

Pose as a Modality: A Psychology-Inspired Network for Personality Recognition with a New Multimodal Dataset

Bin Tang¹, Ke-Qi Pan², Miao Zheng², Ning Zhou², Jia-Lu Sui², Dandan Zhu¹,
Cheng-Long Deng^{2*}, Shu-Guang Kuai^{2, 3, 4*}

¹School of Computer Science and Technology, Shanghai Institute of Artificial Intelligence for Education, Lab of Artificial Intelligence for Education, East China Normal University, Shanghai, China

²School of Psychology and Cognitive Science, Institute of Brain and Education Innovation, Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, Shanghai Key Laboratory of Brain Functional Genomics, Key Laboratory of Brain Functional Genomics, East China Normal University, Shanghai, China

³Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China

⁴NYU-ECNU Institute of Brain and Cognitive Science, New York University Shanghai, Shanghai, China
bintang@stu.ecnu.edu.cn, ddzhu@mail.ecnu.edu.cn, {cldeng, sgkuai}@psy.ecnu.edu.cn

Abstract

In recent years, predicting Big Five personality traits from multimodal data has received significant attention in artificial intelligence (AI). However, existing computational models often fail to achieve satisfactory performance. Psychological research has shown a strong correlation between pose and personality traits, yet previous research has largely ignored pose data in computational models. To address this gap, we develop a novel multimodal dataset that incorporates full-body pose data. The dataset includes video recordings of 287 participants completing a virtual interview with 36 questions, along with self-reported Big Five personality scores as labels. To effectively utilize this multimodal data, we introduce the Psychology-Inspired Network (PINet), which consists of three key modules: Multimodal Feature Awareness (MFA), Multimodal Feature Interaction (MFI), and Psychology-Informed Modality Correlation Loss (PIMC Loss). The MFA module leverages the Vision Mamba Block to capture comprehensive visual features related to personality, while the MFI module efficiently fuses the multimodal features. The PIMC Loss, grounded in psychological theory, guides the model to emphasize different modalities for different personality dimensions. Experimental results show that the PINet outperforms several state-of-the-art baseline models. Furthermore, the three modules of PINet contribute almost equally to the model's overall performance. Incorporating pose data significantly enhances the model's performance, with the pose modality ranking mid-level in importance among the five modalities. These findings address the existing gap in personality-related datasets that lack full-body pose data and provide a new approach for improving the accuracy of personality prediction models, highlighting the importance of integrating psychological insights into AI frameworks.

Introduction

Personality traits are fundamental psychological constructs that influence human behavior and social interactions (Sun et al. 2022). The Big Five personality model (Chen et al.

*Corresponding author.

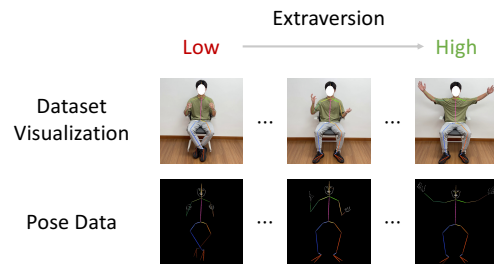


Figure 1: Dataset illustration. The first row shows the visualization results using AlphaPose (Fang et al. 2022), and the second row shows the dataset's pose data.

2021), which includes five dimensions known as OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), is widely recognized and utilized in psychological research. In recent years, the use of machine (deep) learning to analyze multimodal data for detecting the Big Five personality traits has attracted considerable attention from researchers (Leekha et al. 2024; Ryumina et al. 2024). Unlike traditional scale-based measures, multimodal personality recognition offers a more effortless assessment process, reduces the likelihood of user manipulation, and enhances efficiency by simultaneously evaluating multiple aspects (Hickman et al. 2022; Escalante et al. 2020). As a result, multimodal personality recognition has become increasingly important in fields such as psychology, education, and artificial intelligence (Wu et al. 2020).

Despite extensive research in computational personality recognition (Mehta et al. 2020), model accuracy has remained unsatisfactory, making the improvement of predictive performance a key focus for researchers. Psychological research highlights a strong connection between personality traits and pose. Individuals with different personality traits exhibit distinct postural characteristics (Kelly and Ngo Tran 2023; Thomas et al. 2022). For example, those with higher levels of extraversion tend to use broader and faster gestures, along with increased movements of the head, hands, and

legs (Conway et al. 2023). Similarly, varying levels of neuroticism correlate with the relative velocity of pose movements (Koppensteiner 2013). Given that previous studies have largely overlooked the inclusion of pose data in computational personality models, we hypothesize that incorporating such data could significantly enhance the accuracy of personality predictions. However, two major challenges remain.

The first challenge is the lack of datasets that include frontal full-body pose data. Existing datasets related to personality traits either lack pose data (Zafeiriou et al. 2017), capture only upper-body pose (Escalante et al. 2020), provide side-view full-body pose (Giritlioğlu et al. 2021), or collect pose data without the context of speech (Miranda-Correa et al. 2018). Consequently, no dataset currently available offers high-quality full-body pose data suitable for training models.

The second challenge is that current personality prediction models, such as CRNet (Li et al. 2020) and ResNet (Leekha et al. 2024), support inputs from text, speech, and facial modalities, but they do not support pose data. Moreover, existing methods predominantly rely on the visual computing paradigm, performing only simple fusion of visual, audio, and text modalities, but do not consider the different contributions of each modality to each dimension of the Big Five personality traits. Psychological research shows that different dimensions of the Big Five personality traits manifest differently across modalities. For example, judgments of Agreeableness primarily rely on facial data, Conscientiousness on body-related data, and Extraversion on full-body data (Hu and O’Toole 2023; Hu, Baragchizadeh, and O’Toole 2020). Therefore, existing models may not achieve good predictions of the Big Five personality traits.

