Personality-aware Human-centric Multimodal Reasoning: A New Task, Dataset and Baselines

Anonymous ACL submission

Abstract

Personality traits, emotions, and beliefs shape individuals' behavioral choices and decision-making processes. However, for one thing, the affective computing community normally focused on predicting personality traits but overlooks their application in behavior prediction. For another, the multimodal reasoning task emphasized the prediction of future states and behaviors but often neglected the incorporation of individual personality traits. In this work, we introduce a new task called Personality-aware Human-centric Multimodal Reasoning (PHMR) (T^1), with the goal of forecasting the future behavior of a particular individual using multimodal information from past instances, while integrating personality factors. We accordingly construct a new dataset based on six television shows, encompassing 225 characters and 12k samples. To establish a benchmark for the task, we propose seven baseline methods: three adapted from related tasks, two pretrained model, and two multimodal large language models. The experimental results demonstrate that incorporating personality traits enhances human-centric multimodal reasoning performance. To further solve the lack of personality annotation in real-life scenes, we introduce an extension task called Personality-predicted Human-centric Multimodal Reasoning task (T^2) along with the corresponding dataset and method. We will make our dataset and code available on GitHub. [1]

1 Introduction

An individual's personality traits, emotions, and beliefs are essential components of their individual differences, and they shape their behavioral choices and decision-making processes in different situations. Personality encapsulates an individual's psychological characteristics and behavioral tendencies. The reactions of individuals, characterized by diverse personality traits, to the identical situations can exhibit significant variations. For example, when faced with an event of "Person *1* sing a song in public for Person *0* ", a *cautious* and *sensitive* Person *0* may feel embarrassed, whereas a *confident* and *lively* Person *0* may feel excited.

However, on the one hand, in the field of affective computing, there is often a focus on predicting and analyzing personality traits, without applying them to the prediction of future events related to individuals. On the other hand, recently a category of human-centric multimodal reasoning tasks, including Social-IQ (Zadeh et al., 2019), VLEP (Lei et al., 2020), and PMR (Dong et al., 2022), has emerged with the aim to infer individuals' psychological states and behaviors by leveraging the available multimodal information. Existing studies primarily concentrated on objective information related to individuals, while neglecting the incorporation of individual personality traits in reasoning.

In this work, we introduce a new task called Personality-aware Human-centric Multimodal Reasoning (PHMR) (T^1). The goal of our task is to forecast the most probable behavior of a particular individual in future scenarios within intricate social interactions that encompass multiple individuals and long-term interactions. This is achieved by integrating multimodal signals from past moments and considering their personality traits.

To support this task, we construct a dataset based on six television shows from TVQA (Lei et al., 2018), which we refer to as the Personality-aware Human-centric Multimodal Reasoning Dataset (PHMRD). We acquire the personality of the main characters through the PDB[2] website, and utilize them as the personality annotations in our dataset. Following the mainstream settings of multimodal reasoning tasks, we also define our task as a multiple choice problem. As shown in Fig. 1,

[1] We have presented 10 examples of PHMRD in anonymous GitHub.

[2] https://www.personality-database.com

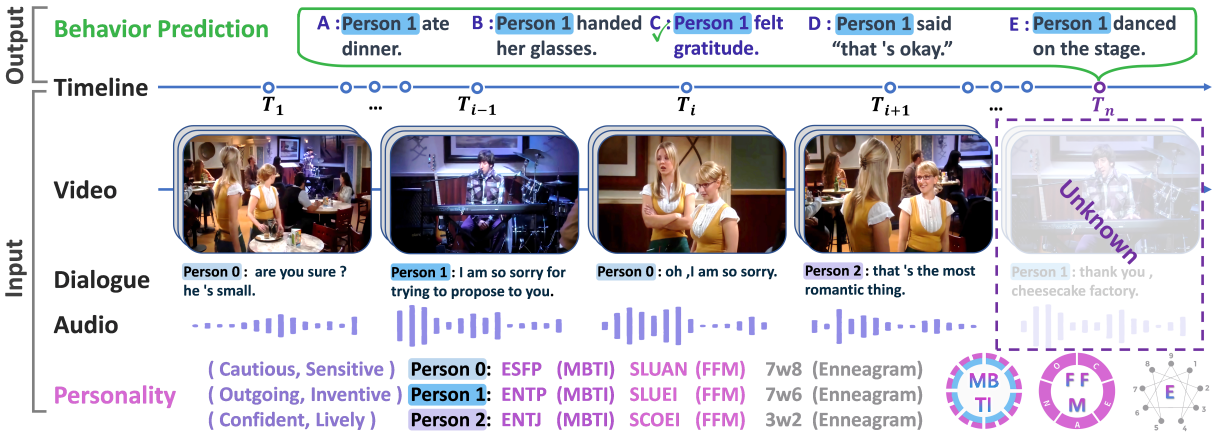

Figure 1: Illustrations of PHMR and PHMRD. Person *0*, *1*, and *2* represent the characters appearing in the video clip. Our task is to predict the most plausible behavior description at a T_n based on video, dialogue, audio and personalities information.

Person *0*, *1*, and *2* represent different characters within the video clip. Given the following information: 1) the image frame sequence of the video; 2) the utterances of dialogue; 3) the audio waveform of the video; 4) the personalities of the corresponding characters. Our task aims to select an option that best describes the most possible behavior of Person *0* in the future. For example, it should be inferred that "Person *0* felt gratitude" is the most possible option.

We benchmark the PHMR task by designing seven baseline methods: three adapted from related tasks, two pre-trained multimodal model, and two multimodal large language models. We employ a pre-trained multimodal model to extract multimodal features and incorporate personality traits, to predict the most possible behavior. The personality is represented in the form of embeddings and are effectively integrated with other multimodal features.

