PersonaMovs: A Multimedia Conversational Dataset for Dynamic Personality Analysis

Anonymous ACL submission

Abstract

Automatic personality detection has evolved from simple text classification to sophisticated multimodal analysis, recognizing the multi-dimensional manifestation of personality beyond textual data. This shift highlights the need for datasets that can accurately capture the complexity of human personality through diverse modalities. We introduce the PersonaMovs (PM), a large, extensive and varied multimedia conversational dataset, built on 305 movies and 14 TV series, featuring over 46k dialogues, 552k utterances, 4016 characters, and 963 hours of video. PM not only addresses the challenges of existing datasets by offering majority-voted personality annotations and detailed relations networks but also paves the way for advanced analysis of personality dynamics across various contexts.

1 Introduction

Personality is a comprehensive yet complex trait that encapsulates individual differences in patterns of thinking, feeling, and behaving (Costa and McCrae, 2002). Detecting personality automatically is of significant importance for improvement of machine's ability to have human-like cognition and engage in more natural interactions with humans, particularly in the context of advancing Artificial General Intelligence (AGI) and various practical applications such as reflective linguistic programming (Fischer, 2023), disease diagnosis (Tseng et al., 2013) and mental health prediction (Feng et al., 2024). In recent years, there has been a burgeoning interest in automatic personality detection, marking a significant shift from traditional methods to innovative computational approaches. At the very beginning, owing to the limitations of multimedia model and computational power, researchers only treat personality prediction as a straightforward text classification task, aiming to decipher personality traits from the digital footprints individuals leave online (Kerz et al., 2022; Yang et al.). However, as shown in Figure 1, researchers have increasingly recognized that personality is manifested via multi-dimensions, with nuances that pure text-based analysis cannot fully capture (Al Maruf et al., 2022; Zhu et al., 2022; Bose et al., 2023). This revelation has propelled the move towards multimodal personality detection as the mainstream methodology.

Figure 1: The Distinctive Features in Three Modalities for Personality Prediction

Personality datasets, integrating text, audio, visual information along with the manner of speaking, face expressions, body language and so on, offer a richer, more nuanced view of human behavior and personality expressions than text-based

datasets alone. This comprehensive approach is essential for developing models that accurately reflect the complexity of human personality. Naturally, a lot of multimodal datasets were released in recent years. There have been a few attempts in multimodal personality dataset construction (Palmero et al., 2021; Junior et al., 2021; Jiang et al., 2020; Chen et al., 2022). There are also many multimodal datasets used to perform other tasks, and some personality prediction works will modify their datasets to adapt the personality context. For instance, TVQA (Lei et al., 2018) is a large dataset which is initially designed to do the visual question answering task. It is used frequently in our research field because of its large scale.

Although current datasets have evolved to include many features necessary for personality prediction, they still exhibit several limitations:

1) **Limited Data Source**: Previous datasets often select one or several famous movies or TV series as the raw data, resulting in a limited number of characters and personality types covered, which hinders the generalizability of model training.

2) **Manual Annotation Issues**: The process of manually annotating personality traits typically relies on a few numbers of volunteers with varying levels of expertise, leading to potential inconsistencies and biases in the annotations.

3) **Dynamic Nature of Personality**: From a psychological standpoint, it is essential to recognize that personality is not a static attribute but one that evolves in response to environmental contexts (Palmero et al., 2021), which current datasets do not adequately capture.

In our study, we endeavor to partially eliminate the aforementioned limitations by providing a scale-up multimodal dataset that contains reliable labels. Specifically, we find a personality database website[1] that offers a large amount of personality types for virtual characters and Zhu et al. (2023) have scraped the personality data from it to annotate TVQA dataset. Compared with previous datasets whose labelling commonly involved five to ten people, our datasets are labelled by about 160 voters on average. It shows the vote distribution rather than a single personality type which is more persuasive and operable. As for how to get the personality dynamics, incorporating relationship networks into personality prediction models offers a solution to this issue. Such networks provide a rich context for observing and understanding individual behaviors, preferences, and traits, reflecting the interconnectedness of personality with social and environmental factors.

Against these backdrops, we introduce the PersonaMovs (PM), a comprehensive dataset that starkly contrasts with existing offerings in several key aspects. PM is built on 305 movies and 14 TV series (894 episodes in total) in different genres, including more than **46k dialogues, 552k utterances, 4016 characters** and **963 hours videos**. With the rich annotation, our dataset supports 4 personality traits models (MBTI, Big Five, Enneagram and Instinctual Variant), 7 kinds of Social Relations and 8 attitudes for the Emotion Relations. Our analysis highlights substantial quantity and diversity in content, adequate experiments on different models with all modalities and personality dynamics discovery.

Our contributions are as follows:

- We introduce PersonaMovs, the most comprehensive and varied multimodal personality dataset to date, surpassing existing datasets in **scope** and **diversity**. This dataset uniquely combines movie and TV genres with personality analysis via audio, video, and text, along with crowdsourced personality, emotion, and social relations labels, unlocking new avenues in personality research[2].