To address these challenges, this study first constructs a multimodal personality recognition dataset that includes whole-body pose data, along with facial, frame, audio, and textual modalities. The dataset comprises full-body videos of 287 participants responding to 36 specific interview questions, annotated with their self-reported Big Five personality scores. We utilize the widely adopted AlphaPose (Fang et al. 2022) to extract pose information from the visual data (Figure 1). Building on this dataset, we propose the Psychology-Inspired Network (PINet), which consists of three modules: Multimodal Feature Awareness (MFA), Multimodal Feature Interaction (MFI), and Psychology-Informed Modality Correlation Loss (PIMC Loss). The MFA module employs the Vision Mamba Block (Zhu et al. 2024) for the first time to fully capture visual personality features both spatially and temporally. During the MFI process, the Fusion Mamba Block (Xie et al. 2024) efficiently integrates multimodal information while maintaining linear complexity. The PIMC Loss, grounded in psychological theories, guides the model to emphasize different modalities for different personality dimensions. Notably, PINet serves as an innovative and bidirectional bridge between AI and psychology.

In summary, the contributions of this work are as follows:

- We construct a new dataset comprising multimodal data and self-reported Big Five personality scores, addressing the gap in existing personality-related datasets that lack

full-body pose data.

- We introduce a novel multimodal personality recognition model, the Psychology-Inspired Network (PINet). For the first time, PINet incorporates Mamba technology into personality recognition, significantly enhancing the performance of the visual encoder and multimodal fusion modules.
- We propose the innovative Psychology-Informed Modality Correlation Loss (PIMC Loss), which guides the model to emphasize different modalities for various personality dimensions.
- Experimental results show that PINet outperforms seven state-of-the-art baseline models. Additionally, integrating pose data significantly improves the performance of various models in predicting personality traits, confirming its crucial role as a modality closely associated with personality traits.

Related Work

Personality-related Dataset

Currently, several personality-related datasets provide multimodal data, but most are limited to common modalities such as speech, text, and facial expressions. For example, the Essays II (Tausczik and Pennebaker 2010) and MyPersonality (Stillwell and Kosinski 2015) datasets focus solely on text, while the First Impressions V2 (Escalante et al. 2020) and Aff-wild (Zafeiriou et al. 2017) dataset includes face, audio, and text. Although the AMIGOS dataset (Miranda-Correa et al. 2018) includes full-body pose data, it is collected while users are watching videos, which differs significantly from the pose typically exhibited during speech. Therefore, this dataset is not suitable for conventional personality trait modeling and recognition. To date, we have identified only one dataset, SIAP (Giritlioğlu et al. 2021), that offers pose data during speech. This dataset records users from frontal, left-rear, and right-front angles. However, the frontal view only provides upper-body images, missing full-body pose data. Although the left-rear and right-front views offer full-body pose data, these side views compromise the accuracy of pose data, which negatively impacts model performance. The study based on the SIAP shows that pose is the least effective modality in predicting the Big Five personality traits among all modalities. In contrast, our findings indicate that pose data has moderate effectiveness among all modalities, underscoring the importance of high-precision pose data. Additionally, the SIAP dataset is relatively small, containing only 180 video clips. This limited sample size may further affect the model’s performance.

Multimodal Personality Recognition Methods

Benefiting from the rapid development of encoders, there are many multimodal personality recognition methods, such as PersEmoN (Zhang, Peng, and Winkler 2019), SENet (Hu, Shen, and Sun 2018), GatedGCCN (Song et al. 2022), HRNet (Wang et al. 2020), EmoFormer (Ryumina et al. 2024), ResNet (Leekha et al. 2024), and CRNet (Li et al. 2020).

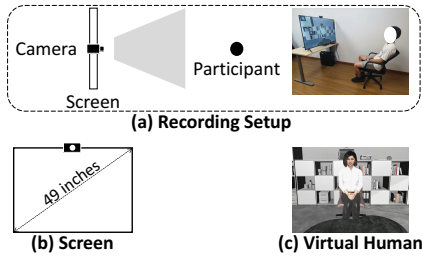


Figure 2: The illustration of the data collection scenarios.

However, these methods either employ a simple CNN or a residual block as the backbone network, or they directly concatenate multimodal features without considering inter-modal correlation. Additionally, previous studies have treated personality recognition merely as a basic classification or regression task, neglecting the internal relationships between the five personality dimensions and multimodal data. With the advent of vision transformers (Han et al. 2022), the performance of encoders has significantly improved. However, their computational complexity has escalated to quadratic levels. This increased complexity constrains their ability to process more image frames under the same computing power, thereby hindering their capability to capture subtle changes in pose and action. In the multimodal feature fusion module, although multi-head attention (Bhajanapalli et al. 2020) considers the interaction between different modalities, it incurs high complexity, leading to more stringent requirements for computing devices.

To this end, we introduce Mamba (Gu and Dao 2023) into the multimodal personality recognition task for the first time and improve both the encoder and the fusion module. Previous studies (Zhu et al. 2024; Yang, Xing, and Zhu 2024) have demonstrated that Mamba can greatly reduce computational complexity. Additionally, we propose the novel PIMC Loss, the first loss function for personality recognition inspired by psychological theory.

Dataset

Participants

This dataset includes data from 287 university students (114 males, 173 females; mean age = 21.65 ± 2.22 years). All participants are native Chinese speakers and have normal or corrected-to-normal visual acuity. They voluntarily take part in the study and sign an informed consent form before the commencement of data collection. The data collection process is approved by the Ethics Committee of the University.