The experimental results reveal that the incorporation of personality traits can enhance reasoning performance in both unimodal and multimodal settings. Our proposed approach outperforms the adapted baselines, demonstrating improved performance. Our multimodal ablation experiments reveal that incorporating personality can yield performance improvements in unimodal setting for PHMR.

To further solve the lack of personality annotation in real-life scenes, we introduce an additional task called Personality-predicted Human-centric Multimodal Reasoning (T^2) along with the corresponding dataset. Taking into account the notion that an individual's personality can be inferred from their behavior and speech, we commence by predicting their personality using multimodal information. Subsequently, we leverage these predictions as surrogates for personality annotations to enhance the reasoning process (T^1). The experimental results show that our method can effectively predict these personality, and achieves satisfactory multimodal reasoning performance without relying on personality annotations.

2 Task

We introduce our primary task as Personality-aware Human-centric Multimodal Reasoning (T^1). Our focus lies on scenarios involving multiple individuals in long-term interactions, where each person's behavior is influenced by not only the actions of others, but also their own personality traits.

Specifically, four types of information are provided for a video clip $\mathcal{VC}$:

1. Video ($\mathcal{V}$): The image frame sequences of the video.
2. Audio ($\mathcal{A}$): The audio waveforms of the video.
3. Dialogue ($\mathcal{D}$): Textual utterances of dialogue, character, and corresponding time span T^D.
4. Personality ($\mathcal{P}$): Annotated personality for characters appearing in $\mathcal{VC}$.

For a fragment $\mathcal{VC}$, it can be divided into n time segments $T_1 \sim T_n$. We remove all multimodal information at the moment T_n. Given the personalities $\mathcal{P}$ and multimodal information $\mathcal{V}$, $\mathcal{A}$, $\mathcal{D}$ at the $T_1 \sim T_{n-1}$, the purpose of this task is to predict the most probable behavior $\mathcal{B}$ of a specific individual

	Train	Dev	Test	Total
# Sample	8768	1938	1910	12616
Video (s)	74.67	75.84	75.32	74.95
# Dialogue	15.57	15.46	15.07	15.48
# Personality	3.72	3.81	3.74	3.73
avg. Choice	15.24	15.40	15.21	15.26
avg. Dialogue	6.79	6.60	6.58	6.73

Table 1: Statistic of our PHMRD dataset. avg. represent average word of.

TV Show	Train	Dev	Test	Total
# Friends	2602	559	558	3719
# BBT	2548	549	553	3650
# Castle	1525	329	333	2187
# Met	843	214	190	1247
# House	745	167	165	1077
# Grey	505	120	111	736
Sum	8768	1938	1910	12616

Table 2: Statistics regarding the sample numbers of six television shows. BBT, Met, House, and Grey represent The Big Bang Theory, How I Met You Mother, House M.D. and Grey, respectively.

Personality	**Vocabulary**
FFM	[RLOAI, SCUEN, $\cdots$, RCUEI, SLOAN]
MBTI	[INTJ, ESFP, $\cdots$, ENFJ, ISTP]
Enneagram	[1w2, 2w1, $\cdots$, 9w1, 1w9]

Table 3: Statistical of the vocabularies of different personalities.

at time T_n . This can be expressed as follows:

$$\mathcal{B} = \mathrm{PHMR}(\mathcal{D}, \mathcal{V}, \mathcal{A}, \mathcal{P}), \quad (1)$$

where PHMR represents the Personality-aware Human-centric Multimodal Reasoning task.

We model PHMR as a multiple choice task, following the established settings of other multimodal reasoning tasks (Zadeh et al., 2019; Lei et al., 2020; Dong et al., 2022) , where the objective is to select the most probable answer from a set of five options, denoted as Choices ($\mathcal{C}$).

3 Dataset

3.1 Dataset Source

We choose TVQA (Lei et al., 2018) as the data source. TVQA is a large-scale video question answering dataset based on television shows, encompassing QA pairs related to object relation and human activities. Our task emphasizes human-centric question answering and primarily concerns the behavior and psychology of individuals.

3.2 Dataset Construction Process

Data Filtering To make the data more suitable for our task settings, we use ChatGPT (gpt-3.5-turbo) to exclude QA pairs that are not related to the individual's behavior and psychology. In the given instance, we consider the query "What color is the object?" to be unrelated to individuals, while the inquiry "Why does person *0* like this object?" is regarded as pertinent to human subjects. The prompt and additional examples are provided in Appendix A. The above example can be accurately evaluated by ChatGPT. To evaluate the results of ChatGPT, we manually label two hundred samples. Subsequently, a kappa test is performed on the manually labeled samples and results obtained from ChatGPT. The kappa result is 0.89, indicating a strong consistency between ChatGPT and the manually labeled samples.

Rewriting of Behavior Description We model the multimodal reasoning task as a multiple choice problem, following the mainstream setting of multimodal reasoning tasks (Social-IQ, VLEP, PMR). The correct option in PHMRD is derived from the rewriting of correct option in TVQA, and the process for obtaining the incorrect options followed the same methodology. In our preliminary experiment, it was observed that when the options are presented in the form of QA pairs, the model tends to over emphasize the importance of questions. Therefore, we manually rewrite its QA pairs by converting them into declarative sentences to describe the behaviors. For example, given the question in TVQA: "What does the person *0* felt when Person *1* sang a song in public?", and the answer: "Person *0* felt embarrassed. ", the rewritten result is "Person *0* felt embarrassed when Person *1* sang a song in public". An example of PHMRD is presented in Fig. 1. We utilize ChatGPT to rewrite the behavior description according to the aforementioned settings. We select 500 samples and manually perform the rewriting process to obtain the manually rewriting the behavior description. To measure the inter-annotator agreement, we utilize the MASI[3] (Passonneau, 2006) metric, and the resulting score is 0.72.

