and test sets, respectively, which is significantly better than other excellent comparison methods.

3.3 Effectiveness of different feature fusion methods

Given that not all modal features contribute positively to predicting the Big Five personality traits, we introduce the Channel Attention Feature Fusion Module (Channel Attn). As shown in Table 4, our Channel Attn outperforms other fusion approaches, highlighting its ability to effectively assign weights to each modality, prioritize crucial modal features, and seamlessly integrate diverse modal information.

3.4 Effectiveness of the temporal attention-block module

Table 5 shows a comparison of the performance with and without the proposed temporal attention block module. The findings demonstrate that integrating the

temporal attention block improves our framework's performance, highlighting the importance of key temporal frames.

Table 4. Performance of different feature fusion strategy (Validation set).

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

Table 5. Comparison of network with temporal attention-block (Validation set).

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

Table 6. Performance comparison between independent and shared CNN modules for local patch feature extraction in the Geo two-stream network. Results are reported as average accuracy (%) on the validation set.

Method	Architecture Type	Accuracy (%)
Shared CNN Modules	Shared weights across P patches	89.47
Independent CNN Modules (Ours)	Separate weights for each patch	**91.85**

3.5 Design rationale for independent CNN modules

To extract discriminative local appearance features from facial regions centered around keypoints, we adopt P independent CNN modules, each specialized for a specific facial patch. While these modules share the same architecture, they do not share parameters. This design is motivated by the observation that different facial regions (e.g., eyes, mouth, nose) exhibit distinct texture patterns and contribute unevenly to personality perception. Parameter sharing across patches could lead to underfitting and diluted region-specific representations, thereby limiting the model's ability to capture subtle variations critical for apparent personality analysis.

Although this design introduces increased computational cost compared to a shared-weight architecture, the modules are lightweight and operate on small local patches, thus keeping the overall resource usage manageable. Moreover, this independence allows the network to learn patch-specific filters, enabling finer-grained attention to personality-relevant micro-expressions. As shown in Table 6, replacing independent modules with a shared-weight variant results in a performance drop, which highlights the benefit of region-specific learning.

4 Conclusion

In this paper, we introduce a novel multi-modal feature learning framework aimed at predicting apparent personality traits. Our approach is built around a graph structure learning network enhanced with an attention mechanism, which uses learned static facial geometric features to boost performance. The framework begins by utilizing pre-trained models to extract features from multiple modalities, including visual global scene appearance, global facial appearance, audio, and text. These features are then further enriched with Bi-GRU/LSTM layers, incorporating a temporal attention block to capture crucial time-dependent information. The integrated features are processed through a multimodal channel attention module, which feeds into an MLP regression model to produce the final prediction. We conduct comprehensive evaluations on the CFI dataset, demonstrating that our framework achieves strong performance.

While the proposed graph-driven multimodal framework demonstrates strong performance in apparent personality assessment, several promising directions remain open for exploration. First, we plan to investigate more efficient graph construction strategies, such as dynamic or self-evolving graphs, to better capture interpersonal variability and contextual cues. Second, incorporating large-scale pretraining for multimodal alignment (e.g., cross-modal contrastive learning) could further enhance generalization to diverse domains. Moreover, extending the framework to multi-language scenarios and real-world deployment, such as real-time feedback in virtual interviews or social robots, represents a valuable avenue for practical impact. Finally, a deeper analysis of the interpretability and fairness of the learned personality representations will be an important step toward responsible AI applications in personality computing.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

Chengwei Ye is an employee of Homesite Group Inc, GA 30043, United States and Huanzhen Zhang is an employee of Payment Department, Chewy Inc, MA 02210, United States.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Wei, X. S., Zhang, C. L., Zhang, H., & Wu, J. (2017). Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, *9*(3), 303-315. [CrossRef]

[2] Kaya, H., Gurpinar, F., & Ali Salah, A. (2017). Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1-9). [CrossRef]

[3] Güçlütürk, Y., Güçlü, U., Baro, X., Escalante, H. J., Guyon, I., Escalera, S., ... & Van Lier, R. (2017). Multimodal first impression analysis with deep residual networks. *IEEE Transactions on Affective Computing*, *9*(3), 316-329. [CrossRef]

[4] Ventura, C., Masip, D., & Lapedriza, A. (2017). Interpreting cnn models for apparent personality trait regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 55-63). [CrossRef]

[5] Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., & Mittal, A. (2016). Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14* (pp. 337-348). Springer International Publishing. [CrossRef]

[6] Suman, C., Saha, S., Gupta, A., Pandey, S. K., & Bhattacharyya, P. (2022). A multi-modal personality prediction system. *Knowledge-Based Systems*, *236*, 107715. [CrossRef]

[7] Escalante, H. J., Kaya, H., Salah, A. A., Escalera, S., Güçlütürk, Y., Güçlü, U., ... & Van Lier, R. (2020). Modeling, recognizing, and explaining apparent personality from videos. *IEEE Transactions on Affective Computing*, *13*(2), 894-911. [CrossRef]

[8] Mujtaba, D. F., & Mahapatra, N. R. (2021, December). Multi-task deep neural networks for multimodal personality trait prediction. In *2021 international conference on computational science and computational intelligence (CSCI)* (pp. 85-91). IEEE. [CrossRef]

[9] Agrawal, T., Agarwal, D., Balazia, M., Sinha, N., & Bremond, F. (2021). Multimodal personality recognition using cross-attention transformer and behaviour encoding. *arXiv preprint arXiv:2112.12180*.