Data Collection and Annotation

Participants engage in a structured interview conducted by a 3D virtual interviewer designed as an Asian woman with a friendly face. Each participant is seated in a room with a 49-inch television around 1.45 meters in front of them. A Logitech BRIO 500 wide-angle camera (Field of View: 90° ; Resolution: 1080P; Frame Rate: 30 fps) is mounted on the top edge of the television, facing the participant to capture their full body, as shown in Figure 2.

Participants are required to answer 36 questions adapted from the NEO-FFI-3 (Meca-Lallana et al. 2019) scale and NEO-PI-R (Costa and McCrae 2008) scale in sequence from the virtual interviewer. Participants are instructed to limit their response to each question to no more than one minute, with the entire interview lasting 30-60 minutes.

After the interview, all participants complete the NEO-FFI-3 scale, which consists of five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, with 12 questions per dimension, totaling 60 questions. Each question is scored from 0 (strongly disagree) to 4 (strongly agree), with the total score for each dimension ranging from 0 to 48. Following previous research (Liao, Song, and Gunes 2024), we normalize the total score for each dimension to a range of 0 to 1. For more statistics on the dataset, please refer to the supplementary material.

Data Preprocessing

The video data contains homogeneous multimodal data, including visual, audio, and textual information. We use FFmpeg to extract the audio from the videos, FunASR (Gao et al. 2023) to obtain the participants’ spoken text and timestamp information. We then utilize these timestamps to align the three modalities. Subsequently, the multimodal data of 287 participants are divided into 10,332 samples ($287 \text{ participants} \times 36 \text{ questions}$).

We use the widely adopted AlphaPose (Fang et al. 2022) to extract pose information from the visual data. Additionally, MTCNN (Zhang et al. 2016) is used to extract facial data. Finally, the data is organized into five modalities: frame, face, pose, audio, and text.

Methodology

Overview

Multimodal personality recognition can be regarded as a multimodal multi-task regression problem. In this section, we introduce the details of our proposed PINet, which serves as a novel and bidirectional bridge between AI and psychology. We also explain how PINet processes multimodal multi-task learning. The proposed PINet consists of three modules: Multimodal Feature Awareness (MFA), Multimodal Feature Interaction (MFI), and Psychology-Informed Modality Correlation Loss (PIMC Loss), as illustrated in Figure 3. Both the MFA and MFI modules are based on the advanced state space model (SSM), i.e., Mamba (Gu and Dao 2023). For details of the SSM, please refer to the supplementary materials.

Multimodal Feature Awareness (MFA)

In the MFA (Encoder) module, there have been many excellent developments, such as the use of residual blocks for feature extraction and the implementation of Vision Transformers to enhance visual feature extraction. Since both audio and text are one-dimensional sequences, their computational load is relatively smaller compared to the three-dimensional sequences of video, and related research has shown that residual blocks perform well. Therefore, this study strictly follows previous researches (Liao, Song, and Gunes 2024;

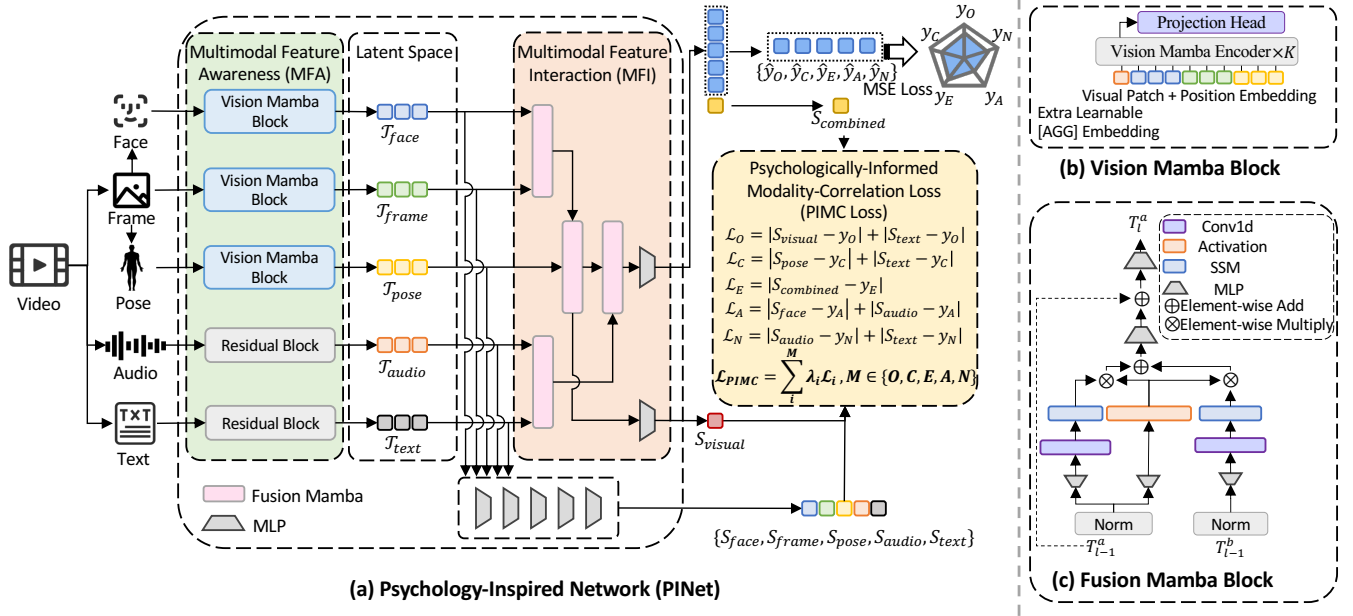


Figure 3: (a) The architecture of our PINet. (b) The Vision Mamba Block in MFA. (c) The Fusion Mamba Block in MFI with the symmetrical parts of the Fusion Mamba Block omitted.

