PersoMoni: A Comprehensive Video-Based Benchmark Dataset for Fine-grained Personality Assessment with 15 Trait Dimensions

Feng-Qi Cui, Jinyang Huang*, Sirui Zhao, Kun Li, Zhi Liu, Meng Li,
Ziyu Jia, Dan Guo*, Meng Wang, *Fellow, IEEE*

***Abstract*—Understanding personality from visual behavior remains challenging due to the limitations of existing personality trait identification (PTI) benchmarks, which are typically based on short, crowd-annotated clips and thus capture only coarse, impressionistic judgments. These datasets lack temporal continuity, clinical validity, and, crucially, the fine-grained sub-trait structure emphasized in psychological assessment. To address these gaps, we introduce *PersoMoni*, a clinically grounded benchmark designed for fine-grained, temporally rich personality computing. The dataset contains 168 full-length psychological interviews conducted by licensed counselors, producing over 23000 aligned facial video segments. Each participant is annotated using the full BFI-2 taxonomy, providing continuous labels for five major traits and, for the first time, fifteen validated sub-traits, which marks the first benchmark to extend personality granularity from 5 to 15 dimensions. This design enables long-horizon modeling, regression-based evaluation, and detailed analysis of subtle behavioral markers in ecologically valid settings. Leveraging this benchmark, we reveal two limitations of current PTI systems, *i.e.*, limited sensitivity to sparse and weak but psychologically meaningful behavioral cues, and modeling difficulty to inter-individual variability across long interactions. To tackle these issues, we propose *PSM-Net*, a personality-aware sparse modeling framework that couples 3D CNN–based encoding with sparsity-driven temporal refinement to enhance trait-discriminative signals. Extensive experiments across compensatively diverse architectures, *i.e.*, 2D CNNs, 3D CNNs, and Transformers, show that the proposed method consistently outperforms existing approaches on both the Big Five traits and all fifteen BFI-2 sub-traits. The dataset and source code are released at https://github.com/QIcita/PersoMoni.**

***Index Terms*—Personality Trait Identification, Fine-grained Behavioral Modeling, Video-based Personality Computing, Sparse Temporal Learning.**

I. Introduction

AUTOMATIC Personality trait identification (PTI) from visual behavior has become a key emerging direction in affective computing [1], [2], reflecting a broader shift from traditional expert-driven assessments toward scalable and behavior-informed intelligent diagnosis. Although the Big Five Personality Traits (BFPT) model provides a robust theoretical foundation [3], conventional methods using these guidelines, *e.g.*, self-report inventories and structured clinical interviews, still remain labor-intensive, subjective, and thus difficult to deploy at scale. These limitations correspondingly catalyze growing interest in computational approaches that infer personality directly from naturalistic cues, including facial dynamics [4]–[6], prosody [7], and other nonverbal behaviors [2], [8]. Personality judgment and social perception research suggests that these cues are not deterministic indicators of personality, but observable behavioral signals, from which relatively stable trait tendencies can be partially inferred across social interaction contexts [9], [10]. This transition from manual evaluation to data-driven auto behavioral modeling enables more efficient psychological screening, further supports remote and real-time assessment, and opens new opportunities in mental health [11], [12], human–computer interaction [13], [14], and personalized intelligent systems [15]–[17].

Despite progress, current research in personality computing remains constrained by several important limitations. Most widely used benchmarks, *e.g.*, the ChaLearn First Impressions dataset [18], are built from short, socially constrained video clips with crowd-annotated labels. As a result, they often reflect superficial first-impression judgments rather than deeper personality structure. Moreover, from a psychological perspective, personality is more likely to be expressed through subtle nonverbal cues unfolding over time, including gaze regulation, micro-expressions, hesitation behaviors, speech rhythm, and postural shifts [19]. However, pioneer state-of-the-art (SOTA) short-video benchmarks often lack the temporal depth needed to capture stable and psychologically meaningful behavioral patterns. Although thin-slice and zero-acquaintance studies indicate that brief observations can reflect certain personality cues [20], such clips are usually insufficient for fine-grained personality inference, which relies on richer and more sustained behavioral evidence [21].

Beyond the limited temporal depth discussed above, existing SOTA datasets also suffer from insufficient label granularity. Most of them provide only coarse trait-level annotations, neglecting the fine-grained sub-trait dimensions specified in

Feng-Qi Cui, Jinyang Huang, Meng Li, Dan Guo, and Meng Wang, are with the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, and the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China.

Feng-Qi Cui and Sirui Zhao are with the MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, and the School of Information Science and Technology, University of Science and Technology of China, Hefei, China.

Kun Li is with the College of Information Technology (CIT), United Arab Emirates University (UAEU), Al Ain, Abu Dhabi, United Arab Emirates.

Zhi Liu is with the Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo, Japan.

Ziyu Jia is with the Beijing Key Laboratory of Brainnetome and Brain-Computer Interface, and the Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Corresponding author*: Jinyang Huang (Email: hjy@hfut.edu.cn), Dan Guo (guodan@hfut.edu.cn).

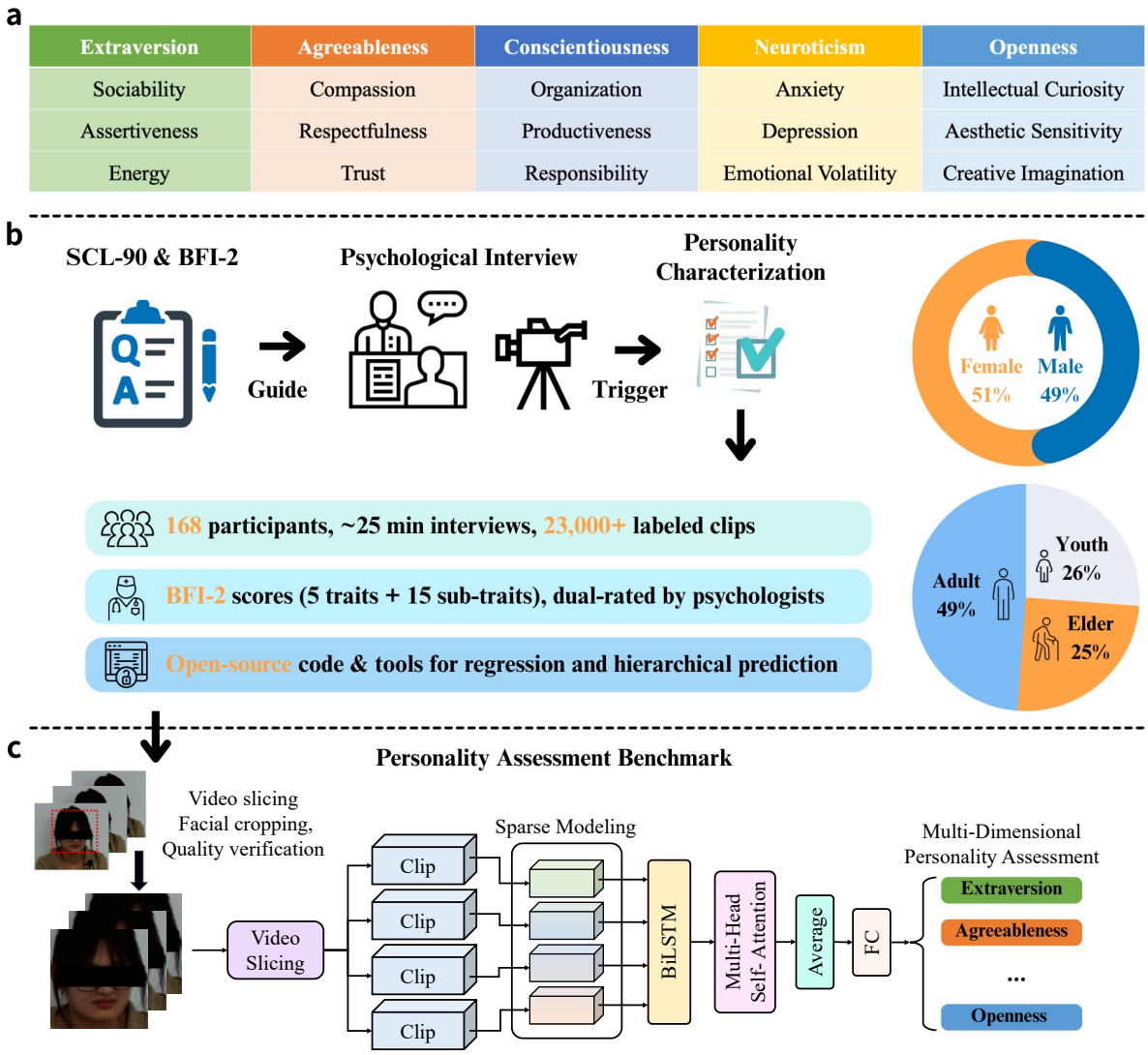


Fig. 1: Overview of the proposed personality assessment benchmark. (a) The hierarchical structure of BFI-2 personality representation, including 5 major traits and 15 corresponding sub-traits. (b) Data acquisition pipeline based on guided psychological interviews, together with participant statistics and annotation information of the *PersoMoni* dataset. (c) The personality regression task pipeline.