- We study seven model architectures from different model families. Our results show that PersonaMovs is more **difficult** compared to other datasets, not only because it has a larger amount of multimedia data, but also due to its diversity and similarity to real life.

- For the first time, we categorize 15 types of relations to depict the dynamics of character interactions on a scene-by-scene basis, enabling a granular analysis of personality **dynamics** through relations networks. Guided by the relations networks, we identify psychological phenomenons in both short-term and long-term conversational contexts, which largely explain the personality dynamics statistically.

2 Dataset Design

This paper introduces a new multimedia personality dataset, PersonaMovs, which is the largest of exist-

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

[2] https://anonymous.4open.science/r/sample-of-MMPD-F26F/

Dataset	Dialogues	Utters. / Dial	Characters	Source
MEmoR	8.53k	64.23	7	The Big Bang Theory
FriendsPersona	0.71k	27.61	7	Friends
CPED	12k	1	392	40 TV shows
UDVIA	188	65.31	147	Dyadic Interaction
The ChaLearn FI	10k	Unknown	3000	YouTube
TVQA	29.4k	2.2	Unknown	6 TV shows
PersonaMovs	46.21k	12.42	4000+	300+ Movies and 14 TV Shows

Table 1: Comparison of different datasets and our PersonaMovs

ing multimedia datasets. In this section, we provide a specific description about our dataset in terms of design principles and the structure in details.

2.1 Design Principles

Personality refers to the combination of characteristics or qualities that form an individual's distinctive character. It encompasses a wide range of traits, behaviors, thoughts, and emotional patterns that evolve from biological and environmental factors (Lepri et al., 2012). A particular personality can determine various outward observable properties or features, including consistent behavioral patterns, communication style, emotional expression and so on. These traits manifest in how an individual consistently acts and reacts in different situations, their manner of speaking and body language, the openness or restraint of their emotional displays, their ways of relating to others, their approach to making decisions, and their preferences in activities, hobbies, and social engagements.

2.1.1 Multiple Personality Models are Needed

In constructing such a dataset for personality prediction, incorporating four distinct personality models, provides a comprehensive framework for understanding the multifaceted nature of human Personality. To this end, evolving four distinct personality models—Myers-Briggs Type Indicator (MBTI), Big Five, Enneagram, and Instinctual Variant—into our dataset construction is essential. Each of these models provides a unique lens through which to view and interpret personality traits, offering complementary insights that are critical for a holistic understanding. By integrating these four models, we aim to construct a dataset that not only captures the complexity of human personality but also facilitates nuanced predictions. This comprehensive framework acknowledges the diversity of human experience and the need for multidimensional analysis to truly understand and predict personality dynamics. The complete definitions can be found in Appendix A.

2.1.2 Personality as Fluidity Rather than Stability

Each personality model offers unique insights and covers different aspects of personality, making them collectively valuable for a multidimensional approach to personality prediction. In addition to these models, we introduce two main categories of relations among characters (more details in Appendix B):

1) **Social Relations**: These provide a comprehensive framework to observe and interpret the nuances of personality in action. We identify seven types of social relations from the perspectives of psychology and sociology. This approach acknowledges that personality is not solely a matter of internal traits and instincts, but is also fundamentally shaped and expressed through interactions with others in various domains of life.

2) **Emotion Relations**: The social relations above are relatively unchangeable, not depicting the attitudes towards someone else. So we define another 8 types for the emotion relations, as the aid for the comprehension of personality. We choose *fondness, jealousy, aversion, pity, respect, hostility, envy and gratitude* as our annotators for the emotional relations, which concludes the diverse attitudes in human's daily life. To annotate these two relations, we select a binary tuple (e.g., *Gil and Adriana: (Romantic, Fondness)*) to represent the relations combination for each pair of characters based on different scenes.

2.2 Structure of PersonaMovs

Aiming to deliver a tidy and readable structure, there is no more suitable file types than JSON format. We distribute different scenes in a single JSON file with index. For each movie or TV show, the video clips with the corresponding JSON and audio files are stored in the same directory and each data point in this dataset is centered around a dialogue scene that involves several interlocutors

based on the original scripts.

As shown in Figure 2, each video clip of our dataset is tagged with a "Scene" identifier, which likely refers to a specific segment or moment within a larger narrative or dataset. The "Dialogue" field contains an array of objects, each providing a detailed description of a scene and dialogues between characters. The dialogues are presented with time corresponding timestamps, too. The "Relationship" field within this object provides a summary or interpretation of their interaction, in this case, indicating a professional relationship with an element of fondness between Gatsby and Daisy. Finally, the "Personality" section provides personality profiles for the characters mentioned in the scene, where their personality type distribution are listed along with a "Votes Distribution" field.

3 Methodology

In this section, we outline the methodology employed to gather, process, and annotate the data for our study. We begin by detailing the sources of our multimodal video data and personality labels, focusing on how we efficiently align subtitles with original scripts to ensure accurate temporal and character associations. And we also present our annotation process, explaining how we leverage the ChatGPT API to automatically annotate social and emotional relations among characters within the text data.