Li et al. 2020) in handling audio and text using residual blocks.

However, the Vision Transformer in the visual modality struggles to capture subtle movements due to its quadratic complexity. The recent Vision Mamba addresses this limitation by extracting visual features with linear complexity. Additionally, the Video Mamba (Chen et al. 2024) further extends the Vision Mamba to achieve 3D video understanding. Based on this, we use the Video Mamba as our Vision Mamba Block (VMB), which is shown in Figure 3 (b).

Specifically, we first use 3D convolution (*i.e.*, $1 \times 16 \times 16$) to project the input videos $X_{video} \in R^{T \times H \times W \times 3}$ into L non-overlapping spatiotemporal patches $X^p \in R^{L \times D}$, where $X_{video} = \{X_{face}, X_{frame}, X_{pose}\}$, $X^p = \{X_{face}^p, X_{frame}^p, X_{pose}^p\}$, $L = t \times h \times w$ ($t = T, h = \frac{H}{16}, w = \frac{W}{16}$) (*i.e.*, L is the length of sequence), D is the dimension of sequence. The sequence of tokens input to the following VMB is:

$$X_{visual} = [X_{AGG}, X^p] + P_s + P_t, \quad (1)$$

where X_{AGG} is the learnable embedding of a special aggregation token whose corresponding output in VMB (X_{out}^0) is used as the aggregated representation for the entire input sequence. Following previous works (Akbari et al. 2021), we add a learnable spatial position embedding $P_s \in R^{(hw+1) \times D}$ and the extra temporal one $P_t \in R^{n \times D}$ to retain the spatiotemporal position information, since the SSM modeling is sensitive to token position. The tokens X are then passed through by K stacked VMBs, and X_{AGG} token at the final layer is fed into the latent space as visual features (T_{visual}):

$$T_{visual} = VMB(X_{visual}), \quad (2)$$

where $VMB(\cdot)$ represents the Vision Mamba Block, $T_{visual} = \{T_{face}, T_{frame}, T_{pose}\}$. It is worth noting that the reason why these features are denoted by \mathcal{T} is that for Multimodal Feature Interaction (MFI), these features are still treated as tokens. And the attention features for audio and text modalities are obtained by:

$$\begin{aligned} T_{audio} &= RB(X_{audio}), \\ T_{text} &= RB(X_{text}), \end{aligned} \quad (3)$$

where $RB(\cdot)$ is the Residual Block, X_{audio} and X_{text} denote the input audios and texts, respectively. The multimodal attention features $\mathcal{T} = \{T_{visual}, T_{audio}, T_{text}\}$ is split into two branches, one branch enters the MFI and the other branch is used to obtain the unimodal score S , which can be formulated as:

$$S_i = MLP(T_i), \quad (4)$$

where $i = \{face, frame, pose, audio, text\}$, $MLP(\cdot)$ represents the MLP with sigmoid. The purpose of obtaining the unimodal score S is to calculate the PIMC Loss.

Multimodal Feature Interaction (MFI)

Multimodal feature fusion has always been a challenging problem. The simplest approach is direct concatenation, which, however, overlooks the complementarity and redundancy between modalities. The multi-head attention mechanism addresses this issue effectively but suffers from quadratic computational complexity, leading to low computational efficiency. Fusion mamba (Xie et al. 2024), on the other hand, offers a solution to the multi-head attention mechanism's challenges by employing linear complexity. Initially, it could handle only two modalities. In this study,

we expand fusion mamba to suit multimodal personality recognition tasks by integrating four fusion mamba blocks into the Multimodal Feature Interaction module.

Particularly, we project the features (tokens) of the two modalities into a shared space, employing a gating mechanism to encourage complementary feature learning while suppressing redundant features. Additionally, to enhance local features, we incorporate depth-wise convolution within the module, thereby amplifying the encoding capability of local features during the fusion process. The generation of \mathcal{T}^a and \mathcal{T}^b follows the process outlined in the Mamba block. Subsequently, we obtain the gating parameter z by projecting \mathcal{T}'_{l-1} and use z to modulate \mathcal{T}^a and \mathcal{T}^b . The output sequence \mathcal{T}'_l is obtained by residual connection and MLP. The schematic for the Fusion Mamba Block is shown in Figure 3 (c). For the pseudocode of the Fusion Mamba Block, please refer to Algorithm 1 in the supplementary material. It is worth noting that Algorithm 1 represents the process of integrating the feature from \mathcal{T}^b into \mathcal{T}^a . In other words, a Fusion Mamba consists of two Fusion Mamba Blocks, which are used to integrate one modality into the other, ensuring symmetry between them. The linear layer is designed to further fuse the outputs of the two Fusion Mamba Blocks, which is omitted in Figure 3 (a).

The output of PINet \hat{y} can be represented by the following formula:

$$\hat{y} = MLP(MFI(\mathcal{T})), \quad (5)$$

where \hat{y} represents the set of predicted values for the five dimensions of the Big Five personality traits $\{\hat{y}_O, \hat{y}_C, \hat{y}_E, \hat{y}_A, \hat{y}_N\}$, $MFI(\cdot)$ represents the Multimodal Feature Interaction.

Psychology-Informed Modality Correlation Loss (PIMC Loss)

Previous studies treat the personality recognition task merely as a regular regression task and ignore the correlations between modalities and personality dimensions. However, psychological studies have shown that different dimensions of the Big Five personality traits may exhibit significant influence in one or more specific modalities rather than uniformly across all modalities (Hu and O'Toole 2023; Hu, Baragchizadeh, and O'Toole 2020).