established personality assessment frameworks such as BFI-2 [22]. As a result, they are less able to support interpretable personality analysis and clinically meaningful prediction [23]. Importantly, BFI-2 is not merely a descriptive extension of the Big Five, but a psychometrically validated hierarchical instrument that structures broad personality domains into more specific facet-level traits to enhance bandwidth, fidelity, and predictive utility [22]. Therefore, existing benchmarks fall short in two important ways, *i.e.*, they do not provide enough temporal richness to capture stable behavioral patterns, nor enough label granularity to represent nuanced personality structure. These two deficiencies together constrain the ability of computational models to learn psychologically meaningful personality signals, while increasing the risk of overfitting to surface-level cues or dataset-specific biases [24].

To address these limitations, we introduce ***Personality Monitoring Dataset*** (***PersoMoni***), **a structured benchmark tailored for fine-grained video-based personality computing**. As shown in Fig. 1, *PersoMoni* is built on 168 full-length psychological interviews conducted by licensed counselors under semi-structured protocols, enabling rich and naturalistic social interactions. Such a setting is more appropriate for eliciting trait-relevant nonverbal behaviors than short, highly constrained, or crowd-sourced videos [10]. Unlike previous datasets, *PersoMoni* provides expert-validated personality annotations based on the BFI-2 taxonomy [25], covering continuous scores for **five higher-order traits and fifteen clinically meaningful sub-traits**. In addition, the interviews are segmented into over 23000 high-resolution face-aligned video clips. These characteristics make *PersoMoni* distinct in three aspects, *i.e.*, long-duration naturalistic interaction, expert-grounded hierarchical labeling, and fine-grained personality modeling. As a result, it can support a broader range of tasks, including behavioral dynamics research, trait regression, se-

quence modeling, and interpretable personality analysis, which further promote the implementation of affective computing.

Considering that prior deep personality recognition models typically use temporal modeling mechanisms, they inevitably fail to capture the sparse, irregular patterns through which personality-related behaviors appear [26]. To overcome these limitations from pioneer methods, building upon this dataset, we further develop ***Personality-aware Sparse Modeling Network* (*PSM-Net*)**, a sparse-aware modeling network that explicitly models the dynamic and selective nature of personality expression. Specifically, *PSM-Net* integrates spatiotemporal representation learning via lightweight 3D convolutions [27], [28], sparse temporal selection inspired by efficient video modeling methods to accommodate inter-individual variability in movement style, expressiveness, and facial appearance. Then, by emphasizing personality-salient segments while suppressing redundant frames, the model is able to focus on the subtle yet stable behavioral markers that reflect underlying personality structure.

By combining clinically reliable annotations, naturalistic long-duration videos, fine-grained trait structure, and dedicated temporal modeling techniques, *PersoMoni* and *PSM-Net* finally build a new benchmark for personality computing in high dimensions. This work not only demonstrates SOTA predictive performance but also opens new research directions in interpretable, psychologically grounded video personality analysis, which enables more reliable computational assessment of complex human traits in real-world settings.

Totally, our contributions can be summarized as follows:

- We introduce *PersoMoni*, a large-scale video-based personality dataset that comprises 168 full-length psychological interviews conducted by licensed counselors. Each participant is annotated with expert-validated BFI-2 personality scores, completely covering both the Big Five and 15 sub-traits. Unlike prior short-clip or crowdsourced datasets, *PersoMoni* try to capture long-form, naturalistic social behavior with clinically meaningful trait structure, thus establishing a rigorous foundation for interpretable personality computing.
- We propose *PSM-Net*, a novel framework that models personality expression as a sparse, selectively activated behavioral process. To accommodate inter-individual variability in expressiveness and behavioral style, the network organically integrates 3D spatiotemporal encoding, sparsity-driven temporal refinement, and dynamic instance normalization. This design enables the model to focus on subtle, psychologically salient cues distributed across extended interactions.
- A comprehensive benchmark protocol is established to evaluate models across both high-level Big Five traits and fine-grained BFI-2 sub-dimensions. Extensive experiments across multiple video backbones demonstrate that our method consistently surpasses strong baselines, confirming the value of clinically annotated long-duration videos and the effectiveness of our sparse temporal personality modeling approach.

II. Related Work

A. Psychological Foundations of Visual Personality Inference

In personality psychology, traits are generally viewed as relatively stable dispositional tendencies rather than directly observable momentary states. While personality itself is not directly visible, it is often expressed through recurring patterns of social behavior and interpersonal interaction [29]. This provides the theoretical basis for visual personality inference, where observable nonverbal behaviors are used as behavioral manifestations of underlying trait tendencies.

Studies on personality judgment and social perception suggest that facial expressions, gaze, head movements, and interpersonal style can convey personality-relevant information, particularly when sufficient interactional context is available [20], [21], [30]. In this sense, visual cues should not be treated as direct representations of personality, but as behavioral evidence from which stable trait tendencies may be inferred [10]. Moreover, BFI-2 provides a psychometrically validated hierarchical framework that represents both broad personality domains and finer facet-level traits. Using this framework allows the proposed benchmark to remain grounded in established personality assessment practice while supporting more interpretable and fine-grained computational modeling [25].

B. Personality-related Tasks and Datasets

Personality trait identification (PTI) aims to infer internal psychological traits from externally observable behaviors, digital traces, or multimedia content [31], [32]. Within computational psychology, the BFPT framework remains the dominant target space due to its cross-cultural robustness and hierarchical organization [25]. Accordingly, PTI models typically predict either continuous trait scores or discretized personality labels from multimodal observations [33].

The development of annotated benchmarks has substantially advanced audiovisual PTI, from early short self-introduction and social-media videos [18], [24] to more naturalistic interaction settings and richer multimodal cues [23], [34]. Related resources have also incorporated physiological or affective signals, extending PTI toward broader multimodal personality analysis [35]–[37].

Despite these advances, existing personality datasets still present three major limitations. First, most video samples remain short and socially constrained, which makes it difficult to capture the temporally distributed behavioral patterns relevant to personality expression [38]. In addition, crowdsourced annotations often reflect first-impression judgments rather than clinically grounded assessments [39]. Second, current benchmarks usually focus on coarse Big Five traits and rarely incorporate hierarchical sub-trait structures, limiting their usefulness for fine-grained and psychologically interpretable personality analysis [25], [40]. Third, many existing resources lack structured interaction scenarios, expert evaluation, and sufficient attention to demographic variation, resulting in limited interpretability and weak alignment with psychological and clinical practice [15], [41], [42].

