Multimodal Fine-Grained Apparent Personality Trait Recognition: Joint Modeling of Big Five and Questionnaire Item-level Scores

Ryo Masumura, Shota Orihashi, Mana Ihori, Tomohiro Tanaka, Naoki Makishima, Satoshi Suzuki, Saki Mizuno, Nobukatsu Hojo

NTT Corporation, Japan
ryo.masumura@ntt.com

Abstract

This paper presents a novel method for automatically recognizing people's apparent personality traits as perceived by others. In previous studies, apparent personality trait recognition from multimodal human behavior is often modeled to directly estimate personality trait scores, i.e., the "Big Five" scores. In the model training phase, ground-truth personality trait scores were usually determined from personality test results scored by many other people using fine-grained questionnaires, however, rich information in the personality test results have not been leveraged for anything other than determining the ground-truth Big Five scores. The scores assigned to each questionnaire item are thought to include more meta-level differences in personality characteristics. Therefore, we propose joint modeling methods that can estimate not only the Big Five scores but also questionnaire item-level scores. This enables us to improve awareness of multimodal human behavior. In addition, we present a newly created self-introduction video dataset with 50-item Big Five questionnaire results since previous apparent personality trait recognition datasets do not provide such personality test results. Experiments using the created dataset demonstrate that our proposed joint modeling methods with a multimodal transformer backbone can improve to estimate Big Five scores and effectively estimate questionnaire item-level scores. We also verify that the estimation performance reached human evaluation performance.

1 Introduction

Recognizing people's personality traits has gained increasing interest. One of the most widely used personality traits is the "Big Five" (Goldberg 1990; McCrae and John 1992), which is used to measure the personality in five continuous dimensions: *openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*. Two types of personality traits have been considered; self-assessed personality traits and apparent personality traits perceived by others (i.e., first impression). These two types of personality traits are often measured through questionnaire-based personality tests. While the personality test results for self-assessed personality traits can be attained by one self-trial, those for apparent personality traits need to be individually given by many people. This is considered expensive in terms of time and effort. Therefore, researchers have studied multimodal apparent personality trait recognition for automatically recognizing apparent personality traits by sensing human multimodal behavior (Mehta et al. 2020; Zhao, Tang, and Zhang 2022; Escalante et al. 2022; Ilmini and Fernando 2024).

In previous studies, several methods for multimodal personality trait recognition have been investigated (Pianesi et al. 2008; Alam, Stepanov, and Riccardi 2013; Güçlütürk et al. 2016; Gorbova et al. 2018; Kampman et al. 2018; Li et al. 2020; Principi et al. 2021; Aslan, Güdükbay, and Dibeklioğlu 2021; Suman, Saha, and Bhattacharyya 2022; Wang et al. 2023; Liao, Song, and Gunes 2022; Curto et al. 2021). Deep-learning-based methods for learning effective representations from multimodal human behavior without introducing hand-crafted features are now widely used (Güçlütürk et al. 2016; Gorbova et al. 2018; Kampman et al. 2018; Li et al. 2020; Principi et al. 2021; Aslan, Güdükbay, and Dibeklioğlu 2021; Suman, Saha, and Bhattacharyya 2022; Wang et al. 2023; Liao, Song, and Gunes 2022; Curto et al. 2021). With these methods, human-behavior information is individually extracted using modal-specific encoders, such as speech, visual, and text encoders, and then integrated to estimate personality traits. The main research topics in previous studies were feature-extraction methods from individual modals or fusion methods of multimodal information, e.g., early or late fusions (see Sec. 2). While handling multimodal input information has been the focus of much attention, a modeling method of output personality traits has not received much attention. A personality trait recognition model is usually trained to directly estimate personality trait scores. In the model training phase, ground-truth personality trait scores were often determined from personality test results scored by many other people using a fine-grained questionnaire. In other words, rich information in the personality test results has not been leveraged for anything other than determining the ground-truth Big Five scores.

However, it is wasteful not to use fine-grained questionnaire-based personality test results. Many questionnaire items (e.g., 20–360 items for Big Five) are used in such a questionnaire, and multiple test scores can be collected in the apparent personality trait recognition task (Goldberg 1990; McCrae and John 1992). The questionnaire item ex-

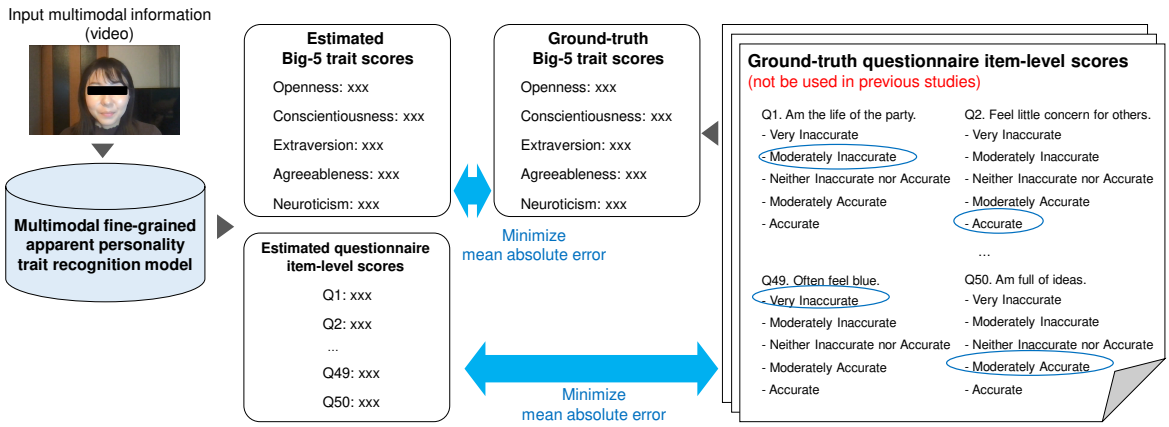

Figure 1: Overview of our proposed multimodal fine-grained apparent personality trait recognition.

amples are "he/she is the life of the party" or "he/she feels little concern for others". These scores are thought to include more meta-level differences in personality characteristics and the human behavior reflected by each answer is thought to be slightly different. Thus, it is considered effective to learn fine-grained internal information in boosting awareness of multimodal human behavior and improving Big Five score estimation. The reason fine-grained questionnaire-based personality test results are not used is that few public datasets include such results. Common datasets, such as ChaLearn first impression (Ponce-López et al. 2016; Escalante et al. 2017) and UDIVA (Palmero et al. 2021), do not provide questionnaire-based personality test results. We are curious about whether the questionnaire item-level scores can be estimated or not.