[3]A measure of the distance between collections used to quantify the degree of overlap between annotations of collection data.

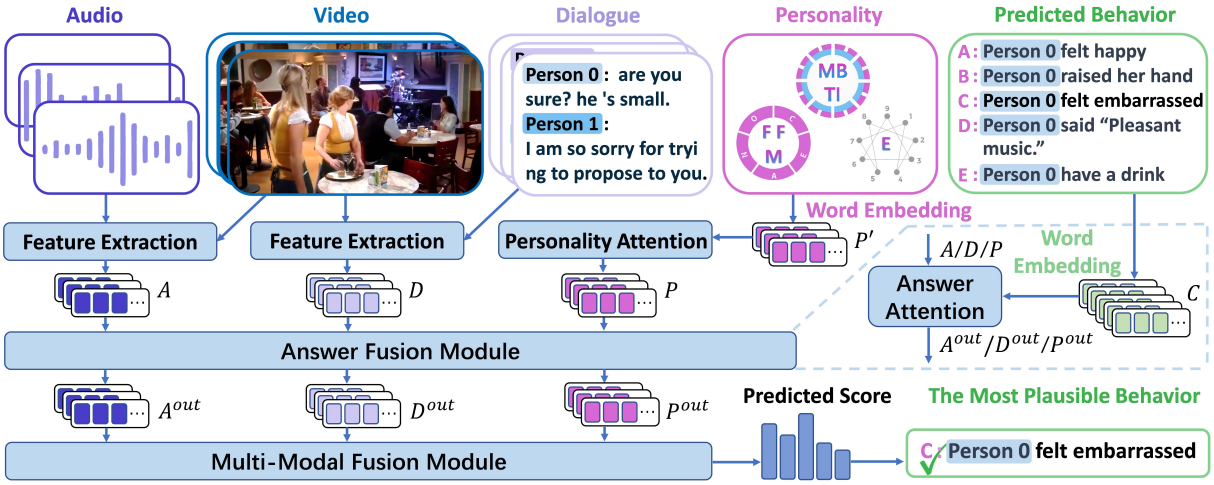

Figure 2: $\text{PRM}_{pretrain}$ model for PHMR.

Personality Annotation We obtain the personality of the characters in six television shows from the Personality Database (PDB) website. Each character's personality is determined through crowd-sourced voting on this site. We adopt the personality that constitutes the largest proportion of votes as the character's personality.

Taking into consideration that personality is assessed using multiple evaluation criteria, we have annotated the mainstream methods in psychology, namely Myers-Briggs Type Indicator (`MBTI`), Five Factor Model (`FFM`), and `Enneagram`, as shown in Table 3. For example, Penny (a character in the television show The Big Bang Theory) possesses ESFP (`MBTI`), SCUAN (`FFM`), and 7w8 (`Enneagram`) personalities, and is an *confident* and *lively* individual. More details can be found in the Appendix B.

3.3 Dataset Statistics

It consists of 225 distinct individuals, each characterized by their own unique personalities, extracted from six television series. Table 1 provides an overview of the statistics pertaining to the PHMRD dataset. On average, each segment in the PHMRD dataset has a duration of 74.95 seconds and featured approximately 3.73 characters. The specific number of samples that conform to the task settings within each television show is presented in Table 2.

4 Method

We propose a preliminary framework, Personality-aware Reasoning Model (**PRM**), to validate the effectiveness of personality in our PHMR task. The overall structure of **PMR** is shown in Fig. 2.

4.1 Multimodal Signal Representation

Since a multimodal model pre-trained on large scale data can better represent features from different modalities , we utilize a pre-trained model, Merlot Reserve (Zellers et al., 2022), which is the SOTA of TVQA, for feature extraction. We employ the notation $\mathcal{V}$, $\mathcal{A}$, $\mathcal{D}$, $\mathcal{P}$, $\mathcal{C}$ to represent the video, audio, dialogue, personality, and choices mentioned in Section 2.

For video modality $\mathcal{V}$, we employ RESERVE-L that utilizes a 24-layer ViT-L/16 (Dosovitskiy et al., 2021) to encode each frame independently. It obtains $V' \in \mathbb{R}^{n_V \times d}$, where n_V represents the length of the image frame sequence, and d=1024 denotes the feature dimension.

For audio modality $\mathcal{A}$, we utilize an Audio Spectrogram Transformer (Gong et al., 2021) to encode each subsegment independently. This is achieved by dividing the audio into intervals of 5 seconds. It yields $A' \in \mathbb{R}^{n_A \times d}$, where n_A represents the number of subsegments after the segmentation process.

For dialogue $\mathcal{D}$, the context of $\mathcal{D}$ is represented as $D' \in \mathbb{R}^{n_D \times l_D \times d_E}$. Here, n_D denotes the number of dialogue utterances, and l_D signify the number of words in a single utterance. Specifically, the representation for each word in the utterance is initialized from Merlot Reserve's embedding matrix.

To represent personality $\mathcal{P}$, we first construct vocabularies corresponding to personality traits based on their discrete distribution. The personality embedding matrix is randomly initialized and optimized in training. We have the option to utilize a single personality trait, such as [`ESFP`], or to utilize their collective set, such as [`ESFP`; `SLUAN`; `7w8`]. Ultimately, we obtain the personality feature $P' \in$

$\mathbb{R}^{n_P \times l_P \times d_E}$ of relevant the characters in the $\mathcal{VC}$. Here, n_P denotes the number of relevant characters in $\mathcal{VC}$, and l_P represents length of P.