Therefore, we aim to design a loss function that guides the model to focus on different modalities for different personality dimensions during training, rather than treating all modalities equally. We customize the loss for each Big Five personality dimension based on its definition. Specifically, we hypothesize that openness, which reflects an individual's acceptance of new ideas, experiences, and cultural diversity, is primarily captured by visual and text modalities. Conscientiousness, related to goal-setting, planning, self-discipline, and execution, is hypothesized to be determined by pose and text. Extroversion, indicating the degree of focus on the external world, sociability, and emotional expression, is believed to be determined by the combined features of the individual. Agreeableness, which describes how an

individual interacts with others, being prosocial and cooperative, is felt to be represented by facial expressions and voice. Neuroticism, involving emotional stability and responsiveness, is thought to be mainly determined by pitch and speech content. Based on these assumptions, we design the Psychology-Informed Modality Correlation Loss (PIMC Loss), which includes five loss functions $\mathcal{L}_O, \mathcal{L}_C, \mathcal{L}_E, \mathcal{L}_A,$ and \mathcal{L}_N , for the five dimensions are simply defined as follows:

$$\begin{aligned} \mathcal{L}_O &= MAE(S_{visual}, y_O) + MAE(S_{text}, y_O), \\ \mathcal{L}_C &= MAE(S_{pose}, y_C) + MAE(S_{text}, y_C), \\ \mathcal{L}_E &= MAE(S_{combined}, y_E), \\ \mathcal{L}_A &= MAE(S_{face}, y_A) + MAE(S_{audio}, y_A), \\ \mathcal{L}_N &= MAE(S_{audio}, y_N) + MAE(S_{text}, y_N), \end{aligned} \quad (6)$$

where $MAE(\cdot)$ represents the mean absolute error, $S_i (i = \{face, frame, pose, audio, text\})$ represents the prediction score from unimodal data, $y_i (i = \{O, C, E, A, N\})$ represents the ground truth for the five dimensions of the Big Five personality traits. The S_{visual} and $S_{combined}$ represent the prediction scores from the visual modality and all modalities combined, respectively. They are cleverly obtained through the MFI mechanism, as shown in Figure 3 (a). The PIMC Loss is defined as:

$$\mathcal{L}_{PIMC} = \sum_i^M \lambda_i \mathcal{L}_i, M \in \{O, C, E, A, N\}, \quad (7)$$

where λ_i represents learnable parameters for each modality. The loss function for the whole task is:

$$\mathcal{L}_{MSE} = MSE(\hat{y}, y), \quad (8)$$

where $MSE(\cdot)$ represents the mean square error. The final loss function \mathcal{L} is defined as follows:

$$\mathcal{L} = \mathcal{L}_{PIMC} + \mathcal{L}_{MSE}. \quad (9)$$

Through the design of PIMC Loss, multimodal personality recognition is approached as multi-task learning. Each task is highly correlated and cooperates effectively to enhance the overall model performance, a finding confirmed by the results of ablation experiments with PIMC Loss.

Experiments

Experimental Details

To fairly compare the performance of different models, all experiments are strictly based on the open-source audio-visual benchmark (Liao, Song, and Gunes 2024), with only the addition of text modality processing. In particular, several models that perform well on the self-reported dataset are used as baselines including CRNet (Li et al. 2020), ResNet (Leekha et al. 2024), GatedGCN (Song et al. 2022), EmoFormer (Ryumina et al. 2024), HRNet (Wang et al. 2020), SENet (Hu, Shen, and Sun 2018), and PersEmon (Zhang, Peng, and Winkler 2019). The dataset is divided into training, validation, and test sets in a ratio of approximately 3:1:1 based on participants, with 6,228 samples (173 participants),

2,052 samples (57 participants), and 2,052 samples (57 participants), respectively. We use the stochastic gradient descent (SGD) (Li 2017) with an initial learning rate of 0.001, momentum of 0.9, and weight decay of 0.005. All experiments are conducted on the PyTorch platform using six Nvidia A6000 GPUs.

Evaluation Metrics

To evaluate the performance of the trained models, we employ four common evaluation metrics: Pearson Correlation Coefficient (PCC), Concordance Correlation Coefficient (CCC), Accuracy (ACC), and Mean Squared Error (MSE) (Song et al. 2021; Liao, Song, and Gunes 2024). For PCC, CCC, and ACC, higher values indicate better performance, whereas for MSE, a lower value is preferred.

Comparison with Baselines

Model	PCC \uparrow	CCC \uparrow	ACC \uparrow	MSE \downarrow
PersEmon	0.0359	-0.0212	0.8575	0.0320
SENet	0.0567	0.0549	0.8628	0.0297
GatedGCN	0.0777	0.0345	<u>0.8651</u>	0.0289
HRNet	0.0924	0.0284	0.8557	0.0327
EmoFormer	0.1046	0.0320	0.8640	0.0293
ResNet	0.1054	0.0542	0.8506	0.0354
CRNet	<u>0.1141</u>	<u>0.1124</u>	0.8577	0.0321
Our PINet	0.2509	0.2706	0.8679	0.0277

Table 1: Performance comparison of our proposed PINet with baselines. The **bold** result indicates the best performance, while the underlined result indicates the second-best performance.

Table 1, Table 2, and Table 3 present the performance comparison between PINet and other methods on our dataset. Table 1 shows the average results across the five personality trait dimensions, while Table 2 and Table 3 provide detailed results for each dimension of the traits.