In this paper, we propose a multimodal fine-grained apparent personality trait recognition method that jointly estimates people's apparent personality trait scores, i.e., Big Five scores, and questionnaire item-level scores. In our task definition, the questionnaire-based personality test results are available in the model training phase. Figure 1 demonstrates an overview of our proposed multimodal fine-grained apparent personality trait recognition. There are two main points to this research.

- The first point is a newly created apparent personality trait recognition dataset since previous representative datasets (Ponce-López et al. 2016; Escalante et al. 2017; Palmero et al. 2021) did not provide questionnaire-based personality test results. We collected over 10,000 self-introduction videos, each of which is scored using a 50-item questionnaire of the Big Five (Goldberg 1993; Apple and Neff 2012) by five people.
- The second point is the joint modeling methods of the Big Five and questionnaire item-level scores. This paper demonstrates two types of joint modeling methods. One is a multi-task joint model that outputs Big Five scores and questionnaire item-level scores simultaneously by providing two recognition heads. The other is a cascaded joint model that first estimates questionnaire item-level scores and then estimates Big Five scores by cascading two recognition heads. We model them using a multimodal transformer architecture (Liao, Song, and Gunes 2023; Sun et al. 2019; Shi et al. 2022; Bapna et al. 2021). We expect that the joint modeling enhances awareness of multimodal human behavior and improves Big Five score estimation.

To the best of our knowledge, this is the first study that jointly estimates Big Five and questionnaire item-level scores. Experiments using the dataset demonstrate that our proposed joint modeling methods with a multimodal transformer backbone can improve to estimate Big Five scores and effectively estimate questionnaire item-level scores. We also verify that the estimation performance reached human evaluation performance.

2 Related Work

Datasets for Personality Trait Recognition There have been several datasets for investigating modeling methods of personality trait recognition. In multimodal datasets using video inputs, YouTube Vlog (Biel and Gatica-Perez 2013), Emergent LEAder (Sanchez-Cortes et al. 2013), ChaLearn first impression (Ponce-López et al. 2016; Escalante et al. 2017) and UDIVA (Palmero et al. 2021) datasets provided Big Five annotations. Among them, YouTube Vlog and ChaLearn first impression datasets handled apparent personality traits perceived by others. The former annotated Big Five trait scores using a 10-item questionnaire and the latter directly annotated them without using questionnaire items. However, both datasets did not provide questionnaire-based personality test results, so conventional modeling methods using these datasets did not leverage such rich information (Güçlütürk et al. 2016; Gorbova et al. 2018; Kampman et al. 2018; Li et al. 2020; Principi et al. 2021; Aslan, Güdükbay, and Dibeklioğlu 2021; Suman, Saha, and Bhattacharyya 2022; Wang et al. 2023; Liao, Song, and Gunes 2022; Curto et al. 2021). In addition, Emergent LEAder and UDIVA datasets did not provide questionnaire-based personality test results although they handled self-reported personality traits. Different from them, this paper demonstrates a new dataset that can access questionnaire-based personal-

Figure 2: Examples of our recorded video.

id	key	question (he or she ...)
1.	E+	is the life of the party.
2.	A-	feels little concern for others.
3.	C+	is always prepared.
4.	N-	gets stressed out easily.
5.	O+	has a rich vocabulary.
...	...	...
46.	E-	is quiet around strangers.
47.	A+	makes people feel at ease.
48.	C+	is exacting in my work.
49.	N-	often feels blue.
50.	O+	is full of ideas.

Table 1: A 50-item Big Five questionnaire.

ity test results and investigate modeling methods that utilize the rich information.

Multimodal Modeling for Personality Trait Recognition In multimodal modeling, feature-extraction methods from individual modals or fusion methods of multimodal information are important topics. The most famous fusion method in multimodal personality trait recognition is to concatenate outputs from modal-specific encoders (Gorbova et al. 2018; Principi et al. 2021; Suman, Saha, and Bhattacharyya 2022; Wang et al. 2023). On the other hand, recent successful multimodal modeling is to use multimodal transformer architecture (Liao, Song, and Gunes 2023; Sun et al. 2019; Shi et al. 2022; Bapna et al. 2021) that can effectively consider cross-modal interactions of outputs from modal-specific encoders. In multimodal personality trait recognition fields, few studies introduced a mulitmodal transformer and showed its effectiveness (Curto et al. 2021). Therefore, this paper uses the multimodal transformer for the backbone network and examines joint modeling methods that can estimate both people's apparent personality trait scores and questionnaire item-level scores. We also investigate the effectiveness of each modal information and how pre-training of modal-specific encoders impacts the improvement of model training.

3 Dataset

This section details our newly created self-introduction video dataset for evaluating fine-grained multimodal apparent personality trait recognition. We collected self-introduction videos from Japanese participants and annotated them with apparent personality traits.