The representation of the multiple choice $\mathcal{C}$ is expressed as $C \in \mathbb{R}^{n_C \times l_C \times d_E}$. Here, n_C represents the number of multiple choice options, and l_C signify the number of words in a single choice, respectively. Specifically, the representation for each word in the utterance is initialized from Merlot Reserve's embedding matrix.

4.2 Personality-aware Multimodal Reasoning

First, the candidate behaviors are integrated with multimodal and personality features. In particular, we concatenate P and C and feed them into Answer-Attention. We apply mean-pooling on P_C to produce the final P^{out}. Similarly, D and A are processed to obtain D^{out} and A^{out}, respectively.

Subsequently, the multimodal features are fused with the personality features. We first concatenate P^{out} with D^{out} and A^{out}, respectively. Then, multimodal information is fused to obtain the modal answer. Finally, we utilize a softmax layer to derive the final result as follows:

$$Ans' = [D^{out}; P^{out}] + [A^{out}; P^{out}], \quad (2)$$

$$Ans = \text{Softmax}(Ans'). \quad (3)$$

In addition to the above method specifically designed for the task, we also modify and adapt several methods from related tasks (TVQA, PMR) to suit our task. Details can be found in Sec. 6.2.

5 Extension: Human-centric Multimodal Reasoning with Predicted Personality

As previously mentioned, acquiring personality annotations in real life is challenging. To overcome this challenge, we propose an extension task called Personality-predicted Human-centric Multimodal Reasoning (T^2).

As shown in Fig. 3, given the availability of multimodal information, we first predict the most probable personality traits $\mathcal{P}'$ of a specific individual in the $\mathcal{VC}$. We accomplish this through the following process:

$$\mathcal{P}' = \text{Personality-Prediction}(\mathcal{D}, \mathcal{V}, \mathcal{A}). \quad (4)$$

Second, we substitute the annotated personality $\mathcal{P}$ with the predicted one $\mathcal{P}'$ in T^1 and re-evaluate its performance by

$$\mathcal{B} = \text{PHMR}(\mathcal{D}, \mathcal{V}, \mathcal{A}, \mathcal{P}'). \quad (5)$$

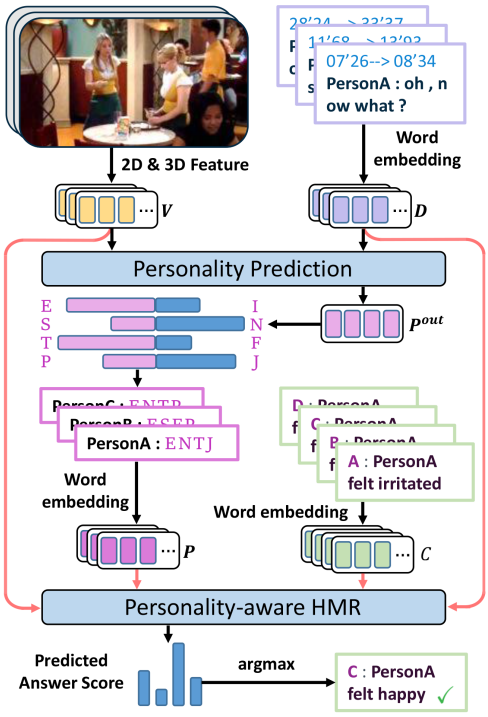

Figure 3: $\text{PRM}_{pretrain}$ for Personality prediction and Personality-aware HMR.

	Train	Dev	Test	Total
# Sample	8152	1746	1747	11645
# Personality	3.29	3.25	3.27	3.28

Table 4: Basic statistic of our MPPD dataset.

In order to achieve precise predictions of personality traits and effectively utilize these predictions for future forecasting, it is imperative to ensure the complete separation of the MPPD and PHMRD. This precautionary step is taken to prevent any data leakage, ensuring that the MPPD remains distinct and independent from the PHMRD. To accomplish this, we consider all clips within the TVQA dataset as clips within the MPPD. Subsequently, we exclude any clips that are also present in the PHMRD.

Furthermore, we eliminate characters from the dataset that possess insufficient information, such as Person *0*, who only engages in greetings or constantly eats. This process yields a list of characters associated with each clip. Once we obtain the corresponding personality traits based on these characters, we replace the names of the individuals with name tags, intentionally disrupting the association between the name and the personality traits. This precautionary measure prevents the model from memorizing the link between specific names and corresponding personality traits. The statistical de-

tails of the dataset are summarized in Table 4.

6 Experiments

6.1 Experimental Settings

The PHMRD and MPPD dataset is divided into training, validation, and testing sets at a ratio of 14:3:3. We use the Accuracy as the primary evaluation metric.

We train our **PRM** model on four RTX 3090 GPUs. For all the models, we train in 20 epochs with Early Stopping. We set the learning rate to $8e-5$ for linear layer in feature fusion, $4e-5$ for personality attention in personality feature extraction, and $1e-5$ for the rest, with linear warmup process. We report the average performance of four runs.

6.2 Compared Methods

We establish three rule-based methods, three adapted multimodal reasoning models, a modality separation variant, and two large video models, as compared methods to our **PRM** model.

Rule Based Methods. **Random**, **Longest** and **Shortest** selects a random, the longest and the shortest option, respectively.

Adapted Methods. **FVTA-adapt** (Liang et al., 2018) concentrates on aligning video and text modalities locally. **TVQA-adapt** (Lei et al., 2018) integrates personality and modality information to acquire features for each modality. **Reserve-adapt** (Zellers et al., 2022) fuses video with text and audio during pre-training, and combines them with personality features to generate the final prediction results.