The average results across the five dimensions demonstrate that our PINet model significantly outperforms other baseline models across all four evaluation metrics. Notably, the PCC and CCC scores of PINet are higher than those of the second-best model, CRNet, by 0.1368 and 0.1508, respectively, representing improvements of 119.9% and 134.2% (Table 1). These findings are further supported by the detailed results for each dimension of the traits. Among all models, PINet achieves the highest PCC and CCC scores across all five dimensions (Table 2), and it also achieves the best performance in ACC for four dimensions (Open, Consc, Extrav, Neuro) and in MSE for five dimensions (Table 3). These results indicate that, our proposed PINet model significantly improves the accuracy of predicting the traits. For the experimental results of PINet and the baselines on the public dataset, please refer to the supplementary material.

Ablation Study

We further conduct ablation studies to assess the importance of PINet’s three modules. To evaluate the significance of the VMB module, we replace it with a conventional Vision

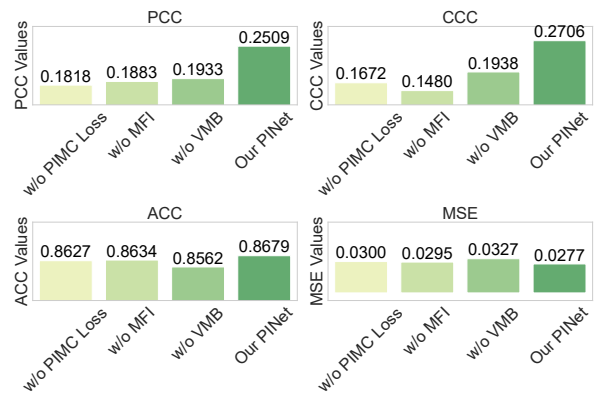


Figure 4: Ablation study with PIMC Loss, MFI and VMB.

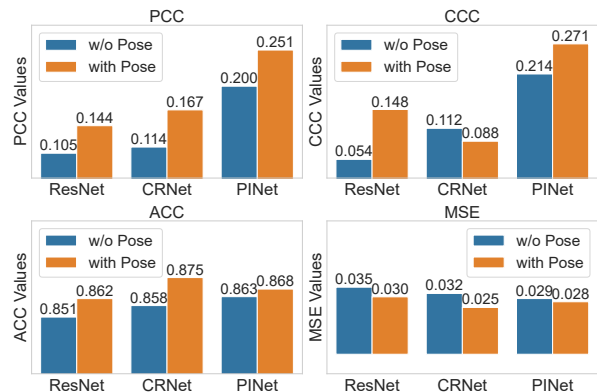


Figure 5: Ablation study with pose data.

Transformer (w/o VMB) while keeping the other modules of PINet unchanged. Similarly, to assess the MFI module, we replace it with a simple concatenation method (w/o MFI). For evaluating the PIMC Loss module, we directly remove it from the model (w/o PIMC Loss). The average results across the five dimensions of the traits (as shown in Figure 4) show that, compared to the PINet, removing any of the three modules leads to a significant drop in PCC and CCC values, a slight decrease in ACC, and an increase in MSE. Moreover, the values of the three modules are quite close in PCC, ACC, and MSE. The results for each dimension are similar to the average results across all dimensions (please refer to supplementary Table 1 and Table 2). These findings demonstrate that all three modules are crucial for improving the prediction of the traits and that their contributions are comparable.

In addition, we conduct another ablation study to evaluate the contribution of pose data to prediction accuracy using the common ResNet, CRNet, and our PINet. We compare the prediction results of models with and without pose data for the traits. The average results across the five dimensions are shown in Figure 5. For all three models, those with pose data significantly outperform those without pose data in PCC, CCC, ACC, and MSE, except for the CCC for CRNet. Notably, in terms of PCC, the values of the PINet, CRNet, and ResNet models with pose data increase by 36.8%,

Model	Pearson Correlation Coefficient (PCC) \uparrow					Concordance Correlation Coefficient (CCC) \uparrow				
	Open	Consc	Extrav	Agree	Neuro	Open	Consc	Extrav	Agree	Neuro
PersEmon	0.0333	0.0508	0.0192	0.0448	0.0316	0.0002	0.0431	-0.1732	0.0344	-0.0103
SENet	0.0306	0.0019	0.1176	0.0468	0.0864	0.0296	0.0018	0.1139	0.0453	0.0837
GatedGCN	0.0403	0.0818	0.1183	0.0579	0.0903	0.0197	0.0184	0.0681	0.0393	0.0267
HRNet	0.0619	0.0885	0.1193	0.1323	0.0599	0.0190	0.0272	0.0367	0.0407	0.0184
EmoFormer	0.0584	0.1139	0.1097	0.1487	0.0921	0.0179	0.0359	0.0387	0.0491	0.0183
ResNet	0.0288	<u>0.1159</u>	0.1164	<u>0.1522</u>	<u>0.1138</u>	0.0148	0.0596	0.0598	0.0783	0.0585
CRNet	0.1258	0.0885	0.1334	0.1512	0.0747	0.1239	0.0842	0.1314	0.1489	0.0736
Our PINet	0.3335	0.1463	0.3146	0.2788	0.1812	0.3447	0.1907	0.3184	0.2821	0.2172

Table 2: PCC and CCC values of our PINet and baseline models for predicting each dimension of the personality traits.