To address these limitations, we introduce *PersoMoni*, a clinically grounded personality dataset collected from con-

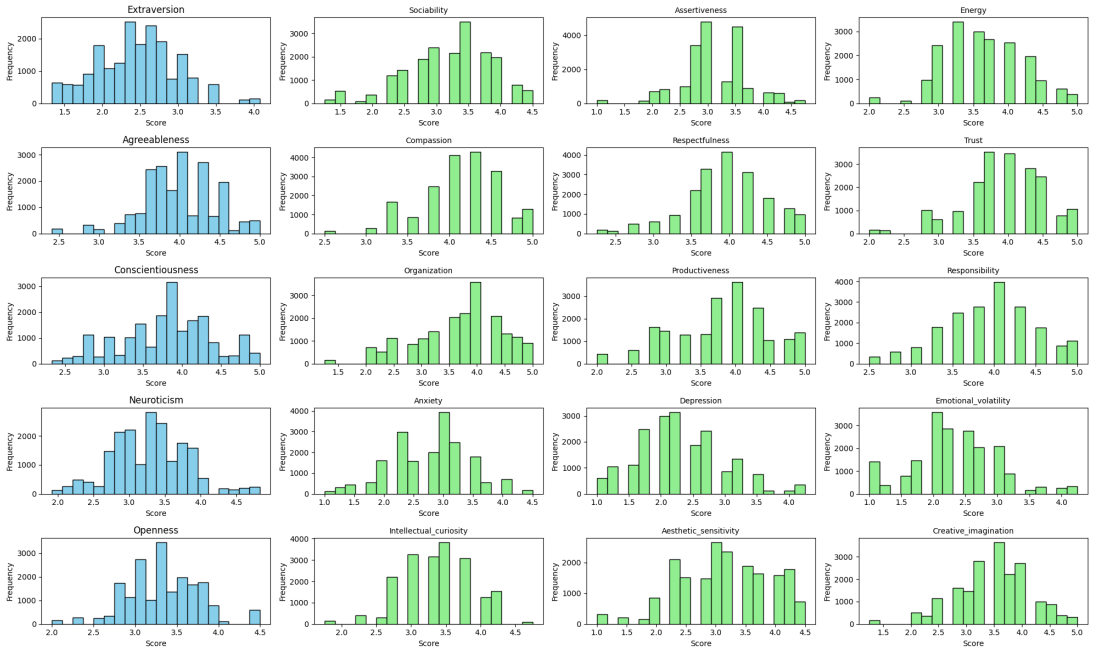


Fig. 2: Distribution of the BFI-2 personality traits in the *PersoMoni* dataset.

trolled psychological interview scenarios. *PersoMoni* incorporates professionally administered BFI-2 assessments [25], enabling analysis of both the Big Five traits and 15 sub-traits. It is built from long-form psychological interviews and further segmented into face-centered clip samples for model training. This design preserves the behavioral richness of extended interactions while providing localized observational units for learning personality-relevant temporal cues. Compared with prior short-video or crowd-annotated resources, this design is better aligned with the observation of trait-relevant behavioral regularities over time, and thus supports more reliable, interpretable, and psychologically grounded personality recognition under naturalistic interaction settings.

C. Personality Information Processing Methods

Alongside the development of PTI benchmarks, personality analysis methods have evolved from handcrafted social signal modeling toward deep, data-driven behavior understanding. Early approaches mainly relied on manually designed audiovisual descriptors, *e.g.*, facial action units, gaze, prosody, and interaction patterns [43], and used linear regression or Support Vector Machines (SVM) to map these features to personality ratings [44]. Although informative, such methods offered limited capacity for modeling temporal dynamics and fine-grained behavioral variation.

With the rise of deep learning, CNN-based models became widely adopted for extracting visual cues from frames or short clips, while tools such as OpenFace improved the reliability of behavior-related measurements through facial landmarks, head pose, and Action Units [45], [46]. Static-image personality studies further showed that appearance cues may carry personality-related signals, but their predictive capacity remains inherently limited by the lack of temporal context [44], [47]. Because personality-related judgments are typically supported by sustained behavioral patterns rather than isolated visual snapshots [20], [21], [48], temporal modeling has become central to video-based PTI. Early systems used RNNs and GRUs to aggregate short-frame sequences [43], while later methods introduced attention-based selection and audiovisual fusion to highlight psychologically salient moments and improve robustness [24], [49], [50]. More recent architectures adopted hybrid classification–regression designs, semantic alignment modules, and Transformer-based modeling to capture longer-range dependencies [17], [51]–[53].

Despite this progress, existing PTI methods still face two important challenges. On the one hand, current benchmarks often provide only coarse trait supervision, making it difficult to support fine-grained and psychologically interpretable inference. On the other hand, long-form behavioral sequences contain substantial redundancy, while personality-relevant cues are often weak, sparse, and unevenly distributed over time. This motivates the need for a modeling framework that can both exploit long-range behavioral evidence and selectively emphasize personality-salient temporal patterns. To this end, our work introduces a clinically grounded benchmark together with a sparse temporal personality modeling framework that explicitly captures long-term behavioral stability while mitigating spurious correlations.

III. New Dataset

A. Data Collection, Processing and Annotation

Modeling fine-grained personality traits from visual signals is challenging because relevant cues are often subtle, weak, and temporally distributed. In addition to the scarcity of suitable benchmarks, it is difficult to ensure both behavioral naturalness and annotation reliability in data collection. In this work, the long-form interview serves as the behavioral

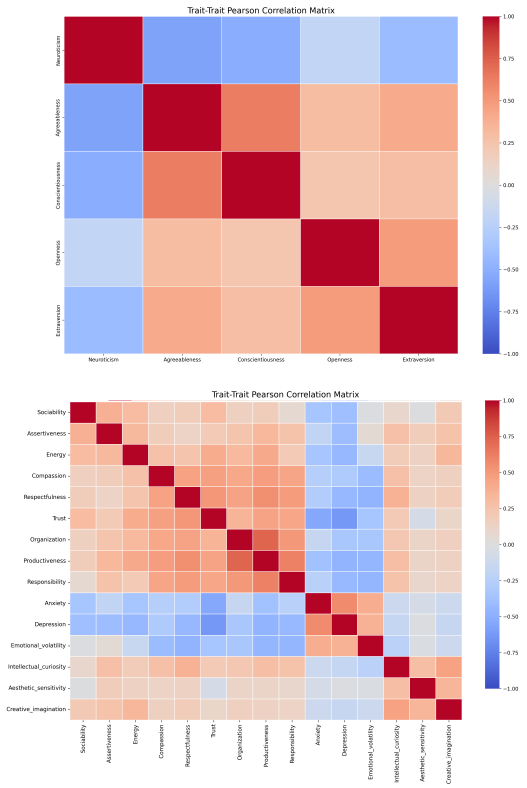


Fig. 3: Pearson correlation matrices among personality dimensions in the *PersoMoni* dataset.

source context, while short contiguous clips are used as local observational units for model training.

Existing data collection paradigms generally fall into two categories. The first involves assigning personality labels to characters in pre-existing video materials, *e.g.*, movies or vlogs. While this approach benefits from rich contextual cues and spontaneous behaviors, it suffers from uncontrolled cinematographic bias, non-neutral character roles, and fragmented editing. The second strategy adopts controlled laboratory environments, where participants are instructed to perform personality-related expressions or are exposed to emotion-inducing stimuli [54]. Although such settings allow better control over experimental variables, they often yield exaggerated or suppressed behaviors that lack ecological validity and conversational context.

1) Data Collection: In contrast, as shown in Fig. 1 b, to balance behavioral naturalness and annotation precision, we adopt a semi-structured psychological interview protocol. Each participant engages in an interactive interview with a licensed psychological counselor, who poses open-ended questions concerning self-concept, interpersonal relationships, and personal values. This setting provides a relatively natural and semantically rich interaction context, while avoiding the role bias of edited media and the artificiality of highly scripted laboratory tasks. Furthermore, prior to the interview, participants complete the SCL-90 [55] and BFI-2 to assist in tailoring the questioning. Through real-time feedback and follow-up guidance, the counselor facilitates authentic verbal and nonverbal expression, including spontaneous smiles, gaze

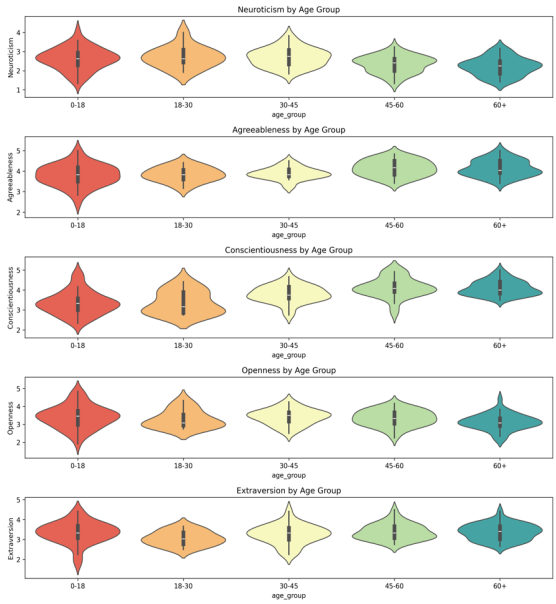


Fig. 4: Statistical distributions of Big Five personality traits across different age demographics in *PersoMoni*.

shifts, facial muscle movements, and head posture changes. In particular, these behaviors are treated as observable cues distributed across interaction, rather than as isolated facial events directly mapped to stable traits.