3.1 Source of Data

Our data source contains mainly two parts, the multimodal video data and personality labels. For video data, we include 14 different genres of TV series and movies via an open-source website[1], and for the scripts and subtitles, we also find other open-source websites[2][3] for research offering the free scripts and subtitles of many famous movie and television programs. Considering the insufficient labeling method of existing works, we collect the personality annotations from personality database website as well as the voting distribution and align them to correctly scripts.

3.2 Data Alignment Process

As subtitle contain temporal information and original scripts associate utterances with characters, we are supposed to align them properly as efficient as possible. However, most of the existing multimodal datasets annotate the timestamps manually with taking up a great deal of time. There are also some works which utilize different automatic tools to align the utterances with their corresponding information. For instance, Lian et al. (2024) use an Automatic Sound Recognition (ASR) tool called Gentle[4] to get the timestamps for the utterances. To streamline the process of aligning dialogue utterances with their respective timestamps and speakers from subtitles, we propose an efficient method leveraging a fuzzy matching algorithm (see Appendix C). Following successful alignment, we proceed to segment the video content into distinct scenes according to the timestamps. Besides, we use FFmpeg[1] to extract the audio track from the video clips and output it as a *.mp3* file.

3.3 Annotation Process

We construct a process to automatically annotate the social and emotion relations among characters by using ChatGPT API (OpenAI, 2023). Only text data are supposed to be processed, thus we choose *gpt-3.5-turbo-1106* pre-trained model to annotate our dataset. Since we preprocess the text data and divide them into scenes, we design a prompt to ask ChatGPT for identifying both social and emotion relations for every single scene. When annotating our data, we encounter a challenge in representing unidirectional affectionate relationships, where A likes B, but B does not reciprocate the feelings. While social relations do not present this issue, emotion relations require a solution to capture this directional information. We address this by interpreting the relative position of different characters in the tuple. For example, A and B (family, fondness) indicates that A has positive feelings towards B. Conversely, if B has positive feelings towards A, the tuple would be B and A, (family, fondness). This approach allows us to clearly represent the directionality of emotional relations.

Based on the definitions of relations, we design this prompt for relations annotation (Fig 4). The prompt categorizes relations into seven social and eight emotion types, ensuring comprehensive coverage of human interactions. Note that we require ChatGPT to generate the responses following our format strictly so that we could better manipulate them flexibly.

[1] https://yts.mx/
[2] https://www.simplyscripts.com/
[3] https://subscene.com/
[4] https://github.com/lowerquality/gentle
[1] https://ffmpeg.org/