Model	Accuracy (ACC) \uparrow					Mean Squared Error (MSE) \downarrow				
	Open	Consc	Extrav	Agree	Neuro	Open	Consc	Extrav	Agree	Neuro
PersEmon	0.8526	0.8482	0.8790	0.8515	0.8562	0.0339	0.0360	0.0227	0.0345	0.0328
SENet	0.8557	0.8630	0.8789	0.8541	0.8623	0.0328	0.0301	0.0230	0.0334	0.0294
GatedGCN	<u>0.8591</u>	<u>0.8692</u>	0.8793	0.8535	<u>0.8644</u>	0.0310	0.0304	0.0219	<u>0.0323</u>	<u>0.0291</u>
HRNet	0.8514	0.8531	0.8771	0.8430	0.8538	0.0341	0.0343	0.0233	0.0388	0.0331
EmoFormer	0.8535	0.8646	<u>0.8824</u>	0.8563	0.8631	0.0313	<u>0.0292</u>	<u>0.0216</u>	0.0340	0.0305
ResNet	0.8440	0.8438	0.8722	0.8409	0.8521	0.0383	0.0379	0.0261	0.0396	0.0354
CRNet	0.8525	0.8505	0.8773	0.8578	0.8503	0.0340	0.0355	0.0240	0.0324	0.0347
Our PINet	0.8622	0.8694	0.8857	<u>0.8572</u>	0.8649	0.0303	0.0267	0.0209	0.0321	0.0283

Table 3: ACC and MSE values of our PINet and baseline models for predicting each dimension of the personality traits.

45.9%, and 25.5%, respectively, compared to models without pose data. For CCC, the values of the PINet and ResNet models increase by 172.5% and 26.3%, respectively. These findings are further supported by the results for each dimension of the traits (please refer to supplementary Table 3 and Table 4). These results demonstrate the importance of pose data in improving the prediction of the traits.

Furthermore, we examine the relative importance of the pose modality compared to other modalities, as shown in Figure 6. We train the revised PINet model, without the MFI and PIMC Loss modules, using only a single modality (face, frame, pose, audio, text) at a time and calculated the PCC for each modality model. We then calculate the relative PCC by dividing them by the PCC value obtained when training PINet with all modalities, yielding values between 0 and 1. The results show that pose’s average relative PCC ranked third, higher than face and frame but lower than text and audio. For each dimension of the traits, the relative PCC of pose ranks third for Open, first for Consc, tied for fourth for Extra, fourth for Agree, and third for Neuro. These findings suggest that the importance of pose is moderate among all modalities.

Conclusion and Future Work

In this study, we first create a dataset consisting of frontal full-body video recordings of 287 participants during interviews, along with self-reported Big Five personality scores as labels. This dataset addresses the gap in existing personality-related datasets that need more high-quality whole-body pose data. Second, to better utilize multimodal data such as pose, we propose a new network called PINet, which consists of three modules: MFI, VMB, and PIMC Loss, based on findings from psychological research. The PINet network assigns different weights to different modal-

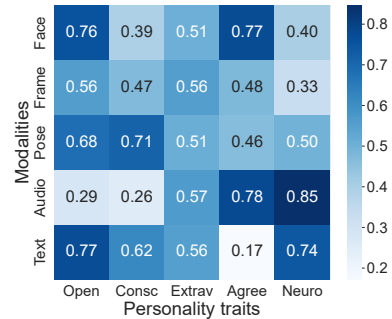


Figure 6: The relative PCC values of the single modality to each personality trait.

ities when predicting each dimension of the Big Five personality traits. Experimental results demonstrate that our proposed PINet outperforms other state-of-the-art baseline models, and the three modules of PINet contribute similarly to the overall performance, which indicates that the psychology-based PINet model can significantly enhance the accuracy of Big Five personality trait predictions. Additionally, the pose modality significantly improves the accuracy of Big Five personality trait predictions. Among the five modalities (face, frame, pose, audio, text), the importance of the pose modality ranks in the middle.

In future research, we plan to further expand the dataset by collecting observer-reported Big Five personality scores to verify the importance of the pose modality and investigate the differences in model prediction between self-reported and observer-reported personality traits. Additionally, we will test our PINet model with the observer-reported label to validate the generalizability of our PINet.

Acknowledgments

This work is supported by the National Science and Technology Innovation 2030 Major Program [grant number 2022ZD0205103], the National Natural Science Foundation of China [grant number 32100879], and the Fundamental Research Funds for the Central Universities.