[10] Diwakar, M. P., & Gupta, B. (2024, January). VGGish Deep Learning Model: Audio Feature Extraction and Analysis. In *International Conference on Data Management, Analytics & Innovation* (pp. 59-70). Singapore: Springer Nature Singapore. [CrossRef]

[11] Qu, S., Yang, Y., & Que, Q. (2021). Emotion Classification for Spanish with XLM-RoBERTa and TextCNN. In *IberLEF@ SEPLN* (pp. 94-100).

[12] Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association.

[13] Eddine Bekhouche, S., Dornaika, F., Ouafi, A., & Taleb-Ahmed, A. (2017). Personality traits and job candidate screening via analyzing facial videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 10-13). [CrossRef]

[14] Gürpinar, F., Kaya, H., & Salah, A. A. (2016, December). Multimodal fusion of audio, scene, and face features for first impression estimation. In *2016 23rd International conference on pattern recognition (ICPR)* (pp. 43-48). IEEE. [CrossRef]

[15] Güçlütürk, Y., Güçlü, U., van Gerven, M. A., & van Lier, R. (2016). Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14* (pp. 349-358). Springer International Publishing. [CrossRef]

[16] Zhang, C. L., Zhang, H., Wei, X. S., & Wu, J. (2016, October). Deep bimodal regression for apparent personality analysis. In *European conference on computer vision* (pp. 311-324). Cham: Springer International Publishing. [CrossRef]

[17] Vo, N. N., Liu, S., He, X., & Xu, G. (2018). Multimodal mixture density boosting network for personality mining. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part I 22* (pp. 644-655). Springer International Publishing. [CrossRef]

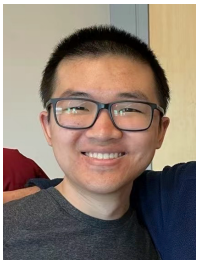
[18] Principi, R. D. P., Palmero, C., Junior, J. C. J., & Escalera, S. (2019). On the effect of observed subject biases in apparent personality analysis from audio-visual signals. *IEEE Transactions on Affective Computing*, *12*(3), 607-621. [CrossRef]

[19] Zhao, X., Liao, Y., Tang, Z., Xu, Y., Tao, X., Wang, D., ... & Lu, H. (2023). Integrating audio and visual modalities for multimodal personality trait recognition via hybrid deep learning. *Frontiers in Neuroscience*, *16*, 1107284. [CrossRef]

[20] Ponce-López, V., Chen, B., Oliu, M., Corneanu,

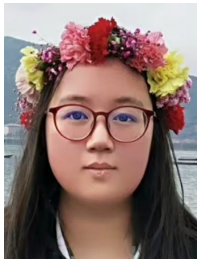
C., Clapés, A., Guyon, I., ... & Escalera, S. (2016). Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14* (pp. 400-418). Springer International Publishing. [CrossRef]

Kangsheng Wang received the B.S. degree in Big Data Management and Applications from University of Science and Technology Beijing, China, in 2019. He is currently pursuing a Master's degree in Electronic Information at the School of Computer and Communication Engineering, University of Science and Technology Beijing, majoring in Computer Technology. (Email: jackie@xs.ustb.edu.cn)



Chengwei Ye received the B.S. degree in Automation from Tongji University and the M.S. degree in Data Science from Worcester Polytechnic Institute, MA 01609, United States, in 2020. (Email: chengweiye18@gmail.com)

Huanzhen Zhang received the B.E. degree in Automation from Changsha University of Science and Technology and the M.S. degree in Computer Science from Worcester Polytechnic Institute, MA 01609, United States, in 2020. (Email: hzhang2@chewy.com)



Linuo Xu received the B.S. degree in Big Data Management and Applications from Dongbei University of Finance and Economics, China, in 2020. She is currently pursuing a Master's degree in Engineering at the School of Information, Yunnan University of Finance and Economics, Yunnan 650000, China.

Shuyan Liu received the B.S. degree in Auto from University of Science and Technology Beijing, China, in 2020. He is currently pursuing a Master's degree in Engineering at the School of Information Science and Technology, Yunnan University, Yunnan 650000, China. (Email: 12024115044@stu.ynu.edu.cn)