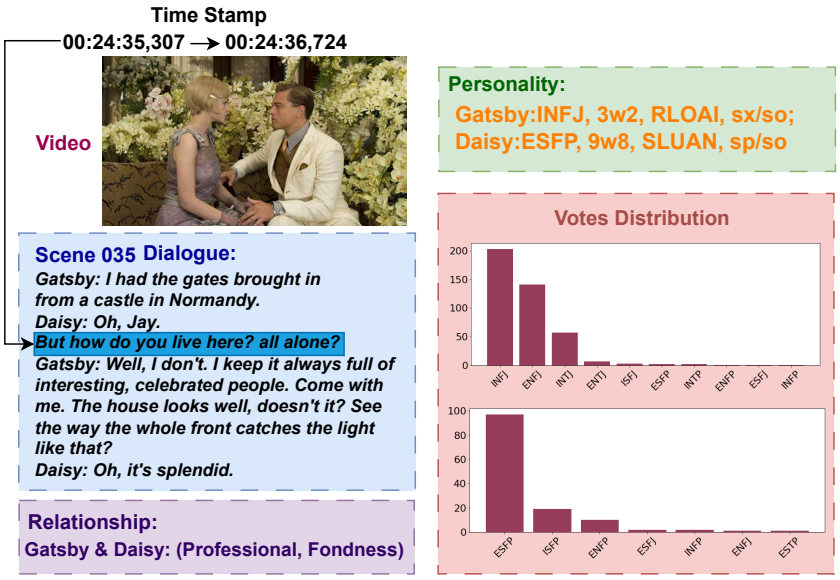


Figure 2: A sample from PersonaMovs

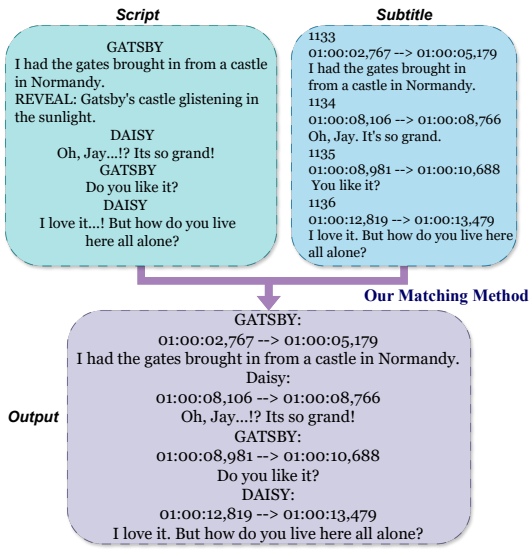


Figure 3: Process of data alignment

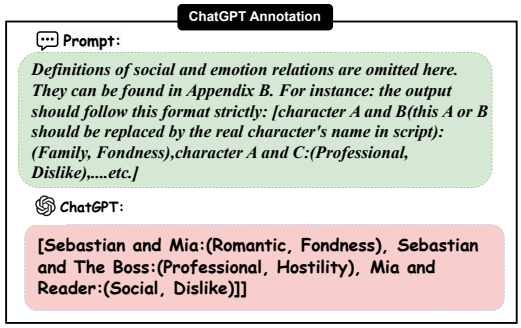


Figure 4: Prompt design for relations annotation

4 Evaluation

We present the basic statistics of our dataset in the first part, and then we evaluate the accuracy of our alignment and annotation process to ensure their reliability. Additionally, we not only test our dataset on different advanced models but also do ablation experiment on both modality and relations annotation. To discover interesting topics about personality dynamics, we focus on those changes of personality and discover several interesting psychological phenomenons.

4.1 Dataset Statistics

As we mentioned before, PersonaMovs is not only a large dataset containing a huge amount of text, audio and video corpus but also its data is highly diverse in terms of personality types, movie and television production genres, and relationship types. Fig 5 are the distribution of two types of relations, which indicates the diversity in terms of interaction scenarios.

4.1.1 Algorithm Evaluation

To evaluate the performance of our character-to-subtitle matching algorithm, we randomly sample a test case comprising over 50 dialogues and 600 utterances from a variety of genres, including 10 films and TV series. We manually check

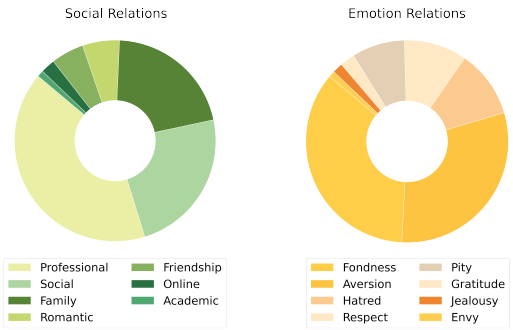


Figure 5: Distribution of social and emotion relations

the aligned characters' name based on the script. Our primary metrics for assessment is accuracy. The algorithm demonstrates an accuracy of about 88%, indicating a high level of accuracy in correctly identifying character names within subtitles across diverse content types. Compared to existing ASR matching algorithm, our approach gains an improvement by 5% in accuracy. Besides, our algorithm shows a very strong efficiency comparing the ASR method, of which accelerating almost 7 times.

Method	Movies	TV	Exec. Time (s)
Gentle (ASR)	82.71%	85.21%	26.51
Our algorithm	87.53%	88.98%	3.55

Table 2: Accuracy and running time per dialogue of subtitle matching algorithm

4.1.2 Annotation Accuracy

Using ChatGPT to annotate relations for the characters is not a completely worthwhile method. To measure the automatic annotation accuracy, we sampled 235 scenes randomly and involved 5 human labelers on relations annotation. These labelers are in their mid-twenties, undergraduate or higher education background, proficient in English with majors in psychology, filmography and sociology, who were instructed to select one of the designated social and emotion relations after aligned video. We continue to compare the automatically annotated results to the human-labeled ground truth. The outcome shows that both social and emotional relationship annotations are dependable, with the accuracy reaching 95% and 84% respectively.

Task	Movies	TV	Total
Social Relations	98.21%	93.91%	95.78%
Emotion Relations	82.04%	84.46%	84.01%

Table 3: Accuracy of relations annotation.

The dataset's foundation on crowdsourced voting allows for an in-depth analysis of subjective biases in personality perception. Researchers can investigate how different demographics (age, gender, cultural background) perceive personality traits and emotions in characters, revealing biases that may exist in personality assessment. This could also extend to studying the impact of viewer's own personality traits on their perceptions of characters, thus contributing to a deeper understanding of projection and identification processes in media consumption.

4.2 Experiment Results

4.2.1 Dataset Difficulties

We test our dataset on popular models including BERT (Devlin et al., 2019), D-DGCN (Yang et al., 2023), Roberta (Liu et al., 2019), AttRCNN (Xue et al., 2018), GPT-3.5 (OpenAI, 2023), GPT-4 (OpenAI, 2024) and MCT (Sun and Zhang, 2023). Table 4 shows the accuracy of our dataset is apparently lower than other competing datasets. One of the main challenges we observed was the complexity and diversity of our dataset compared to other multimedia datasets.

Method	Modalities	FP	TVQA	PM
BERT	T only	61.14	60.61	**52.94**
D-DGCN	T only	69.56	70.21	**68.47**
Roberta	T only	62.58	69.24	**60.37**
AttRCNN	T only	65.01	67.25	**62.44**
GPT-3.5	T only	69.21	66.89	**64.08**
GPT-4	T & V	79.14	78.33	**76.90**
MCT	T, A & V	71.67	69.93	**68.47**

Table 4: Accuracy of different methods on Friends Persona (FP), TVQA, and PersonaMovs (PM). T, A, & V stand for text, audio and video respectively. Lowest accuracy in each row is bolded.