References

- Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui, Y.; and Gong, B. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34: 24206–24221.
- Bhojanapalli, S.; Yun, C.; Rawat, A. S.; Reddi, S.; and Kumar, S. 2020. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*, 864–873. PMLR.
- Chen, G.; Huang, Y.; Xu, J.; Pei, B.; Chen, Z.; Li, Z.; Wang, J.; Li, K.; Lu, T.; and Wang, L. 2024. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*.
- Chen, X.; He, J.; Swanson, E.; Cai, Z.; and Fan, X. 2021. Big five personality traits and second language learning: A meta-analysis of 40 years’ research. *Educational Psychology Review*, 1–37.
- Conway, M.; Bolanis, D.; Leclerc, J.; Chiovitti, S.; Tran, S. S.; Roddick, C.; McLaughlin, K.; Woods-Fry, H.; and Giannopoulos, C. 2023. Perceiving Assertiveness and Anger from Gesturing Speed in Different Contexts. *Journal of Nonverbal behavior*, 47(1): 19–56.
- Costa, P. T.; and McCrae, R. R. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2): 179–198.
- Escalante, H. J.; Kaya, H.; Salah, A. A.; Escalera, S.; Güçlütürk, Y.; Güçlü, U.; Baró, X.; Guyon, I.; Junior, J. C. J.; Madadi, M.; et al. 2020. Modeling, recognizing, and explaining apparent personality from videos. *IEEE Transactions on Affective Computing*, 13(2): 894–911.
- Fang, H.-S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.-L.; and Lu, C. 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7157–7173.
- Gao, Z.; Li, Z.; Wang, J.; Luo, H.; Shi, X.; Chen, M.; Li, Y.; Zuo, L.; Du, Z.; Xiao, Z.; and Zhang, S. 2023. FunASR: A Fundamental End-to-End Speech Recognition Toolkit. In *INTERSPEECH*.
- Giritlioğlu, D.; Mandira, B.; Yilmaz, S. F.; Ertenli, C. U.; Akgür, B. F.; Kınıklioğlu, M.; Kurt, A. G.; Mutlu, E.; Gürel, Ş. C.; and Dibeklioğlu, H. 2021. Multimodal analysis of personality traits on videos of self-presentation and induced behavior. *Journal on Multimodal User Interfaces*, 15(4): 337–358.
- Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.
- Hickman, L.; Bosch, N.; Ng, V.; Saef, R.; Tay, L.; and Woo, S. E. 2022. Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107(8): 1323.
- Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.
- Hu, Y.; Baragchizadeh, A.; and O’Toole, A. J. 2020. Integrating faces and bodies: Psychological and neural perspectives on whole person perception. *Neuroscience & Biobehavioral Reviews*, 112: 472–486.
- Hu, Y.; and O’Toole, A. J. 2023. First impressions: Integrating faces and bodies in personality trait perception. *Cognition*, 231: 105309.
- Kelly, S. D.; and Ngo Tran, Q.-A. 2023. Exploring the Emotional Functions of Co-Speech Hand Gesture in Language and Communication. *Topics in Cognitive Science*.
- Koppensteiner, M. 2013. Motion cues that make an impression: Predicting perceived personality by minimal motion information. *Journal of experimental social psychology*, 49(6): 1137–1143.
- Leekha, M.; Khan, S. N.; Srinivas, H.; Shah, R. R.; and Shukla, J. 2024. VyaktitvaNirdharan: Multimodal Assessment of Personality and Trait Emotional Intelligence. *IEEE Transactions on Affective Computing*, 15(4): 2139–2153.
- Li, X.-L. 2017. Preconditioned stochastic gradient descent. *IEEE transactions on neural networks and learning systems*, 29(5): 1454–1466.
- Li, Y.; Wan, J.; Miao, Q.; Escalera, S.; Fang, H.; Chen, H.; Qi, X.; and Guo, G. 2020. Cr-net: A deep classification-regression network for multimodal apparent personality analysis. *International Journal of Computer Vision*, 128: 2763–2780.
- Liao, R.; Song, S.; and Gunes, H. 2024. An Open-Source Benchmark of Deep Learning Models for Audio-Visual Apparent and Self-Reported Personality Recognition. *IEEE Transactions on Affective Computing*, 15(3): 1590–1607.
- Meca-Lallana, V.; Higuera, Y.; Brañas, M.; Carrascal, P.; Rodríguez-De la Fuente, O.; Maurino, J.; and Ballesteros, J. 2019. Personality traits and perception of health-related quality of life in people with multiple sclerosis: An analysis of the NEO-FFI-3 and MSIS-29 questionnaires (P4. 2-080). *Neurology*, 92(15_supplement): P4–2.
- Mehta, Y.; Majumder, N.; Gelbukh, A.; and Cambria, E. 2020. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4): 2313–2339.
- Miranda-Correa, J. A.; Abadi, M. K.; Sebe, N.; and Patras, I. 2018. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE transactions on affective computing*, 12(2): 479–493.

- Ryumina, E.; Markitantov, M.; Ryumin, D.; and Karpov, A. 2024. OCEAN-AI framework with EmoFormer cross-hemiface attention approach for personality traits assessment. *Expert Systems with Applications*, 239: 122441.
- Song, S.; Jaiswal, S.; Sanchez, E.; Tzimiropoulos, G.; Shen, L.; and Valstar, M. 2021. Self-supervised learning of person-specific facial dynamics for automatic personality recognition. *IEEE Transactions on Affective Computing*, 14(1): 178–195.
- Song, S.; Shao, Z.; Jaiswal, S.; Shen, L.; Valstar, M.; and Gunes, H. 2022. Learning person-specific cognition from facial reactions for automatic personality recognition. *IEEE Transactions on Affective Computing*, 14(4): 3048–3065.
- Stillwell, D.; and Kosinski, M. 2015. myPersonality Project website.
- Sun, X.; Huang, J.; Zheng, S.; Rao, X.; and Wang, M. 2022. Personality assessment based on multimodal attention network learning with category-based mean square error. *IEEE Transactions on Image Processing*, 31: 2162–2174.
- Tausczik, Y. R.; and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54.
- Thomas, S.; Ferstl, Y.; McDonnell, R.; and Ennis, C. 2022. Investigating how speech and animation realism influence the perceived personality of virtual characters and agents. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 11–20. IEEE.
- Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.
- Wu, C.-H.; Wang, Y.; Parker, S. K.; and Griffin, M. A. 2020. Effects of chronic job insecurity on Big Five personality change. *Journal of Applied Psychology*, 105(11): 1308.
- Xie, X.; Cui, Y.; Jeong, C.-I.; Tan, T.; Zhang, X.; Zheng, X.; and Yu, Z. 2024. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *arXiv preprint arXiv:2404.09498*.
- Yang, Y.; Xing, Z.; and Zhu, L. 2024. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168*.
- Zafeiriou, S.; Kollias, D.; Nicolaou, M. A.; Papaioannou, A.; Zhao, G.; and Kotsia, I. 2017. Aff-wild: valence and arousal 'In-the-Wild' challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 34–41.
- Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.
- Zhang, L.; Peng, S.; and Winkler, S. 2019. PersEmoN: A deep network for joint analysis of apparent personality, emotion and their relationship. *IEEE Transactions on Affective Computing*, 13(1): 298–305.
- Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.