During the interview, all participants are recorded using front-facing cameras and remain seated in a natural posture throughout the session. The framing ensures full coverage of facial expressions and eye contact with temporal continuity and spatial alignment, enabling reliable downstream visual analysis. This recording setup reduces viewpoint variation and preserves continuous behavioral streams for subsequent face-centered temporal modeling.

2) Data Processing: After data collection, each full-length interview video is further converted into face-centered visual samples for subsequent modeling. In this process, the facial region of the participant is automatically localized frame by frame, and temporally consistent face crops by YOLO [56] are extracted and normalized into a unified spatial format. We then organize these processed facial frames into short clip units for downstream learning. This processing reduces irrelevant variations in background and camera geometry while preserving the temporal continuity of facial and gaze-related behaviors, thereby providing more focused local observations of personality-relevant nonverbal cues.

3) Data Annotation: After the interview, personality annotations are conducted following the BFI-2 taxonomy, comprising both five broad personality traits and 15 sub-traits. To reduce labeling errors, two psychology professionals independently annotate each full-length interview. To ensure annotation consistency, we adopt a three-stage quality control protocol, *i.e.*, individual scoring, cross-review, and consensus discussion in case of disagreement. Final scores are obtained by averaging the two raters' assessments of each full-length interview. By combining structured interview elicitation with expert BFI-2 scoring at the full-interview level, the dataset

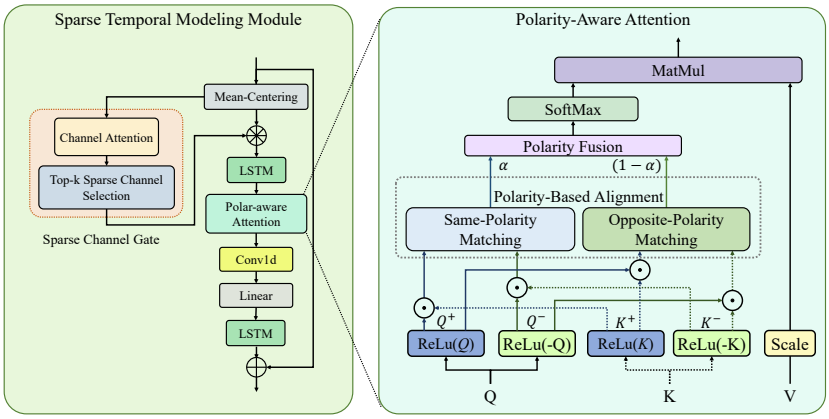


Fig. 5: Architecture of the proposed Sparse Temporal Modeling Module.

links observable behavioral evidence to clinically meaningful and hierarchically organized personality labels.

B. Dataset Statistics and Properties

1) Distribution of the personality values: PersoMoni consists of 168 participants, each contributing one full-length psychological interview video. Ethical approval was obtained prior to data collection, and all participants provided informed consent for academic use of their data. All recordings were captured using 1080p front-facing cameras at 30 frames per second, with durations ranging from 20 to 30 minutes. To support computational modeling, all videos were processed into face-centered clip samples through face detection and temporal segmentation, yielding over 23000 clips. As shown in Fig. 1a, each interview is annotated with a 20-dimensional personality score vector based on the BFI-2 framework, covering five higher-order traits, *i.e.*, Extraversion, Agreeableness, Conscientiousness, Negative Emotionality, and Openness, together with their 15 corresponding sub-traits. The final labels are provided by two licensed psychologists through independent assessment and obtained by averaging their scores.

The dataset demonstrates several desirable properties for fine-grained personality modeling. First, the annotation process is clinically grounded, relying on professional assessment rather than crowd-based impressions, ensuring interpretability and psychological validity. Second, the label structure is fine-grained and hierarchical, enabling both coarse-level and trait-specific analysis. Third, the use of long-form, continuous interview sessions preserves rich temporal dynamics, capturing subtle and evolving behavioral cues that are difficult to observe in short clips. Additionally, the dataset includes a demographically balanced set of participants, with 86 females and 82 males, spanning various age groups and educational backgrounds, which enhances model generalization. Finally, the distribution of personality scores across all traits is approximately normal, covering a wide range of values. This supports not only continuous regression-based modeling, but also potential trait discretization and classification.

To further illustrate the label characteristics, Fig. 2 visualizes the normalized distributions of all 20 traits. The five broad traits exhibit generally symmetric and well-spread distributions, while the sub-traits present more diverse patterns. These characteristics confirm that *PersoMoni* offers a structured yet diverse personality space, making it well-suited for both general and trait-specific PTI.

2) Trait Correlation Structure of Personality: To assess the psychological validity of *PersoMoni*, we analyze the correlation structure of both the aggregated Big-Five domains and the fifteen BFI-2 sub-traits. Fig. 3 presents the corresponding Pearson correlation matrices across all subjects, revealing a clear block structure consistent with the hierarchical organization of BFI-2. This pattern provides an internal validity check for the annotation scheme, suggesting that the collected ratings preserve the expected psychometric organization of the BFI-2 framework. Specifically, sub-traits belonging to Extraversion, Agreeableness, and Conscientiousness show strong positive correlations within each group, indicating coherent and stable behavioral patterns captured by expert annotations. Besides, consistent with the BFI-2 framework [22], [25], Anxiety, Depression, and Emotional Volatility form a distinct cluster within the Negative Emotionality domain. These traits are systematically and negatively correlated with positive traits, *e.g.*, Sociability and Responsibility. Such patterns mirror the established psychometric distinction between the affective vulnerability inherent in Negative Emotionality and the adaptive emotional functioning characteristic of Extraversion and Conscientiousness [57], [58].

Overall, these correlation structures indicate that *PersoMoni* preserves stable and theoretically aligned personality relationships. This supports the reliability of the expert annotations and provides a sound basis for modeling trait dependencies and multi-dimensional personality inference.

3) Age-Related Personality Distribution Patterns: To further examine the population-level psychological properties of *PersoMoni*, we visualize the distributions of the Big Five traits across five age groups using violin plots in Fig. 4. Overall, the observed age-related patterns are broadly consistent with prior lifespan personality research. Specifically, Neuroticism shows a gradual decline from adolescence to mid-adulthood, followed by greater variability in older age, echoing previous findings that negative emotionality tends to decrease across adulthood despite some cross-sample heterogeneity [59]. Agreeableness and Conscientiousness increase steadily with age and become

more stable in middle-to-late adulthood, in line with widely reported developmental trends for these traits [60]. Openness exhibits a mild inverted-U pattern, peaking in early-to-mid adulthood and declining later in life, which also agrees with earlier evidence of relatively higher openness in younger adults and lower levels in older adults [61]. Extraversion remains comparatively stable across the lifespan, although older adults show slightly greater dispersion, consistent with prior work suggesting that age-related variation in Extraversion is generally weaker or less uniform than that of Agreeableness and Conscientiousness. Taken together, these age-dependent distributional patterns suggest that *PersoMoni* captures meaningful personality development trends, supporting both the reliability of its annotations and its value for studying personality variation across age groups.

C. Task Definition

Given a video sequence $V = \{I_1, I_2, \ldots, I_T\}$ captured from a psychological interview, where I_t denotes the t-th frame and T is the total number of frames, the goal of fine-grained personality recognition is to learn a mapping function $f(\cdot)$ that predicts a continuous personality trait vector $\hat{Y} \in \mathbb{R}^D$. Here, $D = 5$ or 15 corresponds to the total number of BFI-2 taxonomy, comprising both five broad personality traits and fifteen associated sub-traits.