Recorded Videos

We collected 10,100 self-introduction videos from 1,010 participants. We decided on ten self-introduction themes, which were "please tell us about your hobbies", "please tell us about your favorite food", "please tell us about your favorite celebrity", "please tell us about the tourist spots that you are glad you visited", "please tell us about your most impressive childhood memories", "please tell us about some interesting people you have met", "please tell us about your favorite season", "please tell us about the place you would like to visit", "please tell us about something you would like to try" and "please tell us about something you are not good at", and recorded ten videos from each participant. All participants were Japanese. The recorded videos are composed of about 12,395 min of recordings, and the average duration of each video is 73.6 seconds. The max and min duration of all videos were 102.1 s and 59.1 s, respectively. All videos were recorded using Zoom on laptop PCs. We recorded the videos at 25 fps in 1280 × 720 resolution. Camera views were frontal and we recorded the upper part of the body. The audio was recorded at 16 kHz. Figure 2 shows our recorded video images. We split the dataset into a training dataset containing 9,030 videos recorded from 903 participants, a validation dataset containing 500 videos recorded from 50 participants, and a test dataset containing 570 videos recorded from the remaining 57 participants. The number of videos in our new dataset is comparable to the ChaLearn first impression dataset which is the largest personality trait recognition dataset (Ponce-López et al. 2016; Escalante et al. 2017).

Annotations of Apparent Personality Traits

All recorded videos were annotated with apparent personality traits. We adopted the "Big Five" (Goldberg 1990; McCrae and John 1992), which is used to measure the personality in five continuous dimensions: *openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*, for the apparent personality traits. To annotate people's apparent personality traits, we recruited 200 assessors who did not know the 1,010 participants. We used a 50-item Big Five questionnaire (Goldberg 1993; Apple and Neff 2012), and videos in the training and validation datasets were scored by 5 randomly selected assessors and those in the test dataset were scored by 10 randomly selected assessors. In the test dataset, 5 annotations were used for assigning ground-truth information, and the other 5 annotations were used for performing human evaluation (the human evaluation results are shown in Section 6). Each assessor watched each recorded video two or three times and answered the questionnaire. We used a 5-point scale for scoring. Table 1 shows the 10 items in the 50-item Big Five questionnaire. Each key in Table 1 represents which personality traits it pertains to. "O", "C", "'E", "A" and "N" represent *openness*, *conscientiousness*,

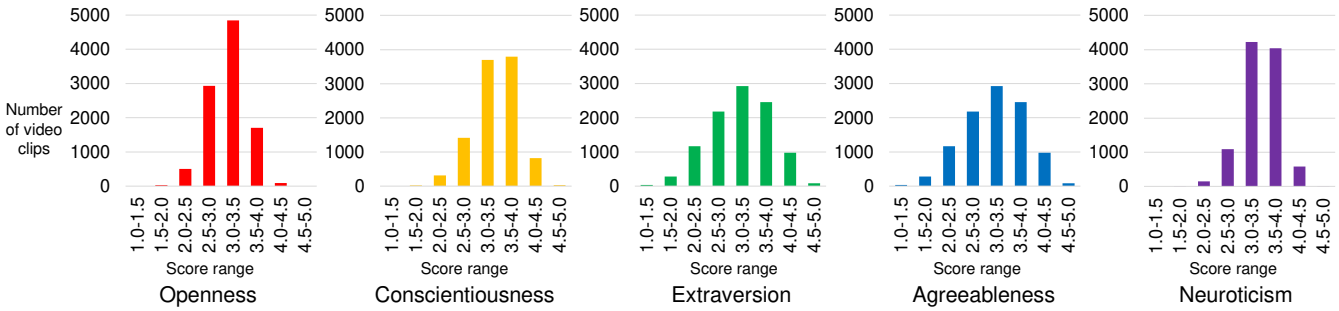

Figure 3: The histograms of the annotated Big Five personality trait scores for our recorded videos.

extraversion, *agreeableness*, and *neuroticism*, respectively. For "+" keyed items, the response "Very Inaccurate" is assigned a value of 1, "Moderately Inaccurate" a value of 2, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 4, and "Very Accurate" a value of 5. For "-" keyed items, the response "Very Inaccurate" is assigned a value of 5, "Moderately Inaccurate" a value of 4, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 2, and "Very Accurate" a value of 1. Note that the annotators were instructed to avoid assigning "Neither Inaccurate nor Accurate" as much as possible. Once numbers are assigned for all of the items in the scale, just average all the values to obtain a total scale score. Figure 3 shows the histograms of the annotated Big Five personality traits of our recorded videos. The scores of individual personality traits are in the range of [1, 5]. Note that these scores are normalized in the range of [0, 1] when using deep-learning-based modeling methods to handle them.

4 Task Definition

This section details the task definition of multimodal fine-grained apparent personality trait recognition and assumptions of a dataset. In this task, personality trait scores $\hat{\boldsymbol{y}} = [\hat{y}_1, \cdots, \hat{y}_K]^\top$ and questionnaire item-level scores $\hat{\boldsymbol{z}} = [\hat{z}_1, \cdots, \hat{z}_E]^\top$ are jointly estimated from an audio-visual video input, which is represented as audio features $\boldsymbol{S} = \{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_M\}$ and their corresponding visual features $\boldsymbol{U} = \{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_N\}$, where $\boldsymbol{s}_m$ is the m-th audio feature, $\boldsymbol{u}_n$ is the n-th visual feature, M is the number of audio features, N is the number of visual features, K is the number of personality traits, and E is the number of questionnaire items. Thus, K is 5 when handling Big Five scores (Goldberg 1990; McCrae and John 1992), and L is 50 when handling 50-item questionnaire (Goldberg 1993; Apple and Neff 2012). Audio features are generally extracted from speech information, and visual features are extracted from human RGB images. When modeling multimodal fine-grained apparent personality trait recognition, $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{z}}$ are estimated using

$$\{\hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}\} = \mathcal{F}(\boldsymbol{S}, \boldsymbol{U}; \boldsymbol{\Theta}), \tag{1}$$

where $\mathcal{F}(\cdot)$ is the model function and $\boldsymbol{\Theta}$ represents the trainable model parameter set. When an automatic speech recognition (ASR) system can be used to convert the $\boldsymbol{S}$ into text $\boldsymbol{W} = \{w_1, \cdots, w_L\}$, $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{z}}$ are estimated using

$$\{\hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}\} = \mathcal{F}(\boldsymbol{S}, \boldsymbol{W}, \boldsymbol{U}; \boldsymbol{\Theta}). \tag{2}$$