A more challenging dataset, such as the one we have developed, offers several advantages in terms of personality detection: 1) Our dataset captures a wide range of real-life situations and intricate contexts, which better mirrors the complexity of human interactions. This realism is crucial for developing models that can perform well in practical applications. 2) Training on a more difficult dataset forces models to learn more nuanced patterns and relationships, leading to better generalization capabilities. 3) A difficult dataset sets a high standard for model evaluation, ensuring that only the most effective models are considered successful. This helps in distinguishing truly advanced models from those that perform well only on simpler

tasks. Additionally, one notable observation from the results is that the MCT model, which leverages three modalities (text, audio, and video), does not outperform the GPT-4 model, which uses only two modalities (text and video). This performance gap suggests that Large Language Model outperforms the small model on this task, even though the latter uses more modalities.

4.2.2 The Importance of Multi-Modality

We conducted a series of ablation experiments to assess the impact of different modalities and relations annotations on the performance of personality prediction models. The experiments were designed to understand how the exclusion of specific modalities or relations annotations affects the overall prediction accuracy.

Method	Modality	Accuracy
MCT	T & A	66.13
	T & V	67.91
	T only	63.43
	T, A & V	68.47
GPT-4-0125	T only	70.20
	T & V	76.90

Table 5: Ablation experiment on different modalities

Table 5 presents the results of ablation experiments where different combinations of video and audio modalities were excluded. The result underscore the critical importance of using multiple modalities to achieve higher accuracy in personality prediction tasks. Models that leverage both audio and video data, in addition to text, consistently outperform those that rely solely on textual data.

Method	With Relations	Without Relations
BERT	53.88	52.94
Roberta	59.21	58.39
GPT-4-0125	73.22	70.20

Table 6: Ablation experiment on relations annotations.

Table 6 shows the results of ablation experiments focusing on the inclusion or exclusion of relations' annotations which finds the relations annotations tend to slightly enhance the performance. This highlights the importance of including rich contextual information to improve the accuracy of personality prediction models.

The multimodal nature of the dataset (incorporating video, audio, textual, and crowd-sourced data) enables comprehensive studies that integrate different data types to understand personality. This could lead to the development of new theories or the refinement of existing ones that account for the complexity of personality as depicted through various media. It could also foster interdisciplinary research, combining insights from psychology, computer science, linguistics, and media studies.

4.2.3 Personality Dynamics

Movies and TV series and their characters often evolve over time, offering a fertile ground for studying personality dynamics. The dataset allows for longitudinal studies on how characters' personalities change in response to narrative events, relationships, and challenges. This could lead to new models that explain personality development and dynamics in complex social settings, bridging narrative theory and psychological research.

We manually select two famous characters to find their potential change of personality, and we discover there are two types of personality shifts. In the short term, people show different personalities depending on their relations with interlocutor. For example, as shown in Figure 6, Jay Gatsby from the famous romantic movie called "*The Great Gatsby*" behaves as an INFJ in front of his beloved Daisy and as an ENFJ in front of his business partners. In the long term, people may change their personality due to major turning point of life. Like Mia Dolan in "*La La Land*", she was always ENFJ, but after the breakup she became an ISFJ. The prediction results generated by GPT-4 align with the peaks in the voting distribution, indicating that this personality shift is observable within our realistic dataset.

According to our finding, we conduct statistical analysis based on our dataset to figure out if there exists certain personalities that are easily attracted to each other. To analyze the patterns of personality attraction, we focus on identifying pairs of personalities that frequently appear together bi-directionally in fondness, aversion, romantic and friendship relations. Figure 7 presents the favorite network with 16 MBTI personality types, providing a clear visual summary of statistical findings. The size of each node is proportional to the number of connections (degree) it has, which means personality types with more relationships are represented by larger nodes. The color of the edges represents the weight of the relationship between the personality types. Darker edges indicate a higher frequency or stronger relationship. Based on these vivid figures, we can discover very interesting psychological phe-