Thus, the task is formulated as a multi-dimensional regression problem and can be represented as:

$$\hat{Y} = f(V; \theta), \tag{1}$$

where θ denotes the learnable parameters of the model. Each component $\hat{y}_i \in \hat{Y}$ represents the model's estimation of the extent to which a specific trait is expressed in the behavioral dynamics within V. The ground-truth label $Y \in \mathbb{R}^D$ is obtained from professional annotations based on full-session observations, and the goal is to minimize the discrepancy between $\hat{Y}$ and Y using Mean Squared Error (MSE) loss. Our goal is to develop a model that captures these distributed temporal signals and encodes them into semantically meaningful representations, enabling accurate regression of both global and trait-specific personality scores from behavioral videos.

D. Baseline Models

To provide a comprehensive benchmark for fine-grained personality regression, we evaluate seven different representative models that are widely used for visual representation learning in affective computing. From a methodological structure perspective, these baseline models span three architectural paradigms as follows:

2D CNN-based methods: ResNet18, ResNet18+LSTM, and ResNet50+LSTM extract spatial features from individual frames, with LSTM variants modeling temporal dependencies across frame sequences.

3D CNN-based methods: R3D and X3D capture short-term spatiotemporal patterns directly through 3D convolutions applied over stacked video frames.

Transformer-based methods: VideoMAE and TimeSformer adopt attention-based architectures to model long-range temporal dependencies across full video clips.

To ensure a fair comparison, we preserve the original backbone of each model while appending a standardized regression head, which consists of a fully connected layer mapping visual features to a 5 or 15-dimensional BFI-2 personality score vector. Furthermore, all models are trained using the same supervision setup and loss function, enabling direct performance comparison across personality dimensions.

E. PSM-Net framework

To construct fine-grained personality representations from long-term facial behavior sequences, we propose *PSM-Net*. The pipeline first segments each full interview video into fixed-length clips and employs a 3D CNN to extract local spatiotemporal features from each clip, capturing subtle motion and expression dynamics. The clip-level features are then concatenated into a long sequence and fed into the proposed Sparse Temporal Modeling Module (STMM), which selectively enhances behavior signals most relevant to personality traits. Subsequently, the features are aggregated and refined through a BiLSTM and a self-attention layer. Finally, to regress continuous personality scores, a lightweight temporal averaging layer and a fully connected prediction head are applied, covering both the 5-dimensional Big Five traits and the 15 BFI-2 sub-traits. The overall architecture of *PSM-Net* is illustrated in Fig. 1 c.

1) Sparse Temporal Modeling Module: Long-form facial behavior sequences often contain substantial redundancy, where many actions are weakly related to personality traits. In contrast, diagnostically meaningful cues tend to be sparse, unevenly distributed, and expressed across multiple temporal scales. To address this challenge, we introduce the Sparse Temporal Modeling Module (STMM), which comprehensively extracts personality-relevant behavioral structures through channel-wise sparsification, hierarchical temporal modeling, and a polarity-aware attention mechanism, as illustrated in Fig. 5. First, given a 3D CNN–encoded sequence:

$$X = \{x_t\}_{t=1}^{T}, \qquad X \in \mathbb{R}^{T \times C}, \tag{2}$$

where T is the temporal length and C is the channel dimension, STMM first applies a channel attention module to estimate an implicit importance distribution:

$$s = \sigma(W_2\, \phi(W_1\, g(X))), \tag{3}$$

where $g(\cdot)$ denotes temporal global average pooling, $W_1 \in \mathbb{R}^{C \times C/r}$ and $W_2 \in \mathbb{R}^{C/r \times C}$ are learnable projections with reduction ratio r, $\phi(\cdot)$ is a nonlinear activation, and $\sigma(\cdot)$ is a sigmoid function producing channel-wise importance scores $s \in \mathbb{R}^C$. A Top-k sparse channel selection then identifies behavior-relevant channels:

$$m_i = \mathbf{1}(s_i \geq \tau_\rho(s)), \qquad i = 1, \ldots, C, \tag{4}$$

where $\rho \in (0, 1)$ is the sparsity ratio and $\tau_\rho(s)$ denotes the threshold that selects the top ρC channels. The resulting sparse representation is:

$$X_s = X \odot m, \tag{5}$$

where $\odot$ denotes channel-wise Hadamard product with temporal broadcasting. Before temporal modeling, STMM performs mean-centering, i.e., subtracting the temporal mean of X_s to suppress global activation bias and emphasize dynamic variations:

$$\tilde{X}_s = X_s - \mu(X_s).$$

The local stage models short-term micro-dynamics via an LSTM:

$$H_l = \mathrm{LSTM}(\tilde{X}_s), \tag{6}$$

where $H_l \in \mathbb{R}^{T\times d}$ is the locally encoded sequence.

After modeling fine-grained micro-dynamics, it remains necessary to address challenges such as cross-segment temporal misalignment and the asymmetric expression of positive and negative behavioral patterns, both of which are prevalent in personality-related behaviors. Although conventional self-attention can capture global dependencies, it treats positive and negative activations uniformly and thus struggles to characterize the "polarized dynamics" typical in personality expression (e.g., approach vs. avoidance, positive vs. negative affective tendencies). To more effectively distinguish polarity-specific temporal patterns and enhance long-range dependency modeling, STMM introduces a polarity-aware attention mechanism in the global stage. First, linear projections generate queries, keys, and values:

$$Q,\, K,\, V = f(H_l), \tag{7}$$

where $f(\cdot)$ is a learnable linear layer. To separate positive and negative activation patterns, STMM decomposes Q and K using ReLU-based polarity operators:

$$Q^+ = \mathrm{ReLU}(Q), \qquad Q^- = \mathrm{ReLU}(-Q), \tag{8}$$

$$K^+ = \mathrm{ReLU}(K), \qquad K^- = \mathrm{ReLU}(-K), \tag{9}$$

corresponding to the polarity-based kernel mappings $\Psi^+(\cdot)$ and $\Psi^-(\cdot)$. Same-polarity and opposite-polarity alignments are computed as:

$$A^{\text{same}} = Q^+(K^+)^\top + Q^-(K^-)^\top, \tag{10}$$

$$A^{\text{opp}} = Q^+(K^-)^\top + Q^-(K^+)^\top. \tag{11}$$

A learnable fusion factor $\alpha \in [0,1]$ balances the two components:

$$A = \alpha\, A^{\text{same}} + (1-\alpha)\, A^{\text{opp}}. \tag{12}$$

A softmax operation yields normalized attention weights:

$$\hat{A} = \mathrm{Softmax}(A), \tag{13}$$

and the global representation is obtained via:

$$H_g = \hat{A}V. \tag{14}$$

To enforce temporal compactness and generate structured trajectories, STMM employs a conv–linear transformation:

$$Z = f_{\text{seg}}(H_g), \tag{15}$$

where $f_{\text{seg}}(\cdot)$ is implemented by a 1D convolution that compresses short temporal windows into segment-level descriptors. A reconstruction module:

$$Y = f_{\text{recon}}(Z), \tag{16}$$

consisting of a linear projection followed by an LSTM, restores a temporally aligned sequence.

Finally, a global baseline is added to stabilize inter-individual variation:

$$\hat{X} = Y + \mu(X), \tag{17}$$

where $\mu(X)$ denotes the temporal mean of the original input sequence. Overall, by integrating sparse channel gating, hierarchical temporal modeling, and polarity-aware attention, STMM effectively distills stable and personality-relevant dynamic cues from highly redundant behavioral sequences.

model	Validation MSE					
	Mean	Neu.	Agre.	Con.	Ope.	Ext.
resnet18 [62]	0.3508	0.4602	0.2504	0.3773	0.3549	0.3115
resnet18+LSTM [63]	0.3496	0.4738	0.2365	0.3954	0.3540	0.2882
resnet50+LSTM [63]	0.3299	0.4341	0.2351	**0.3344**	0.3187	0.3270
R3D [64]	0.3443	0.4708	0.2572	0.3749	0.2994	0.3192
X3D [65]	0.3186	0.3869	0.2214	0.3882	0.2651	0.3311
VideoMAE [66]	0.3177	0.3914	0.2258	0.3906	0.2596	0.3212
TimeSformer [67]	0.3166	**0.3839**	0.2244	0.3934	**0.2580**	0.3235
PSM-Net (Ours)	**0.3135**	0.3857	**0.2186**	0.3632	0.2812	**0.3188**

TABLE I: Quantitative comparison of validation MSE on the five higher-order personality traits from the BFI-2 framework.