To train the model parameters $\boldsymbol{\Theta}$, we prepare a dataset of audio-visual video input with personality test results individually given by many people. The dataset is expressed as

$$\mathcal{D} = \{(\boldsymbol{S}^1, \boldsymbol{U}^1, \mathcal{Q}^1), \cdots, (\boldsymbol{S}^{|\mathcal{D}|}, \boldsymbol{U}^{|\mathcal{D}|}, \mathcal{Q}^{|\mathcal{D}|})\}, \tag{3}$$

where $\mathcal{Q}^d$ is the d-th questionnaire-based personality test results and expressed as

$$\mathcal{Q}^d = \{\boldsymbol{z}_1^d, \cdots, \boldsymbol{z}_C^d\}. \tag{4}$$

Here, $\boldsymbol{z}_c^d = [z_{c,1}^d, \cdots, z_{c,E}^d]^\top$ is the questionnaire item-level scores given by the c-th person against the d-th video input, where C is the number of people that answered the questionnaire and E is the number of items in the questionnaire. The dataset can be converted into

$$\mathcal{D} = \{(\boldsymbol{S}^1, \boldsymbol{U}^1, \boldsymbol{y}^1, \boldsymbol{z}^1), \cdots, (\boldsymbol{S}^{|\mathcal{D}|}, \boldsymbol{U}^{|\mathcal{D}|}, \boldsymbol{y}^{|\mathcal{D}|}, \boldsymbol{z}^{|\mathcal{D}|})\}, \tag{5}$$

where $\boldsymbol{y}^d = [y_1^d, \cdots, y_K^d]^\top$ is the d-th ground-truth personality trait scores and $\boldsymbol{z}^d = [z_1^d, \cdots, z_E^d]^\top$ is the d-th ground-truth questionnaire item-level scores. The k-th score in the personality trait scores and the e-th score in the questionnaire item-level scores are calculated from

$$y_k^d = \frac{1}{C} \sum_{c=1}^{C} \mathrm{PersonalityTrait}(\boldsymbol{z}_c^d; \boldsymbol{\lambda}_k), \tag{6}$$

$$z_e^d = \frac{1}{C} \sum_{c=1}^{C} z_{c,e}^d, \tag{7}$$

where $\mathrm{PersonalityTrait}(\cdot)$ is the function to convert the questionnaire item-level scores into few-dimensional personality trait scores, and $\boldsymbol{\lambda}_k$ denotes pre-defined parameters to compute the k-th personality trait. The function to compute Big Five scores from a 50-item questionnaire is detailed in Section 3.2.

5 Modeling Methods

This section details modeling methods of multimodal fine-grained apparent personality trait recognition. We present a single-task modeling method and our proposed joint-modelling methods. To effectively handle multimodal inputs, this paper adopts a mulitmodal transformer based modeling. A same backbone architecture is used for both methods.

Backbone Multimodal Transformer

In our modeling methods, a multimodal transformer is used as the backbone network architecture. The advantage of this is that different types of features can be handled with the same input method. The architecture consists of four encoder blocks: audio, text, visual, and multimodal encoders. Figure 4 shows the architecture. The audio encoder converts audio features $\boldsymbol{S}$ into audio representations $\boldsymbol{A}$, the text encoder converts text $\boldsymbol{W}$ into text representations $\boldsymbol{T}$, and the visual encoder converts visual features $\boldsymbol{U}$ into visual representations $\boldsymbol{V}$.

The multimodal encoder considers cross-modal interactions of outputs from the audio, text, and visual encoders (Liao, Song, and Gunes 2023). The inputs for the multimodal encoder are

$$\boldsymbol{H}_0 = \begin{cases} \mathrm{TemporalConcat}(\boldsymbol{A}, \boldsymbol{T}, \boldsymbol{V}) & \text{if ASR is performed,} \\ \mathrm{TemporalConcat}(\boldsymbol{A}, \boldsymbol{V}) & \text{else,} \end{cases} \tag{8}$$

$$\boldsymbol{H}'_0 = \mathrm{AddSegment}(\boldsymbol{H}_0; \boldsymbol{\theta}_{\mathrm{segment}}), \tag{9}$$

where $\mathrm{TemporalConcat}()$ is a function that concatenates inputs on the temporal axis, and $\mathrm{AddSegment}()$ is a function that adds a continuous vector in which modal-specific segment information is embedded to distinguish the concatenated vectors. We obtain hidden vectors $\boldsymbol{H}$ by

$$\boldsymbol{H} = \mathrm{TransformerEnc}(\boldsymbol{H}'_0; \boldsymbol{\theta}_{\mathrm{multi}}), \tag{10}$$

where $\mathrm{TransformerEnc}()$ is a function of the transformer encoder blocks that consist of multi-head self-attention layers and position-wise feed-forward networks (Vaswani et al. 2017) and $\boldsymbol{\theta}_{\mathrm{multi}} \in \boldsymbol{\Theta}$ are the trainable parameters of the multimodal encoder. Note that the length of $\boldsymbol{H}$ changes depending on the inputs.