Multimodal Large Language Models. **Chat-UniVi** (Jin et al., 2023) employs a visual model to convert the images into embeddings and transforms the dialogue and personality into prompts. **Gemini** (Anil et al., 2023) use 16 images and prompt with dialogue, personality, and multiple choices.

Modality Separation method. **PRM**$_{vanilla}$ is a variant of **PRM**. It generates independent representations for each of the three modalities. This approach allows us to conduct modality ablation experiments more easily and effectively. To address the challenges posed by the integration of video modality features into the audio and text modalities as observed in the Merlot Reserve experiment, we also introduce a vanilla method.

Model	w/o Per.	w/ Per.	Imp.
Shortest	16.70	-	-
Random	20.00	-	-
Longest	**34.08**	-	-
FVTA-adapt	33.64	34.35	0.71
TVQA-adapt	**43.37**	43.92	0.55
PRM$_{vanilla}$	43.03	**44.14**	**1.11**
Reserve-adapt	52.43	53.07	0.64
PRM	**53.13**	**54.06**	**0.93**

Table 5: Accuracy for rule-based, deep learning methods and video large model. Per. represents personality, while Imp. denotes the enhancement achieved by incorporating personality information.

Model	w/o Per.	w/ Per.	Imp.
Chat-UniVi	23.50	20.61	-2.89
Gemini	33.17	34.08	0.91

Table 6: Accuracy for zero-shot large video model.

6.3 Main Results

In Table 5, we report the results of different compared methods in two settings. In the supervised setting, compared to TVQA-adapt, although the **PRM**$_{vanilla}$ model's performance is suboptimal without incorporating personality, it achieves the best performance when personality information is integrated. Compared to Reserve-adapt, the **PRM** model exhibits a significant improvement. It highlights that personality information is effective for the task and our model can more effectively utilize personality information.

In the zero-shot setting, as shown in Table 6, the existing multimodal large language models face challenges in PHMR and exhibit performance inferior to fine-tuned pretraining models. In contrast to Chat-UniVi, Gemini exhibits enhanced performance by incorporating personality. This enhancement may be attributed to the inclusion of the correspondence between personality and behavior in Gemini's training corpus, such as the inclusion of task-oriented datasets for personality prediction.

6.4 Results of Different Personality Type

Table 7 displays the outcomes derived from experiments involving different personality types. The experimental results demonstrate that the combination of three personalities yields the best performance, followed by `MBTI` and `FFM`, which achieve comparable results.

Personality	Accuracy	Imp.
Baseline	53.13	-
w/ `MBTI`	54.03	0.90
w/ `FFM`	54.04	0.91
w/ `Enneagram`	53.89	0.76
w/ All	**54.06**	**0.93**

Table 7: Accuracy for different personalities for **PRM**.

Modality	w/o Per.	w/ Per.	Imp.
Dialogue	51.02	51.27	0.25
Audio	52.82	53.55	0.73
All	**53.13**	**54.06**	**0.93**

Table 8: Accuracy for different modality for **PRM**. All represents input of all modality information.

The p-value obtained from the statistical significance experiment is lower than 5%, signifying that the improvement attributed to the incorporation of personality information is statistically significant.

6.5 Results of Different Modalities

As showed in Table 8, we conduct the experiments under various modality combinations based on **PRM**. In the unimodal settings, audio outperforms dialogue both in terms of base performance and the enhancement contributed by personality. The multimodal result is higher than the unimodal one, with a more substantial improvement in personality enhancement.

As demonstrated in Table 9, the results for different modalities in **PRM**$_{vanilla}$ are presented. The audio modality exhibits low performance, thus diminishing the overall effectiveness of the vanilla method across all modalities. Nevertheless, the incorporation of personality information yields the most significant enhancement in the context of multimodality.

Modality	w/o Per.	w/ Per.	Imp.
Dialogue	43.08	43.37	0.29
Audio	36.59	36.70	0.11
Video	43.74	44.23	0.49
All	43.03	44.14	**1.11**

Table 9: Accuracy for different modality for **PRM**$_{vanilla}$.

Test Domain	w/o Per.	w/ Per.	Imp.
Friends	52.92	53.40	0.48
BBT	55.57	56.23	0.66
Castle	48.44	49.34	0.90
Met	56.14	59.12	**2.98**
House	48.68	49.49	0.81
Grey	57.65	58.85	1.20

Table 10: Accuracy for different TV shows for **PRM**.

Model	HL$_{\downarrow}$	RL$_{\downarrow}$	AP$_{\uparrow}$
PRM	25.52	25.71	77.68

Table 11: Results of personality prediction. Here, HL means Hamming Loss, and RL means Ranking Loss, AP means Average Precision.

6.6 Results on Different Television Shows

In order to account for the diverse types and quantities of episodes in the six television shows, we divide them into partitions to assess the influence of incorporating personality information. Given the limited number of available episodes for each TV show, training separate models individually proves to be a challenge. Therefore, we opt for a trained model on the complete dataset and evaluate its performance on the test sets of these six television shows. The corresponding results can be found in Table 10. The results show that the introduction of personality in each type of TV shows leads to a improvement, and that the improvement is greater for TV shows with a relatively small number of clips.

6.7 Experiments on the Extension Task T^2

We utilize the **PRM** model to train and predict personality traits on the MPPD dataset. Specifically, we predict the Myers-Briggs Type Indicator (MBTI) and apply the predicted MBTI to T^1.