2) Temporal Aggregation and Personality Regression: After filtering and enhancing the raw clip-level features by STMM, the model obtains a compact and personality-relevant behavioral sequence. Then, to further integrate long-range temporal dependencies and produce a stable personality representation, *PSM-Net* adopts a lightweight yet effective temporal aggregation pipeline. More specifically, the sequence is first processed by a BiLSTM, which captures forward–backward contextual cues essential for personality expression. Compared with a unidirectional structure, the BiLSTM better models continuous behavioral transitions commonly observed in interview scenarios, *e.g.*, gradual shifts in facial tension or sustained engagement patterns, leading to smoother and more coherent temporal representations. Next, a lightweight self-attention module is introduced to adaptively assign importance weights to different temporal segments. This mechanism highlights behavior segments that convey stronger personality cues while suppressing sporadic or noisy movements. Finally, to derive the global personality descriptor, *PSM-Net* applies temporal average pooling to the enhanced sequence, yielding a length-invariant and semantically stable feature vector. This vector is then fed into a linear prediction head to produce continuous personality scores.

IV. Experiments

A. Evaluation Metrics

We evaluate model performance using the MSE, a standard metric for regression tasks. Given the predicted score $\hat{y}_{i,j}$ and

model	Validation MSE														
	Mean	Neuroticism			Agreeableness			Conscientiousness			Openness			Extraversion		
		Soc.	Ass.	Ene.	Com.	Res.	Tru.	Org.	Pro.	Res.	Anx.	Dep.	Emo.	Int.	Aes.	Cre.
resnet18 [62]	0.4670	0.6709	**0.2697**	0.4757	0.2839	0.2695	**0.3616**	0.4036	0.4232	0.5296	0.3460	0.4134	0.8936	0.3884	0.6721	0.6045
resnet18+LSTM [63]	0.4412	0.5599	0.2990	0.5018	0.3329	**0.2395**	0.3628	0.3847	0.3944	0.5414	0.2916	0.3997	0.7948	0.3694	0.5790	0.5677
resnet50+LSTM [63]	0.4314	0.5302	0.2888	0.4500	0.2849	0.2746	0.3781	0.4047	0.4182	0.5168	**0.2692**	0.3959	0.8092	0.3529	0.5559	0.5410
R3D [64]	0.4383	**0.4707**	0.2906	0.4912	0.3204	0.2677	0.3780	**0.3679**	0.4092	0.5348	0.2799	0.3833	0.7499	0.3630	0.7022	0.5653
X3D [65]	0.4317	0.5443	0.2939	0.4563	0.2970	0.2643	0.3653	0.3900	**0.3700**	**0.4974**	0.2967	0.3965	0.8316	0.3433	0.5685	0.5600
VideoMAE [66]	0.4186	0.5118	0.2890	0.4498	0.2826	0.2742	0.3669	0.3981	0.3970	0.5102	0.2739	0.3741	0.7535	0.3239	0.5522	**0.5218**
TimeSformer [67]	0.4182	0.5112	0.2898	0.4496	**0.2821**	0.2739	0.3670	0.3952	0.3931	0.5086	0.2740	**0.3739**	0.7574	**0.3227**	**0.5511**	0.5236
PSM-Net (Ours)	**0.4175**	0.5432	0.2873	**0.4425**	0.2824	0.2641	0.3637	0.3966	0.3867	0.5084	0.2709	0.3824	**0.7097**	0.3244	0.5763	0.5245

TABLE II: Validation MSE results across the 15 BFI-2 sub-traits.

T	Big-5	Sub-15
8	0.3209	0.4256
16	**0.3135**	0.4182
24	0.3146	**0.4175**
32	0.3168	0.4207

TABLE III: Effect of temporal input length T of *PSM-Net*.

ρ	Big-5	Sub-15
0.25	0.3198	0.4246
0.50	**0.3135**	**0.4175**
0.75	0.3157	0.4191

TABLE IV: Effect of sparsity ratio ρ of *PSM-Net*.

Variant	Big-5	Sub-15
Transformer Encoder	0.3178	0.4213
w/o Polarity-aware Attn.	0.3159	0.4190
Full *PSM-Net*	**0.3135**	**0.4175**

TABLE V: Ablation on temporal modeling and polarity-aware attention in *PSM-Net*.

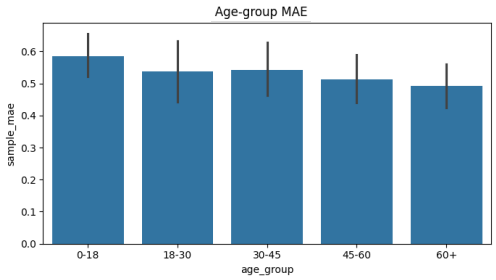


Fig. 6: Age-group MAE comparison across five demographic intervals.

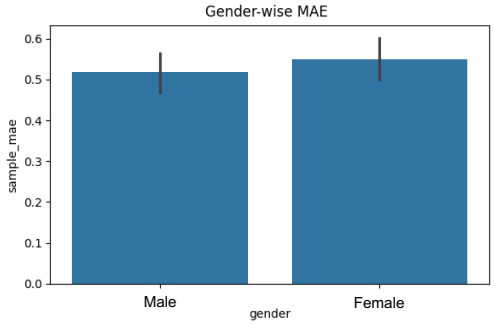


Fig. 7: Gender-wise MAE analysis between male and female subjects.

ground-truth score $y_{i,j}$ for the j-th personality trait of the i-th sample, MSE is computed as:

$$\text{MSE} = \frac{1}{ND}\sum_{i=1}^{N}\sum_{j=1}^{D}(\hat{y}_{i,j} - y_{i,j})^2, \tag{18}$$

where N is the total number of test samples and $D = 5$ or 15 corresponds to the full set of BFI-2 traits. Particularly, a lower MSE indicates better regression accuracy across both global and fine-grained traits.

B. Implementation Details

In the *PersoMoni* dataset, we uniformly sample 16 frames from each clip as the input sequence during training. The dataset split is performed strictly at the participant level while all clips from the same participant are assigned to only one subset, and there is no subject overlap across the training, and test sets. Each frame is resized to 112×112 pixels before being passed through a 3D ResNet-18 encoder.

We train the model using stochastic gradient descent (SGD) with an initial learning rate of 0.001, momentum 0.9, and weight decay 0.05. A cosine learning rate scheduler with linear warm-up over the first 5 epochs is used. The model is trained for 80 epochs with a batch size of 8. For data augmentation, color jitter is applied with a strength of 0.4, while no additional random crop or horizontal flip is used in the current implementation. All experiments are implemented in PyTorch and conducted on a single NVIDIA A6000 GPU.

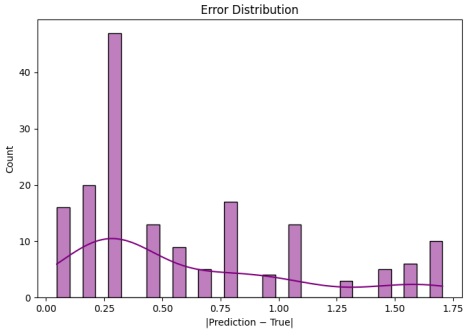


Fig. 8: Overall error distribution of personality score regression on the *PersoMoni* dataset.