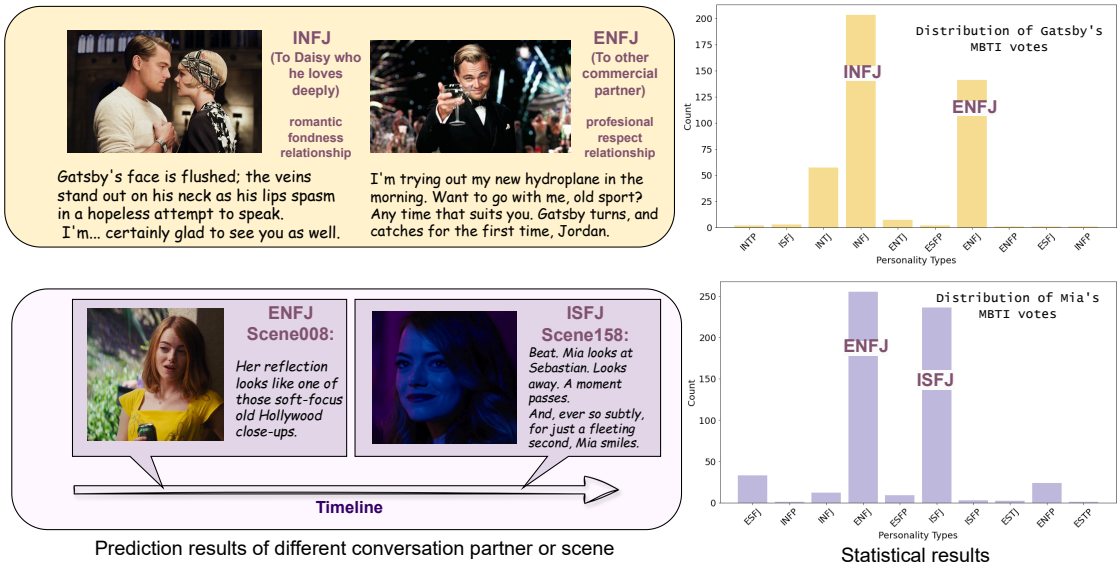
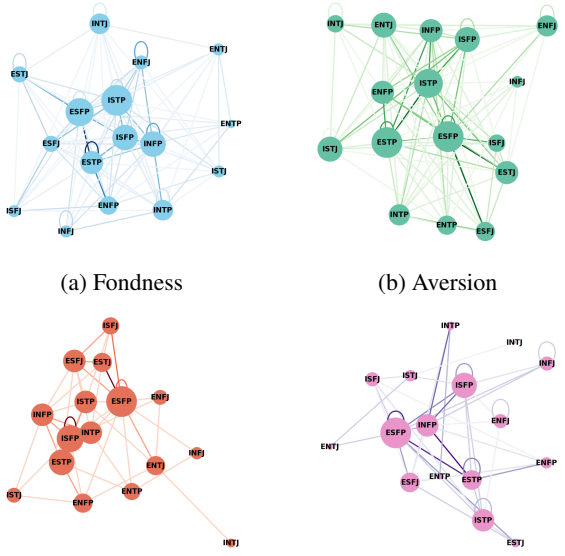


Figure 6: Case study for personality dynamics.

nomenons. For instance, ISTP is the most popular personality since almost every other personality has a fondness relation with it, and ESTP prefers to be around ESFP, ENFP and people with the same personality as themselves. ESFP may not like people with same personality because it has a dark circle on itself.



(a) Fondness (b) Aversion

(c) Romantic (d) Friendship

Figure 7: Favorite Networks with Different Personalities about Four Relations

By including data on social and emotion relations between characters, the dataset opens new pathways for exploring the dynamics of personality through interactions. This aspect can support research into how different personality types influence and are influenced by social networks, both within narrative contexts and long-term conversions. It provides a basis for computational models that simulate personality dynamics in social networks, potentially informing theories on social behavior, conflict resolution, and group dynamics.

5 Conclusion

In this study, we introduce PersonaMovs, an outstanding multimodal dataset tailored for personality prediction. Built upon a foundation of varied movies and TV shows, PM enriches with precise annotations for personality traits based on different psychological personality models and detailed relations networks, capturing the dynamic interplay of characters' interactions and emotional connections. By integrating multimodal data and emphasizing the fluid nature of personality within social contexts, PM opens new avenues for comprehensive analysis of personality dynamics, offering valuable insights into how personality traits manifest and interact in varied narratives.

Copyright Concerns

missions obtained from the copyright holders. This does not imply endorsement by or affiliation with the copyright owners. Use of these materials is limited to the scope of the permission granted and is not intended for commercial distribution.

Limitations

While our PersonaMosaic designed for personality prediction shows superiority in most aspects, it also comes with inherent limitations.

Dialogues and character behaviors extracted from movies or TV shows may not always accurately reflect real-life personality traits due to the scripted nature of these interactions. Fictional characters are often designed to serve a narrative purpose, which might exaggerate or oversimplify certain personality traits for dramatic effect, leading to potential biases in personality prediction.

The process of annotating dialogues, character relationships, and personality traits, even if partially automated, involves a degree of subjectivity. Different annotators might interpret the same dialogue or behavior differently based on their own biases and experiences, leading to inconsistencies in the dataset.

The dataset may predominantly reflect the cultural norms and values of the society in which the content was produced, potentially limiting its applicability across different cultural contexts. Our dataset is based on English movies and TV shows, so it may not interpret other non-English cultural contexts properly.

References

Abdullah Al Maruf, Md. Abdullah-Al Nayem, Md. Mahmudul Haque, Zakaria Masud Jiyad, Al Mamun Or Rashid, and Fahima Khanam. 2022. A survey on personality prediction. In *Proceedings of the 2nd International Conference on Computing Advancements*, ICCA '22, page 407–414, New York, NY, USA. Association for Computing Machinery.

Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. 2023. Movieclip: Visual scene recognition in movies. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2082–2091.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022. CPED: A large-scale chinese personalized and emotional dialogue dataset for conversational ai.

W. Andrew Collins and L. Alan Sroufe. 1999. *Capacity for Intimate Relationships: A Developmental Construction*, Cambridge Studies in Social and Emotional Development, page 125–147. Cambridge University Press.