The attentive pooling converts variable length hidden vectors $\boldsymbol{H}$ into a fixed size vector. The fixed vector is obtained by

$$\boldsymbol{h} = \mathrm{AttentivePooling}(\boldsymbol{H}; \boldsymbol{\theta}_{\mathrm{pool}}), \tag{11}$$

where $\boldsymbol{\theta}_{\mathrm{pool}} \in \boldsymbol{\Theta}$ is the trainable parameters of the attentive pooling, and $\mathrm{AttentivePooling}()$ is the attentive pooling function.

Single-Task Model

In single-task modeling method, apparent personality trait recognition from multimodal human behavior is modeled to directly estimate personality trait scores. Thus, the personality trait scores are calculated as

$$\hat{\boldsymbol{y}} = \mathrm{Sigmoid}(\boldsymbol{h}; \boldsymbol{\theta}_{\mathrm{head}}), \tag{12}$$

where $\mathrm{Sigmoid}()$ is a sigmoid-activation layer with a linear transformation and $\boldsymbol{\theta}_{\mathrm{head}} \in \boldsymbol{\Theta}$ are its trainable parameters.

The loss function to train the model parameters is the mean absolute error between the ground-truth personality trait scores and estimated scores and defined as

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} |\hat{\boldsymbol{y}}^d - \boldsymbol{y}^d|. \tag{13}$$

In a similar way, questionnaire item-level scores can be modeled using a single-task model. Note that a single task-model for the apparent personality trait score and that for the questionnaire item-level scores are independently optimized.

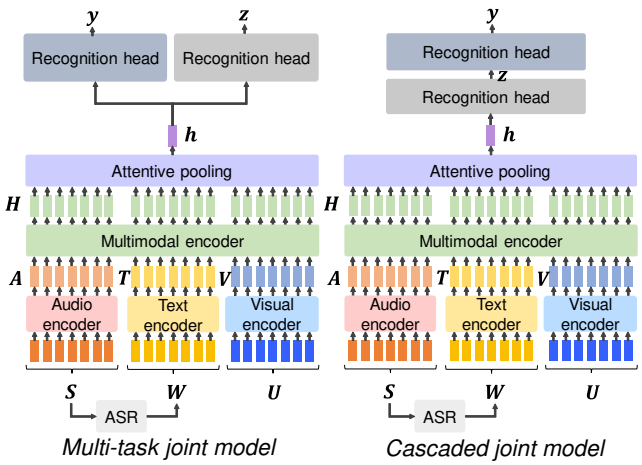

Figure 4: Multi-task joint modeling and cascaded joint modeling methods.

Joint Models

Our proposed joint modeling methods jointly estimate people's apparent personality trait scores and questionnaire item-level scores. We demonstrate multi-task joint modeling and cascaded joint modeling methods. Figure 4 shows detailed architectures.

Multi-Task Joint Model This model jointly estimates personality trait scores and questionnaire item-level scores by providing two prediction heads calculated as

$$\hat{\boldsymbol{z}} = \mathrm{Sigmoid}(\boldsymbol{h}; \boldsymbol{\theta}^{\mathrm{z}}_{\mathrm{head}}), \tag{14}$$

$$\hat{\boldsymbol{y}} = \mathrm{Sigmoid}(\boldsymbol{h}; \boldsymbol{\theta}^{\mathrm{y}}_{\mathrm{head}}), \tag{15}$$

where $\{\boldsymbol{\theta}^{\mathrm{z}}_{\mathrm{head}}, \boldsymbol{\theta}^{\mathrm{y}}_{\mathrm{head}}\} \in \boldsymbol{\Theta}$ are its trainable parameters.

Cascaded Joint Model This model first estimates questionnaire item-level scores, and then estimates personality trait scores by cascading two prediction heads calculated as

$$\hat{\boldsymbol{z}} = \mathrm{Sigmoid}(\boldsymbol{h}; \boldsymbol{\theta}^{\mathrm{z}}_{\mathrm{head}}), \tag{16}$$

$$\hat{\boldsymbol{y}} = \mathrm{Sigmoid}(\hat{\boldsymbol{z}}; \boldsymbol{\theta}^{\mathrm{y}}_{\mathrm{head}}). \tag{17}$$

The difference with the multi-task joint model is that estimated personality test scores directly affect personality trait scores.

Loss Function Both joint models are trained with the same loss function, which is calculated from not only the mean absolute error between the ground-truth and estimated personality trait scores but also the mean absolute error between the ground-truth and estimated personality test scores. It is defined as

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} |\hat{\boldsymbol{y}}^d - \boldsymbol{y}^d| + \frac{\alpha}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} |\hat{\boldsymbol{z}}^d - \boldsymbol{z}^d|, \tag{18}$$

where α is the hyper-parameter to adjust the balance between the two types of scores.