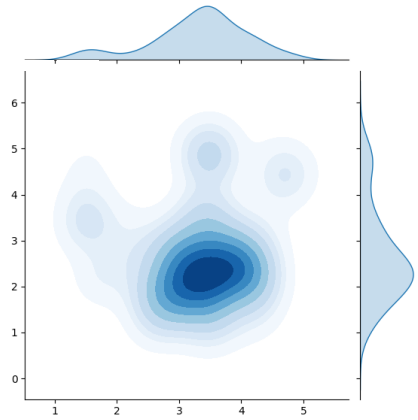


Fig. 9: Joint probability density distribution between model predictions and ground-truth personality scores.

C. Experimental Analysis

1) Overall Results: We evaluate our method against several representative baseline architectures, covering 2D CNNs, 3D CNNs, and Transformer-based models. The corresponding experimental results are summarized in Tab. I and Tab. II, which demonstrate a consistent performance hierarchy, *i.e.*, 2D CNN models tend to underperform compared to 3D CNNs, while Transformer-based architectures achieve stronger results due to their superior temporal modeling capacity. Considering that this trend is broadly consistent with existing findings in video-based affective computing research, it supports the validity of our dataset for eliciting personality-related signals at multiple temporal scales. From Tab. I and Tab. II, we observe that our proposed method outperforms all baselines across most of the 5 or 15 personality traits, especially in sub-traits associated with emotional regulation and interpersonal behavior. It shows strong capabilities in capturing fine-grained temporal dynamics through the integration of sparse temporal modeling and dynamic normalization. Furthermore, these components contribute to better generalization across both global and local trait levels.

Our method not only achieves the best overall prediction accuracy but also exhibits more stable performance across all traits compared to baseline models. While Transformers demonstrate powerful capacity in high-level traits like Openness and Extraversion, they often struggle with emotionally subtle sub-traits. In contrast, our model maintains strong regression fidelity in both categories, indicating improved robustness and interpretability.

Overall, these results support the effectiveness of the proposed *PSM-Net* framework and suggest that the *PersoMoni* dataset provides a reliable foundation for benchmarking fine-grained personality trait recognition under realistic and clinically meaningful conditions.

2) Age-group Robustness Analysis: To evaluate the stability of the model across different age populations, we partition the test set into five age intervals and compute the MAE indexes for each group. As shown in Fig. 6, the results show that the MAE values across different age groups remain consistently low and close to each other, indicating strong robustness with respect to age distribution. The $0-18$ group exhibits slightly higher error, which may be attributed to greater facial variability among adolescents, stronger fluctuations in personality scores, or a smaller number of samples in this age range. In contrast, the 60+ group achieves slightly lower error, suggesting that older samples do not introduce additional noise. Nevertheless, the model maintains stable predictive accuracy across age groups without displaying any notable age-related bias.

3) Gender-wise Performance Comparison: Since gender imbalance may introduce systematic bias in affective and personality prediction models, we further examine performance across male and female subjects [41]. We further examine model performance across gender. By grouping the test samples into male and female, we compare the MAE values between these two groups, as shown in Fig. 7. The performance gap between genders is minimal, and the confidence intervals of the MAE values substantially overlap. This indicates that the model does not exhibit any significant gender-related prediction bias. Apparently, this finding demonstrates desirable fairness in the gender dimension, as the model achieves comparable prediction accuracy for both male and female samples. Given that gender imbalance can easily introduce distributional bias in personality regression tasks, the observed stability also highlights the robustness of our approach and the completeness of the dataset.

4) Error Distribution Characteristics: To gain deeper insight into the overall prediction bias, we plot the histogram of absolute errors together with the kernel density estimation curve. As depicted in Fig. 8, the results reveal that most errors fall within the 0.2–0.5 range, indicating that the model achieves relatively high accuracy for the majority of samples. Meanwhile, the error distribution shows a clear long-tailed pattern, *i.e.*, although large-error samples (greater than 1.0) are rare, they do exist. These outliers may be caused by noise in facial expressions or inherent subjectivity in personality annotations, all of which inevitably make certain samples particularly difficult to fit. Nevertheless, this distribution confirms the model's strong general stability, while also suggesting potential improvement directions, *e.g.*, hard-sample augmentation and more robust feature modeling.

5) Prediction–Ground Truth Joint Distribution: The joint scatter–density visualization in Fig. 9 depicts the global relationship between predicted and ground-truth personality scores. Most samples cluster tightly within a dominant high-

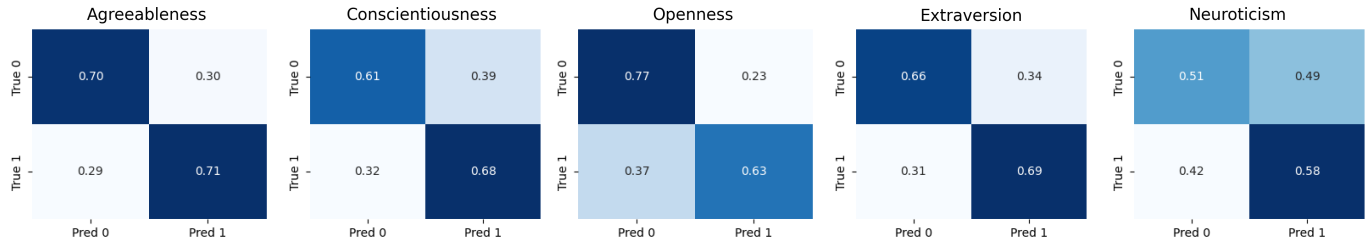


Fig. 10: Confusion matrices of binary personality prediction across the five Big Five traits.

density region, indicating that the model captures the overall mapping between facial behavior and personality traits. Compared with the ground-truth distribution, the predicted values exhibit a slightly narrower spread, suggesting reduced variance in the model outputs. The density center lies marginally above the ground-truth center, implying a mild overestimation tendency rather than underestimation. In addition, a few low-density islands appear around the main cluster, revealing that extreme or noisy samples still cause localized deviations. Overall, the distribution shows that the model fits global personality tendencies well, while modeling fine-grained individual differences remains a challenge.

6) Big Five Personality Classification Application: To further assess the applicability of our dataset to personality recognition tasks, we conduct binary classification experiments under the Big Five personality framework. We divide each Big Five personality trait into two categories, high and low, using the median score of each personality category, and visualize the results using confusion matrices for each trait. Fig. 10 presents the confusion matrices of binary personality prediction across the five Big Five traits. Overall, the classification structures exhibit consistent patterns, with clear diagonal dominance across most traits, indicating that the visual behavioral cues in the interviews contain sufficient information to distinguish between high and low personality levels. The prediction distributions are relatively balanced across classes, suggesting the absence of major bias toward either positive or negative categories.

Among the five traits, Neuroticism shows the weakest separability, with more mixed entries in the confusion matrix, which aligns with the known difficulty of estimating internalized emotional tendencies purely from visual signals. Nevertheless, all traits maintain structurally similar confusion matrices, implying that video-based behavior conveys cross-dimensional consistency in personality-related expressions. As personality scores are continuous and subjective by nature, borderline samples inevitably lead to some misclassification in the binarized setting. These errors most likely arise from the inherent ambiguity near the decision threshold rather than model deficiencies.

7) Ablation on Key Hyper-parameters: We further analyze two key hyper-parameters in our framework, namely the temporal input length T and the sparsity ratio ρ, and additionally provide a compact ablation on the temporal modeling module. As shown in Tab. III, using too short a temporal window leads to inferior performance on both tasks, suggesting that limited temporal context is insufficient to capture subtle and distributed nonverbal cues related to personality. Increasing the input length improves the regression results, but the optimal temporal span is slightly task-dependent, indicating that fine-grained sub-traits may benefit from slightly richer temporal context than broader trait dimensions.

We next examine the effect of the sparsity ratio ρ in the Top-k selection module. As reported in Tab. IV, overly small ρ values cause noticeable performance degradation, while overly large ρ weakens the benefit of sparse selection by retaining too many redundant features. A moderate sparsity ratio achieves the best balance between preserving informative behavior patterns and suppressing irrelevant variations.

Consistent with this trend, Tab. V further shows that replacing the proposed temporal modeling module with a standard Transformer encoder leads to inferior performance, while removing the polarity-aware attention also results in a consistent drop. Overall, these results support our design choice that personality-related cues are sparse but not isolated, and therefore benefit from moderate temporal coverage together with selective sparse modeling and polarity-aware attention.