Paul Costa and Robert McCrae. 2002. Personality in adulthood: A five-factor theory perspective. *Management Information Systems Quarterly - MISQ*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Robert A. Emmons and Michael E. McCullough. 2004. *The Psychology of Gratitude*. Oxford University Press.

Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. 2024. How far are we from agi.

Kevin A. Fischer. 2023. Reflective linguistic programming (rlp): A stepping stone in socially-aware agi (socialagi).

Hang Jiang, Xianzhe Zhang, and Jinho D. Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13821–13822.

Julio C. S. Jacques Junior, Agata Lapedriza, Cristina Palmero, Xavier Baro, and Sergio Escalera. 2021. Person perception biases exposed: Revisiting the first impressions dataset. In *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE.

Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. 2022. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Bruno Lepri, Jacopo Staiano, Giulio Rigato, Kyriaki Kalimeri, Ailbhe Finnerty, Fabio Pianesi, Nicu Sebe, and Alex Pentland. 2012. The sociometric badges corpus: A multilevel behavioral dataset for social behavior in complex organizations. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 623–628.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. 2024. Merbench: A unified evaluation benchmark for multimodal emotion recognition.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

OpenAI. 2023. Gpt-3.5-turbo. https://www.openai.com/research/gpt-3-5. Accessed: 2024-06-03.

OpenAI. 2024. Gpt-4-0125. GPT-4-0125 model.

Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, David Leiva, and Sergio Escalera. 2021. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1–12.

Mingwei Sun and Kunpeng Zhang. 2023. Multimodal co-attention transformer for video-based personality understanding. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1450–1459.

Chiu-yu Tseng, Chao-yu Su, and Tanya Visceglia. 2013. Levels of lexical stress contrast in english and their realization by l1 and l2 speakers. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–5.

Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. 2018. Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, 48.

Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. Multi-document transformer for personality detection. 35(16):14221–14229.

Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 2023. Orders are unwanted: Dynamic deep graph convolutional network for personality detection.

Yangfu Zhu, Linmei Hu, Xinkai Ge, Wanrong Peng, and Bin Wu. 2022. Contrastive graph transformer network for personality detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4559–4565. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yaochen Zhu, Xiangqing Shen, and Rui Xia. 2023. Personality-aware human-centric multimodal reasoning: A new task.

A Definitions of Personality Models

- **Myers–Briggs Type Indicator (MBTI)**: The MBTI categorizes personality into four dimensions. Extraversion (E) vs. Introversion (I): Extraverts are outgoing and energized by social interactions, while Introverts are reserved and energized by solitude. Sensing (S) vs. Intuition (N): Sensors focus on present, concrete information, valuing practicality, whereas Intuitives are imaginative and future-oriented, valuing abstract ideas. Thinking (T) vs. Feeling (F): Thinkers base decisions on logic and fairness, prioritizing objectivity, while Feelers base decisions on personal values and the impact on others, prioritizing harmony. Judging (J) vs. Perceiving (P): Judgers prefer structured and organized lives, liking plans and decisiveness, while Perceivers prefer flexibility and spontaneity, liking to keep their options open. Each MBTI type is defined by a combination of four cognitive functions, which can be either introverted (i) or extraverted (e). Extraverted Sensing (Se): Focuses on the present moment and physical reality, highly attuned to sensory experiences. Introverted Sensing (Si): Relies on past experiences and memories, valuing tradition and consistency. Extraverted Intuition (Ne): Sees patterns and connections, focusing on future possibilities and abstract ideas. Introverted Intuition (Ni): Focuses on internal insights and foresight, seeing underlying meanings and future potentials. Extraverted Thinking (Te): Organizes and structures the external world, prioritizing logic and efficiency. Introverted Thinking (Ti): Analyzes and categorizes information internally, valuing logical consistency and understanding. Extraverted Feeling (Fe): Prioritizes harmony and social values, focusing on the needs and feelings of others. Introverted Feeling (Fi): Values personal beliefs and feelings, making decisions based on inner values and ethics.

- **Big Five Personality Traits**: The Big Five model describes personality using five broad traits. Openness to Experience: High openness involves imagination and insight, while low openness involves practicality and routine. Conscientiousness: High conscientiousness is characterized by organization and dependability, while low conscientiousness is characterized by spontaneity and flexibility. Extraversion: High extraversion includes sociability and assertiveness, while low extraversion (introversion) includes reserve and solitude. Agreeableness: High agreeableness in-

Relations type	Description
Family	Parents (grandparents) and children, siblings, etc.
Friendship	Based on common interest, mutual respect and affection, but not related to the blood.
Romantic	Based on emotional attraction and include dating, marriage, etc.
Professional	Formed in a work environment, such as colleagues, superiors and subordinates, etc.
Social	Formed in a broader social context, such as neighbors, club members.
Academic	Formed in an educational setting, such as between teachers and students, classmates.
Online	Established in online spaces or through social media platforms.

Table 7: Descriptions of Social Relations

Relations type	Description
Fondness	A positive emotion characterized by a person's fondness for another.
Jealousy	Unhappy and angry because someone has something that you want.
Aversion	A negative emotion, referring to a feeling of disfavor towards someone.
Pity	A feeling of sadness for someone else's difficult situation.
Respect	Admiration felt or shown for someone that you believe has good ideas or qualities.
Hostility	An unfriendly or unkindness towards someone or something.
Envy	A discontented feeling when a person desires what someone else has.
Gratitude	An emotion of being thankful for someone else's help or kind actions.