Input Modal	Pre-training of model-specific encoders	*Openness*		*Conscientiousness*		*Extraversion*		*Agreeableness*		*Neuroticism*		Average	
		Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.
A	-	0.341	93.5	0.305	92.4	0.473	89.7	0.349	92.5	0.190	93.0	0.332	92.2
A	*A*	0.493	93.9	0.619	93.4	0.647	91.2	0.572	92.3	0.473	93.5	0.561	93.0
V	-	0.141	91.9	0.261	90.1	0.098	87.1	0.192	89.9	0.034	92.5	0.145	90.3
V	*V*	0.233	93.1	0.310	90.8	0.264	86.4	0.433	92.4	0.233	93.1	0.294	90.8
A, V	-	0.246	93.4	0.253	91.8	0.370	88.6	0.263	91.8	0.124	92.9	0.251	91.8
A, V	*A*	0.504	94.1	0.571	93.0	0.696	91.1	0.576	92.5	0.505	93.3	0.570	92.8
A, V	*V*	0.323	93.5	0.298	91.8	0.492	89.7	0.435	92.0	0.329	93.1	0.375	92.0
A, V	*A, V*	0.544	94.4	0.604	93.5	0.735	91.0	0.615	92.6	0.532	94.0	0.606	93.3
A, T, V	-	0.441	94.0	0.500	92.6	0.527	89.1	0.448	92.3	0.438	93.1	0.471	92.3
A, T, V	*A*	0.553	94.0	0.573	92.3	0.707	91.2	0.599	92.4	0.563	93.6	0.599	92.8
A, T, V	*T*	0.580	94.0	0.624	93.2	0.587	89.1	0.552	92.3	0.556	93.8	0.580	92.6
A, T, V	*V*	0.504	94.0	0.502	92.8	0.583	89.0	0.507	92.2	0.501	93.2	0.519	92.3
A, T, V	*A, T, V*	**0.585**	**94.6**	**0.675**	**94.0**	**0.752**	**92.4**	**0.617**	**92.7**	**0.586**	**94.1**	**0.643**	**93.5**
Human evaluation		0.544	92.9	0.668	92.7	0.770	91.7	0.645	92.4	0.532	92.1	0.634	92.4

Table 2: Evaluation of backbone multimodal transformer and pre-training of modal-specific encoders.

6 Experiments

In our experiments, we used our created dataset detailed in Sec. 3. We verified the effectiveness of our proposed joint modeling method and compared our system with human evaluation.

Setups

Configurations We carried out pre-processing to extract audio and visual features from video input. For the acoustic features, we extracted 80 log Mel-scale filterbank coefficients. The frame shift was 10 ms. For the visual features, face regions in each input frame were detected with CenterNet (Zhou, Wang, and Krähenbühl 2019) trained on the Wider Face dataset (Yang et al. 2016). The face images were cropped and resized to 128×128, and down-sampled to 3 fps. We converted the audio features into text using a transformer-based end-to-end ASR system trained with 20K hours of Japanese speech. The backbone network architecture for the single-task model, multi-task joint model, and cascaded joint model were the same. The configuration is as follows. For the audio encoder, audio features passed two convolution and max pooling layers with a stride of 2, so we down-sampled them to $1/4$ along with the time axis. We stacked 4 transformer encoder blocks. For the visual encoder, the convolutional-neural-network function was composed of the MobileNetV3 architecture (Howard et al. 2019), and stacked two transformer encoder blocks. For the text encoder, we stacked six transformer encoder blocks. For the multimodal encoder, we stacked two transformer encoder blocks. For each encoder, the dimensions of the output continuous representations were set to 256, the dimensions of the inner outputs were set to 1024, and the number of heads in the multi-head attentions was set to 4. Swish activation was used for these encoders. For each prediction head, a fully connected layer with the sigmoid-activation function was used.

Pre-Training of Modal-Specific Encoders We pre-trained the parts of the backbone network architecture. The audio encoder was pre-trained with masked prediction of hidden units (Hsu et al. 2021) using over 20K hours of Japanese speech. The text encoder was pre-trained with a masked language-modeling task (Devlin et al. 2019) using over 100G tokens of text. The visual encoder was pre-trained in two steps: with a face-recognition task using VGGFace2 (Cao et al. 2018) in the first step and a still-image-based facial-expression-recognition task using the FER (Goodfellow et al. 2013), RAF-DB (Li, Deng, and Du 2017), and AffectNet (Mollahosseini, Hasani, and Mahoor 2019) datasets in the second step. Note that these pre-trained parameters were not freezed in the following main training.

Training After the pre-training, all parameters in each model were trained. The mini-batch size was set to 8, and the dropout rate in the transformer blocks was set to 0.1. We used RAdam (Liu et al. 2020) for optimization. The hyper-parameter α in Eq. (18) was set to 1.0. The training steps were stopped based on early stopping using the validation dataset. We trained all models with one NVIDIA A6000 GPU.

Evaluation Metrics We evaluated the multimodal apparent personality trait recognition performance in terms of Pearson's correlation coefficient and accuracy. The accuracy was computed in the same manner as with ChaLearn first impression (Ponce-López et al. 2016; Escalante et al. 2017). Note that the scores were normalized in the range of [0, 1].

Results

Evaluation of Backbone Multimodal Transformer Table 2 shows the Big Five performance when input modals were shifted. In this setting, we used the single-task modeling method. Table 2 also evaluated how pre-training of modal-specific encoders impacted. In the first column, *A*, *T*, *V* represent audio, text, and visual inputs, respectively. The

Modeling method	Joint modeling	Input Modal	*Openness* Corr.	*Openness* Acc.	*Conscientiousness* Corr.	*Conscientiousness* Acc.	*Extraversion* Corr.	*Extraversion* Acc.	*Agreeableness* Corr.	*Agreeableness* Acc.	*Neuroticism* Corr.	*Neuroticism* Acc.	Average Corr.	Average Acc.
Single-task model	-	*A, V*	0.544	94.4	0.604	93.5	0.735	91.0	0.615	92.6	0.532	94.0	0.606	93.3
Multi-task joint model	✓	*A, V*	0.540	94.5	0.641	93.5	0.742	92.0	0.627	93.4	0.560	94.0	0.622	93.4
Cascaded joint model	✓	*A, V*	0.538	94.2	0.586	93.5	0.713	91.2	0.625	93.6	0.558	94.0	0.604	93.2
Single-task model	-	*A, T, V*	0.585	94.6	0.675	94.0	0.752	**92.4**	0.617	92.7	0.586	94.1	0.643	93.5
Multi-task joint model	✓	*A, T, V*	**0.636**	**94.5**	**0.705**	**94.2**	0.743	92.1	**0.663**	**93.6**	**0.636**	**94.5**	**0.677**	**93.8**
Cascaded joint model	✓	*A, T, V*	0.562	94.4	0.659	94.8	**0.760**	91.8	0.620	93.7	0.614	93.8	0.643	93.5
Human evaluation			0.544	92.9	0.668	92.7	0.770	91.7	0.645	92.4	0.532	92.1	0.634	92.4

Table 3: Evaluation of joint modeling methods.