Effect of Personality Prediction In Table 11, the lower is better for hamming loss and ranking loss, while the greater is better for average precision. The result indicates that the predicted personality

Model	Annotated	Predicted	Dif.
PRM	54.03	53.82	**-0.21**

Table 12: Different result between predicted personality retraining and annotated ones. Dif. indicates the difference between them.

can serve as a reasonable approximation of the ground truth.

Effect of Personality-Predicted PHMR To verify the usefulness of the predicted personality, we substitute the annotated personality with the predicted personality and train **PRM** for T^1 from scratch. Table 12 illustrates the performance, showcasing a marginal decrease of 0.21%. The findings show that multimodal information can be utilized to predict personality to alleviate the shortage of personality annotations in real life.

7 Related Work

7.1 Multimodal Reasoning

Multimodal reasoning builds upon unimodal reasoning, with examples such as VCR (Zellers et al., 2019), VCG (Park et al., 2020), and NExT-QA (Xiao et al., 2021), which include both object-centric and human-centric reasoning.

Object-centric reasoning tasks primarily focus on objects within videos, featuring less content about humans. AGQA (Grunde-McLaughlin et al., 2021) predicts object relations using video and relation graphs. Sherlock (Hessel et al., 2022) extrapolates missing information based on existing clues in images. TVQA (Lei et al., 2018) selects answers based on video, dialogue, question, and time span, with some questions related to humans.

Human-centric reasoning tasks often feature individuals appearing in a video only once, making it difficult to obtain personality information without numerous samples of the same person. Social-IQ (Zadeh et al., 2019) focused on multiple-choice questions using video, dialogue, and behavioral and psychological inquiries. WHYCAT (Ignat et al., 2021) assessed motivation based on video and descriptions. PMR (Dong et al., 2022) selected the optimal response using images and questions.

The aforementioned tasks provide the information about the question, whereas others perform reasoning without this information. These tasks usually do not involve individuals's psychological activity. VLEP (Lei et al., 2020)predicted the next likely action based on video and textual description. VAR (Liang et al., 2022) inferred images and descriptions based on videos and surrounding context, following αNLI (Bhagavatula et al., 2020), a task concentrating solely on the text modality.

7.2 Personality Computing

Personality computing (Yang et al., 2022a) is a research field that investigates personality through computational techniques utilizing various sources such as text, multimedia, and social networks. At present, three primary personality evaluation indicators are employed in research: the Five Factor Model (`FFM`) (McCrae and Costa, 1987), Myers-Briggs Type Indicator (`MBTI`) (Myers et al., 1985), and `Enneagram` (Palmer, 1988). The `FFM`'s five dimensions describe predictable surface behavior, while `MBTI`'s four dimensions primarily explain behavior and are closely related to instincts. The `Enneagram` outlines nine core motivations of individuals, each possessing its own patterns of thinking, feeling, and behaving.

The work related to personality computing has two main branches: predictive personality and applied personality. The evolution of the personality prediction task is presented as follows. Early studies (Flekova and Gurevych, 2015) adopted a simpler binary classification evaluation metric. Subsequent studies (Ponce-López et al., 2016; Kampman et al., 2018; Aslan and Güdükbay, 2019) aimed to establish predictions for the Big Five personality traits. In recent years, research efforts (Gjurkovic and Snajder, 2018; Stajner and Yenikent, 2021; Sang et al., 2022) have shifted their focus towards the `MBTI`. There were also some studies (Zhang et al., 2022; Yang et al., 2022b; Tayarani et al., 2022; Plepi et al., 2022) that focus on utilizing personality in practical applications.

8 Conclusion

In this work, we introduce a new Personality-aware Human-centric Multimodal Reasoning (PHMR) task (T^1) and construct a new dataset, PHMRD, based on six television shows. The experimental results indicate that incorporating personality information enhances the performance of human-centric multimodal reasoning. Moreover, an ablation study reveals that three distinct personality traits contribute to varying degrees of performance improvement. To further solve the lack of personality annotation in real-life scenes, we introduce an extended task called Personality-predicted Human-centric Multimodal Reasoning (T^2). The experimental results show that our method can accurately predict personality, and achieves satisfactory multimodal reasoning performance without relying on personality annotations.

Limitations

Personality-aware human-centric multimodal reasoning is a challenging task. This work is a preliminary study for this task, which was not defined as a generation task.

Due to space limitation, Our focus in this work is on the presentation of the task and the dataset. The proposed baseline is quite simple and has plenty of space for improvement.

Personality computing is a relatively broad field, our review of literature may not fully cover the research in this area.

Ethics Statement

We would like to thank Lei et al. (2018)'s valuable work on TVQA. The TVQA is licensed under a licence of MIT, which allows commercial using, modification, distribution, and private using the material for any purpose. We will also make our PHMRD publicly available later. Personality database (PDB) website is an open source, anyone can get information without registration. All of the datasets and models are in English, which benefits English speakers more. We have employed 2 postgraduates experienced in natural language processing for verify the results of ChatGPT. We pay postgraduates around $7-10 per hour, well above the local average wage, and engage in constructive discussions if they are concerned about the process.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Süleyman Aslan and Ugur Güdükbay. 2019. Multimodal video-based apparent personality recognition using long short-term memory and convolutional neural networks. *CoRR*, abs/1911.00381.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Qingxiu Dong, Ziwei Qin, Heming Xia, Tian Feng, Shoujie Tong, Haoran Meng, Lin Xu, Zhongyu Wei, Weidong Zhan, Baobao Chang, Sujian Li, Tianyu Liu, and Zhifang Sui. 2022. Premise-based multimodal reasoning: Conditional inference on joint textual and visual clues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 932–946. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1805–1816. The Association for Computational Linguistics.