Table 8: Description of Emotion Relations

volves trust and altruism, while low agreeableness involves skepticism and competition. Neuroticism: High neuroticism involves emotional instability and anxiety, while low neuroticism involves emotional stability and calmness.

- **Enneagram**: The Enneagram classifies personality into nine types, each representing different motivations and fears. Type 1: The Reformer, driven by a need for perfection. Type 2: The Helper, driven by a need to be loved. Type 3: The Achiever, driven by a need for success. Type 4: The Individualist, driven by a need for uniqueness. Type 5: The Investigator, driven by a need for knowledge. Type 6: The Loyalist, driven by a need for security. Type 7: The Enthusiast, driven by a need for variety and fun. Type 8: The Challenger, driven by a need for control. Type 9: The Peacemaker, driven by a need for harmony. A 2w3 individual is likely to be more ambitious, charming, and goal-oriented than a typical Type 2. They still seek to help others but are also motivated by a desire for success and recognition.

- **Instinctual Variants**: The Instinctual Variants theory describes three primary instinctual drives influencing behavior. Self-Preservation (SP): Focuses on safety, health, and comfort. Social (SO): Focuses on relationships, status, and community. Sexual (SX): Focuses on intimacy, attraction, and one-on-one connections. For instance, an 8w7 with a Sexual variant, is highly charismatic and seeks intense and passionate connections with others. He or she is bold and assertive, often focusing his or her energy on building strong, impactful relationships.

B Definitions of Relations

Human social networks are complex and multifaceted. By categorizing relations, we can better understand the dynamics and nuances of how people interact with each other. Different types of relations provide context for interactions, which is crucial for analyzing social behaviors and patterns, improving social network analysis, and applying this knowledge across various fields and applications. Table 7 and Table 8 provide a structured approach to understanding the complex web of relations that individuals navigate. By categorizing these relations into social and emotional types, we can better analyze and predict personality dynamics in various contexts (Collins and Sroufe, 1999; Emmons and McCullough, 2004).

C Data Alignment Algorithm

The details of data alignment algorithm are as follows:

1. *Preprocess the raw data* Firstly, we divide the scripts into several scenes according to the coherence in language of camera, instead of randomly clipping in a certain time period. This segmentation is guided by explicit scene transition cues found in movie scripts, such as

Algorithm 1 Scripts and Subtitles Matching

Input: $Script, Subtitles$
Output: Updated subtitles with speaker names

$dial\&speakers \leftarrow empty$
$threshold \leftarrow 0.8$
for $scene$ in $Script$ **do**
  for $Dials$ in $scene$ **do**
    Extract $speaker$ and $dial$ from $Dials$
    $dial\&speakers \leftarrow speaker, dial$
  end for
end for
for $subtitle$ in $Subtitles$ **do**
  $match\_score \leftarrow 0$
  $match\_speaker \leftarrow Null$
  for $line$ in $subtitle$ **do**
    for $speaker, dial$ in $dial\&speakers$ **do**
      $score \leftarrow Similar(subtitle, dial)$
      if $score > match\_score$ **then**
        Update $match\_score$ and $match\_speaker$
      end if
    end for
    if $match\_score \geq threshold$ **then**
      Update $line$ with $match\_speaker$
    end if
  end for
  Update $subtitle$
end for
return Updated $Subtitles$

"*CUT TO:*" or scene location indicators. For TV show scripts, which might lack uniform scene transition markers, we identify scene changes by detecting pauses exceeding 3 seconds between utterances.

2. *Match the utterance* This algorithm is rooted in the comparison of utterances from original scripts and subtitles based on a similarity threshold. If the similarity between a pair of utterances meets or exceeds this threshold, the character's name is accurately associated with the utterance.

3. *Rematch with the slide window* Basically, the content in scripts is slightly different with the subtitles, because the director may have improvised on the set. Thus, we introduce a slide window algorithm to evaluate the utterance-level similarity. As shown in Algorithm 2, we set a window to slide over the script and, for each utterance, compare the content inside the window with each subtitle entry to get the similarity of the paragraph in the window.

Algorithm 2 Slide Window Matching

Input: $Script, Subtitles$
Output: Updated subtitles

$window\_size \leftarrow 10$
$threshold \leftarrow 0.8$
$matches \leftarrow empty\_list$
for $i \leftarrow 0$ to $Len(Script) - window\_size$ **do**
  $window \leftarrow slice(scriptTokens, i, i + window\_size)$
  $match\_score \leftarrow 0$
  for $j \leftarrow 0$ to $Len(Subtitles) - 1$ **do**
    $score \leftarrow Similar(window, Subtitles[j])$
    if $score > match\_score$ **then**
      Update $match\_score$
    end if
  end for
  if $match\_score \geq threshold$ **then**
    $matches \leftarrow Subtitles[j]$
  end if
end for
return Updated $Subtitles$ with $matches$