Modeling method	Input modal	Number of questionnaire items	Average Corr.	Average Acc.
Single-task model	*A, T, V*	0	0.643	93.5
Multi-task joint model	*A, T, V*	10	0.650	93.5
Multi-task joint model	*A, T, V*	20	0.651	93.7
Multi-task joint model	*A, T, V*	50	**0.677**	**93.8**

Table 4: Evaluation of multi-task joint models when changing the number of questionnaire items.

Modeling method	Joint modeling	Input modal	Average Corr.	Average Acc.
Single-task model	-	*A, T, V*	0.546	90.8
Multi-task joint model	✓	*A, T, V*	**0.557**	**90.8**
Cascaded joint model	✓	*A, T, V*	0.547	90.7
Human evaluation			0.438	86.6

Table 5: Evaluation of estimating questionnaire item-level scores.

second column represents which modal-specific encoders were pre-trained. The results showed that combining audio, text, and visual inputs was effective. They also showed that pre-training of modal-specific encoders was essential for each modal. The best results were comparable to human evaluation.

Evaluation of Joint Modeling Methods Table 3 shows the Big Five score estimation performance of single modeling methods and joint modeling methods. The results show that the multi-task joint model significantly outperformed the single-task model, and the cascaded joint model was comparable to the single-task model. This suggests that it is beneficial to consider auxiliary information (questionnaire item-level scores) indirectly so that it does not interfere with the end-to-end optimization of the main apparent personality trait recognition task. In addition, we can see that the multi-task joint model was effective even when increasing the number of input modal types. These results demonstrate that the multi-task joint modeling is effective in recognizing apparent personality traits from multimodal human behavior.

Evaluation of Multi-Task Joint Models When Changing the Number of Questionnaire Items Table 4 shows the results of multi-task joint models when changing the number of questionnaire items. In this evaluation, audio, text, and visual modals were used as the input, and audio, text, and visual encoders were pre-trained. The results show that the performance improved as the number of questionnaire items to be jointly modeled increased. This indicates that awareness of multimodal human behavior can be improved by increasing the number of questionnaire items.

Evaluation of Estimating Questionnaire Item-Level Scores Table 5 shows the performance of recognizing questionnaire item scores in terms of correlation and accuracy. In this evaluation, audio, text, and visual modals were used as inputs, and audio, text, and visual encoders were pre-trained. The results show that the multi-task joint model slightly outperformed the single-task model and the cascaded joint model on average. With consideration to Table 2, it is considered that the multi-task joint model is more effective for estimating Big Five than for estimating the questionnaire item-level scores. In addition, the performance of each model outperformed human evaluation. We consider that this is because our trained models could make predictions based on more stable multimodal criteria than humans.

7 Conclusion

This paper demonstrated a multimodal fine-grained apparent personality trait recognition method that jointly estimates people's apparent personality trait scores and questionnaire item-level scores. The main advantage of our proposed method is that considering an internal questionnaire-based personality test results enables us to improve awareness of multimodal human behavior. In addition, this paper demonstrated a newly created self-introduction video dataset with 50-item Big Five questionnaire results. Experiments using our created dataset showed that the multi-task joint model with audio, text, and visual features effectively estimates Big Five scores and questionnaire item-level scores. We verified that the estimation performance reached human evaluation performance.

References

Alam, F.; Stepanov, E. A.; and Riccardi, G. 2013. Personality traits recognition on social network-facebook. *In Proc. AAAI conference on web and social media*, 6–9.

Apple, M. T.; and Neff, P. 2012. Using Rasch Measurement to Validate the Big Five Factor Marker Questionnaire for a Japanese University Population. *Journal of Applied Measurement*, 13: 1–21.

Aslan, S.; Güdükbay, U.; and Dibeklioğlu, H. 2021. Multimodal assessment of apparent personality using feature attention and error consistency constraint. *Image and Vision Computing*, 110: 104163.

Bapna, A.; Chung, Y.; Wu, N.; Gulati, A.; Jia, Y.; Clark, J. H.; Johnson, M.; Riesa, J.; Conneau, A.; and Zhang, Y. 2021. SLAM : A Unified Encoder for Speech and Language Modeling via Speech-Text Joint Pre-Training. *arXiv:2110.10329*.

Biel, J.; and Gatica-Perez, D. 2013. The YouTube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs. *IEEE Transactions on Multimedia*, 15: 41–55.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. VGGFace2: A dataset for recognising face across pose and age. *In Proc. IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 67–74.

Curto, D.; Clapés, A.; Selva, J.; Smeureanu, S.; Júnior, J. C. S. J.; Gallardo-Pujol, D.; Guilera, G.; Leiva, D.; Moeslund, T. B.; Escalera, S.; and Palmero, C. 2021. Dyadformer: A Multi-modal Transformer for Long-Range Modeling of Dyadic Interactions. *In Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2177–2188.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171—4186.

Escalante, H. J.; Guyon, I.; Escalera, S.; Júnior, J. C. S. J.; Madadi, M.; Baró, X.; Ayache, S.; Viegas, E.; Güçlütürk, Y.; Güçlü, U.; van Gerven, M. A. J.; and van Lier, R. 2017. Design of an explainable machine learning challenge for video interviews. *In Proc. International Joint Conference on Neural Networks (IJCNN)*, 3688–3695.