Matej Gjurkovic and Jan Snajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, PEOPLES@NAACL-HTL 2018, New Orleans, Louisiana, USA, June 6, 2018*, pages 87–97. Association for Computational Linguistics.

Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: audio spectrogram transformer. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 571–575. ISCA.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. AGQA: A benchmark for compositional spatio-temporal reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11287–11297. Computer Vision Foundation / IEEE.

Jack Hessel, Jena D. Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of

Lecture Notes in Computer Science, pages 558–575. Springer.

Oana Ignat, Santiago Castro, Hanwen Miao, Weiji Li, and Rada Mihalcea. 2021. Whyact: Identifying action reasons in lifestyle vlogs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4770–4785. Association for Computational Linguistics.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2023. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *CoRR*, abs/2311.08046.

Onno Kampman, Elham J. Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 606–611. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. TVQA: localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1369–1379. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8769–8784. Association for Computational Linguistics.

Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. 2022. Visual abductive reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15544–15554. IEEE.

Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander G. Hauptmann. 2018. Focal visual-text attention for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6135–6143. Computer Vision Foundation / IEEE Computer Society.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

Isabel Briggs Myers, Mary H McCaulley, and Robert Most. 1985. *Manual, a guide to the development and use of the Myers-Briggs type indicator*. consulting psychologists press.

Helen Palmer. 1988. *The Enneagram: Understanding yourself and the others in your life*. Harper & Row.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pages 508–524. Springer.

Rebecca J. Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*, pages 831–836. European Language Resources Association (ELRA).

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7391–7402. Association for Computational Linguistics.

Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian A. Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 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*, volume 9915 of *Lecture Notes in Computer Science*, pages 400–418.

Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li, and Jeffrey Stanton. 2022. MBTI personality prediction for fictional characters using movie scripts. *CoRR*, abs/2210.10994.

Sanja Stajner and Seren Yenikent. 2021. Why is MBTI personality detection from texts a difficult task? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3580–3589. Association for Computational Linguistics.

Mohammad Tayarani, Anna Esposito, and Alessandro Vinciarelli. 2022. What an "ehm" leaks about you: Mapping fillers into personality traits with quantum evolutionary feature selection algorithms. *IEEE Trans. Affect. Comput.*, 13(1):108–121.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE.

Liang Yang, Shuqun Li, Xi Luo, Bo Xu, Yuanling Geng, Zeyuan Zeng, Fan Zhang, and Hongfei Lin. 2022a.

Computational personality: a survey. *Soft Comput.*, 26(18):9587–9605.

Qi Yang, Sergey Nikolenko, Alfred Huang, and Aleksandr Farseev. 2022b. Personality-driven social multimedia content recommendation. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 7290–7299. ACM.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8807–8817. Computer Vision Foundation / IEEE.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. MERLOT RESERVE: neural script knowledge through vision and language and sound. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16354–16366. IEEE.

Le Zhang, Songyou Peng, and Stefan Winkler. 2022. Persemon: A deep network for joint analysis of apparent personality, emotion and their relationship. *IEEE Trans. Affect. Comput.*, 13(1):298–305.

A ChatGPT Prompt and Some Samples

A.1 Data Filtering

PROMPT:" {What color is the object?} as unrelated to individuals, while {Why does person *0* like this object?} is regarded as pertinent to human subjects. Is {QUESTION} related to human behavior or psychology?"

Similarly, the question "What kind of object is on person *0*'s leg?" is considered irrelevant to individuals, whereas the query "What is person *0* talking about?" is deemed relevant to human subjects.

B Personality typology

B.1 MBTI

Myers-Briggs Type Indicator (MBTI) is an introspective self-report questionnaire used in personality typology to identify various psychological preferences in how people view the environment and make judgments.

For evaluation, the MBTI splits personality into four dimensions (E/I, S/N, T/F, and J/P). The four dimensions, in particular, have various emphases, as follows:

- Concentrate on attention, Extrovert (E), and Introvert (I) to show the people's primary energy source.
- Concentrate on cognition, Judgment (J), and Perceiving (P) to demonstrate how humans collect information.
- Concentrate on judgment, Intuition (N), and Sensing (S) to demonstrate how individuals make decisions.
- Concentrate on living, Thinking(T), and Feeling(F) to demonstrate how humans behave.

Each dimension includes two opposing labels that may be combined to produce 16 personalities, as seen in Fig. 4(c).

The 16 personalities have their own characteristics, as shown in Table 13.

B.2 FFM

Five Factor Model (FFM), also known as Big Five Model was the model to comprehend the relationship between personality and academic behaviors.this model was defined by several independent sets of researchers who used factor analysis of verbal descriptors of human behavior.these researchers began by studying relationships between a large number of verbal descriptors related to personality traits.

For evaluation, the FFM splits personality into Five dimensions (I/N, O/U, S/R, A/E, and L/C). The Five dimensions, in particular, have various emphases, as follows:

- Openness to experience (inventive/curious(I) vs. consistent/cautious(N))
- Conscientiousness (efficient/organized(O) vs. extravagant/careless(U))
- Extraversion (outgoing/energetic(S) vs. solitary/reserved(R))
- Agreeableness (friendly/compassionate(A) vs. critical/rational(E))
- Neuroticism (sensitive/nervous(L) vs. resilient/confident(C))

Detailed names of the five personality dimensions in polar opposites as shown in Fig. 4(a).

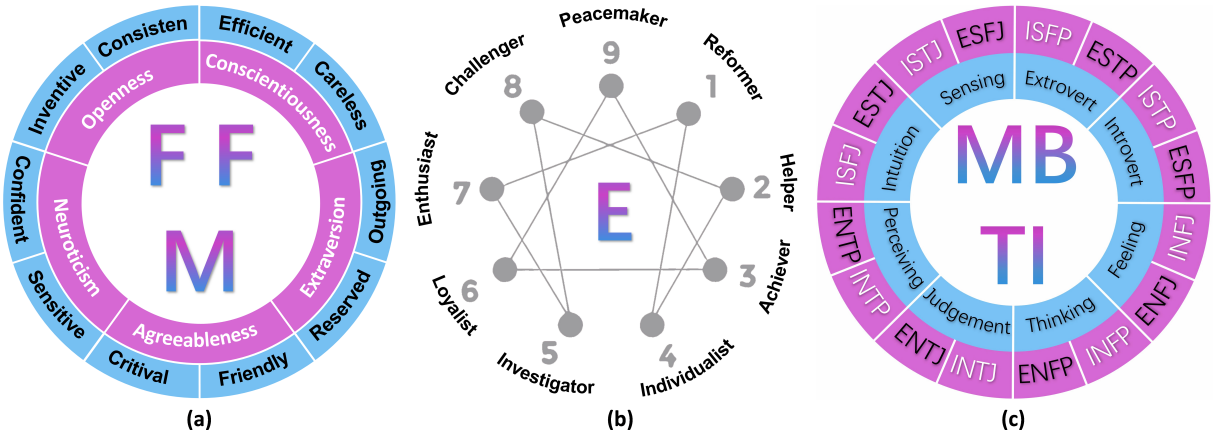

Figure 4: three personalities.

MBTI	Name	Description
ISTJ	Inspector	These people tend to be reserved yet willful, with a rational outlook on life. They compose their actions carefully and carry them out with methodical purpose.
ISFJ	Protector	These people tend to be warm and unassuming in their own steady way. They're efficient and responsible, giving careful attention to practical details in their daily lives.
INFJ	Counselor	They tend to approach life with deep thoughtfulness and imagination. Their inner vision, personal values, and a quiet, principled version of humanism guide them in all things.
INTJ	Mastermind	These thoughtful tacticians love perfecting the details of life, applying creativity and rationality to everything they do. Their inner world is often a private, complex one.
ISTP	Crafter	They tend to have an individualistic mindset, pursuing goals without needing much external connection. They engage in life with inquisitiveness and personal skill, varying their approach as needed.
ISFP	Composer	They tend to have open minds, approaching life, new experiences, and people with grounded warmth. Their ability to stay in the moment helps them uncover exciting potentials.
INFP	Healer	These rare personality types tend to be quiet, open-minded, and imaginative, and they apply a caring and creative approach to everything they do.
INTP	Architect	These flexible thinkers enjoy taking an unconventional approach to many aspects of life. They often seek out unlikely paths, mixing willingness to experiment with personal creativity.
ESTP	Promoter	They tend to be energetic and action-oriented, deftly navigating whatever is in front of them. They love uncovering life's opportunities, whether socializing with others or in more personal pursuits.
ESFP	Performer	These people love vibrant experiences, engaging in life eagerly and taking pleasure in discovering the unknown. They can be very social, often encouraging others into shared activities.
ENFP	Champion	These people tend to embrace big ideas and actions that reflect their sense of hope and goodwill toward others. Their vibrant energy can flow in many directions.
ENTP	Inventor	They tend to be bold and creative, deconstructing and rebuilding ideas with great mental agility. They pursue their goals vigorously despite any resistance they might encounter.
ESTJ	Supervisor	They possess great fortitude, emphatically following their own sensible judgment. They often serve as a stabilizing force among others, able to offer solid direction amid adversity.
ESFJ	Provider	They are attentive and people-focused, and they enjoy taking part in their social community. Their achievements are guided by decisive values, and they willingly offer guidance to others.
ENFJ	Teacher	These warm, forthright types love helping others, and they tend to have strong ideas and values. They back their perspective with the creative energy to achieve their goals.
ENTJ	Fieldmarshal	They are decisive people who love momentum and accomplishment. They gather information to construct their creative visions but rarely hesitate for long before acting on them.

Table 13: sixteen types of MBTI and their descriptions.

Enneagram	Description
Perfect	Emphasizes principles, is not easy to compromise, distinguishes between black and white, has high demands on both oneself and others, and pursues perfection.
Helping	Desire to establish a good relationship with others, people-oriented, willing to accommodate others.
Achievement	Competitive, and measure their own value by achievements, is a workaholic.
Ego	Emotional, afraid of being rejected by others, feeling that others do not understand themselves, doing their own way.
Ideal	likes to think and analyze, has a strong desire for knowledge, but lacks action, and has low requirements for material life.
Doubtful	Be cautious in doing things, not easy to trust others, have many doubts, like group life, and work hard.
Active	Optimistic, like novelty, like to follow the trend, don't like pressure.
Leader	Pursue power, emphasize strength, do not rely on others, and have a sense of justice.
Peaceful	It takes a long time to make decisions, fears disputes, and prays for harmonious coexistence.

Table 14: sixteen types of MBTI and their descriptions.

B.3 Enneagram

The `Enneagram` personality unfolds according to the nine horns of the ancient totem, revealing nine different inner dynamics that make each person inherently unique as an individual.the nine personality types described by the `Enneagram` personality theory are not good or bad; there are recognizable and fundamental differences in the way people with different personality types respond to the world. It is now common knowledge that our personalities are our own, that they filter and interpret what we see and hear, and that the basic principle of Type 9 personality theory is that each of us has one of nine possible "filters" that will keep the blueprint for our lives and the general It is used to protect a certain level within our nature and to form our communication strategy with the outside world.

Names of the nine personalities as shown in Fig. 4(b).