Escalante, H. J.; Kaya, H.; Salah, A. A.; Escalera, S.; Güçlütürk, Y.; Güçlü, U.; Baró, X.; Guyon, I.; Júnior, J. C. S. J.; Madadi, M.; Ayache, S.; Viegas, E.; Gürpinar, F.; Wicaksana, A. S.; Liem, C. C. S.; van Gerven, M. A. J.; and van Lier, R. 2022. Modeling, Recognizing, and Explaining Apparent Personality From Videos. *IEEE Transactions on Affective Computing*, 13: 894–911.

Goldberg, L. R. 1990. An alternative description of personality: the big-five factor structure. *Journal of personality and social psychology*, 1216–1229.

Goldberg, L. R. 1993. The structure of phenotypic personality traits. *American Psychologist*, 48: 26–34.

Goodfellow, I. J.; Erhan, D.; Carrier, P. L.; Courville, A. C.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.; Zhou, Y.; Ramaiah, C.; Feng, F.; Li, R.; Wang, X.; Athanasakis, D.; Shawe-Taylor, J.; Milakov, M.; Park, J.; Ionescu, R. T.; Popescu, M.; Grozea, C.; Bergstra, J.; Xie, J.; Romaszko, L.; Xu, B.; Zhang, C.; and Bengio, Y. 2013. Challenges in representation learning: A report on three machine learning contests. *In Proc. International Conference on Neural Information Processing (ICONIP)*, 117–124.

Gorbova, J.; Avots, E.; Lüsi, I.; Fishel, M.; Escalera, S.; and Anbarjafari, G. 2018. Integrating Vision and Language for First-Impression Personality Analysis. *IEEE Transactions on Multimedia*, 25: 24–33.

Güçlütürk, Y.; Güçlü, U.; van Gerven, M. A. J.; and van Lier, R. 2016. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. *In Proc. European Conference on Computer Vision (ECCV) workshops*, 349–358.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; Le, Q. V.; and Adam, H. 2019. Searching for MobileNetV3. *In Proc. International Conference on Computer Vision (ICCV)*, 1314–1324.

Hsu, W.; Bolte, B.; Tsai, Y. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 29: 3451–3460.

Ilmini, W.; and Fernando, T. 2024. Detection and explanation of apparent personality using deep learning: a short review of current approaches and future directions. *Computing*, 106: 275–294.

Kampman, O.; Barezi, E. J.; Bertero, D.; and Fung, P. 2018. Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction. *In Proc. Association for Computational Linguistics (ACL)*, 606–611.

Li, S.; Deng, W.; and Du, J. 2017. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2584–2593.

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. 2022. An Open-source Benchmark of Deep Learning Models for Audio-visual Apparent and Self-reported Personality Recognition. *arXiv:2210.09138*.

Liao, R.; Song, S.; and Gunes, H. 2023. Multimodal Learning With Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 12113–12132.

Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; and Han, J. 2020. On the Variance of the Adaptive Learning Rate and Beyond. *In Proc. International Conference on Learning Representations (ICLR)*.

McCrae, R. R.; and John, O. P. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 175–115.

Mehta, Y.; Majumder, N.; Gelbukh, A.; and Cambria, E. 2020. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53: 2313–2339.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10: 18–31.

Palmero, C.; Selva, J.; Smeureanu, S.; Júnior, J. C. S. J.; Clapés, A.; Moseguí, A.; Zhang, Z.; Gallardo-Pujol, D.; Guilera, G.; Leiva, D.; and Escalera, S. 2021. Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset. *In Proc. IEEE Winter Conference on Applications of Computer Vision (WACV) Workshops*, 1–12.

Pianesi, F.; Mana, N.; Cappelletti, A.; Lepri, B.; ; and Zancanaro, M. 2008. Multimodal recognition of personality traits in social interactions. *In Proc. international conference on Multimodal interfaces*, 53–60.

Ponce-López, V.; Chen, B.; Oliu, M.; Corneanu, C. A.; Clapés, A.; Guyon, I.; Baró, X.; Escalante, H. J.; and Escalera, S. 2016. ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results. *In Proc. European Conference on Computer Vision (ECCV)*, 400–418.

Principi, R. D. P.; Palmero, C.; Júnior, J. C. S. J.; and Escalera, S. 2021. On the Effect of Observed Subject Biases in Apparent Personality Analysis From Audio-Visual Signals. *IEEE Transactions on Affective Computing*, 607–621.

Sanchez-Cortes, D.; Aran, O.; Jayagopi, D. B.; Mast, M. S.; and Gatica-Perez, D. 2013. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 7: 39–53.

Shi, B.; Hsu, W.; Lakhotia, K.; and Mohamed, A. 2022. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. *In Proc. International Conference on Learning Representations (ICLR)*.

Suman, C.; Saha, S.; and Bhattacharyya, P. 2022. Emotion-Aided Multi-modal Personality Prediction System. *In Proc. International Conference on Neural Information Processing (ICONIP)*, 289–301.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schm, C. 2019. VideoBERT: A joint model for video and language representation learning. *In Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 7463–7472.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *In Proc. Advances in Neural Information Processing Systems (NIPS)*, 5998–6008.

Wang, Y.; Li, D.; Funakoshi, K.; and Okumura, M. 2023. EMP: Emotion-guided Multi-modal Fusion and Contrastive Learning for Personality Traits Recognition. *In Proc. ACM International Conference on Multimedia Retrieval (ICMR)*, 243–252.

Yang, S.; Luo, P.; Loy, C. C.; and Tang, X. 2016. WIDER FACE: A Face Detection Benchmark. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5525–5533.

Zhao, X.; Tang, Z.; and Zhang, S. 2022. Deep Personality Trait Recognition: A Survey. *Frontiers in Psychology*.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as Points. *arXiv:1904.07850*.