V. Conclusion

In this paper, we present *PersoMoni*, a structured and clinically grounded video-based personality computing benchmark dataset. This dataset contains 168 full-length psychological interview videos with expert-annotated scores across 20 personality traits based on the BFI-2 framework. Compared to existing resources, *PersoMoni* offers high-resolution temporal segmentation, continuous regression labels, and a rich spectrum of interpersonal and emotional expressions, thereby enabling more interpretable and fine-grained personality analysis and further supporting various potential tasks, *e.g.*, behavioral dynamics research, trait regression, sequence modeling, and interpretable personality analysis under naturalistic conditions. To leverage this dataset, we propose *PSM-Net*, a Personality-aware Sparse Modeling Network that integrates 3D CNNs for spatial-temporal encoding, and sparse modeling for temporal structure refinement. Extensive experiments on both Big-Five and sub-trait regression tasks demonstrate the effectiveness of our method in capturing subtle personality-related signals and outperforming several competitive baselines.

VI. Future Research Directions

As for our future works, several promising directions can further advance fine-grained personality computing. First, *PersoMoni* can be further enriched with more fine-grained annotations, such as behavior-level labels, interaction patterns,

and temporally localized personality-related events, so as to better connect global trait scores with interpretable behavioral evidence. Second, incorporating hierarchical structures between high-level and sub-traits can better model the semantic correlations defined in personality taxonomies like BFI-2. Third, future work should explore the use of large multimodal models for long-video personality understanding, since they may better capture long-range temporal dependencies, interaction semantics, and cross-event behavioral consistency in full-length interviews. Meanwhile, given the current subject scale of *PersoMoni*, future work should further expand the number and diversity of participants and improve annotation granularity to better support deep spatiotemporal modeling and reduce potential overfitting risks. Finally, future efforts should also explore multimodal extensions, real-world robustness, and interpretability, enabling broader deployment in psychologically grounded applications.

ACKNOWLEDGMENTS

We would like to thank the ethics committee of the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, for supervising the collection and usage of the dataset. As well, we would like to thank all the participants interviewed for the dataset. Prior to data collection, we informed each participant of the requirement for data collection and obtained their signed consents for academic research purposes.

This work is supported by National Natural Science Foundation of China (Grant No. 62302145, 72188101, 62272144), National Key R&D Program of China (NO.2024YFB3311602), the Anhui Provincial Natural Science Foundation (2408085J040), Major Project of Anhui Provincial Science and Technology Breakthrough Program (202423k09020001), the Fundamental Research Funds for the Central Universities (JZ2024AHST0337, JZ2025HGTB0225), the Major Scientific and Technological Project of Anhui Provincial Science and Technology Innovation Platform (Grant No. 202305a12020012), and the New Cornerstone Science Foundation through the XPLORER PRIZE.

REFERENCES

[1] J. Li, H. Zhang, J. Chen, J. Gong, Y. Wang, S. Wang, and Y. Zhao, "Mprnet: A temporal-aware cross-modal encoding framework for personality recognition," *IEEE Transactions on Affective Computing*, 2025.

[2] S. Song, Z. Shao, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Learning person-specific cognition from facial reactions for automatic personality recognition," *IEEE Transactions on Affective Computing*, 2023.

[3] D. J. Ozer and V. Benet-Martinez, "Personality and the prediction of consequential outcomes," *Annu. Rev. Psychol.*, 2006.

[4] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou, "Rethinking the learning paradigm for dynamic facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[5] X. Zhang, Y. Lu, H. Yan, J. Huang, Y. Gu, Y. Ji, Z. Liu, and B. Liu, "Resup: Reliable label noise suppression for facial expression recognition," *IEEE Transactions on Affective Computing*, 2025.

[6] F.-Q. Cui, A. Tong, J. Huang, J. Zhang, D. Guo, Z. Liu, and M. Wang, "Learning from heterogeneity: Generalizing dynamic facial expression recognition via distributionally robust optimization," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

[7] P. Zhang, M. Hu, H. Zhang, C. Wu, and Z. Yang, "Personality-aware multimodal driver emotion recognition towards intelligent connected vehicles," *IEEE Transactions on Affective Computing*, 2025.

[8] J. Huang, Y. Feng, F.-Q. Cui, X. Zhang, Z. Liu, X. Liu, J. Liu, F. Zhang, and M. Li, "Identifying who you are no matter what you write through abstracting handwriting style," *IEEE Transactions on Dependable and Secure Computing*, 2026.

[9] D. C. Funder, "On the accuracy of personality judgment: A realistic approach," *Psychological Review*, 1995.

[10] ——, "Accurate personality judgment," *Current Directions in Psychological Science*, 2012.

[11] Y. Gao, H. Shi, Y. Fu, C. Chu, and T. Kawahara, "Bridging speech emotion recognition and personality: Dataset and temporal interaction condition network," *IEEE Transactions on Affective Computing*, 2025.

[12] R. Wang, J. Huang, J. Zhang, X. Liu, X. Zhang, Z. Liu, P. Zhao, S. Chen, and X. Sun, "Facialpulse: An efficient rnn-based depression detection via temporal facial landmarks," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

[13] K. Gönç and H. Dibeklioğlu, "Affect and personality aided modeling of transcribed speech for depression severity estimation," *IEEE Transactions on Affective Computing*, 2025.

[14] Y. Zhao, H. Zhang, J. Li, S. Song, C. Lian, Y. Liu, Y. Wang, and C. Fu, "Multimodal depression assessment framework integrating personality and gait for older adults with medical conditions," *IEEE Transactions on Affective Computing*, 2025.

[15] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, 2014.

[16] F.-Q. Cui, Z. Lin, X. Rao, A. Tong, S. Li, F. Wang, C. Chen, and B. Liu, "Micacl: Multi-instance category-aware contrastive learning for long-tailed dynamic facial expression recognition," in *2025 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, 2025.

[17] Z. Jia, Y. Liu, H. Wang, and T. Jiang, "Cross-modal knowledge distillation for enhanced unimodal emotion recognition," *IEEE Transactions on Affective Computing*, 2025.

[18] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *European conference on computer vision*, 2016.

[19] J. C. S. Jacques Junior, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. J. van Gerven, R. van Lier, and S. Escalera, "First impressions: A survey on vision-based apparent personality trait analysis," *IEEE Transactions on Affective Computing*, 2022.

[20] P. Borkenau, N. Mauer, R. Riemann, F. M. Spinath, and A. Angleitner, "Thin slices of behavior as cues of personality and intelligence." *Journal of personality and social psychology*, 2004.

[21] N. A. Murphy and J. A. Hall, "Capturing behavior in small doses: A review of comparative research in evaluating thin slices for behavioral measurement," *Frontiers in Psychology*, 2021.

[22] C. J. Soto and O. P. John, "The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power," *Journal of Personality and Social Psychology*, 2017.

[23] M. Leekha, S. N. Khan, H. Srinivas, R. R. Shah, and J. Shukla, "Vyaktitvanirdharan: Multimodal assessment of personality and trait emotional intelligence," *IEEE Transactions on Affective Computing*, 2024.

[24] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Gucluturk, U. Güçlü, X. Baró, I. Guyon, J. J. Junior, M. Madadi *et al.*, "Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos," *IEEE Transactions on Affective Computing*, 2020.

[25] J. Soto, "The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power," *Journal of Personality and Social Psychology*, 2016.

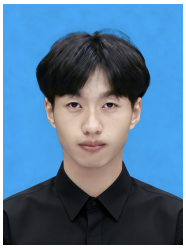
[26] Y. Ji, W. Wu, H. Lin, W. Hu, Y. Hu, L. Kang, X. Chen, and L. He, "Demographic-guided behavior patterns contrast for personality prediction," *IEEE Transactions on Affective Computing*, 2025.

[27] S. Song, S. Jaiswal, E. Sanchez, G. Tzimiropoulos, L. Shen, and M. Valstar, "Self-supervised learning of person-specific facial dynamics for automatic personality recognition," *IEEE Transactions on Affective Computing*, 2023.

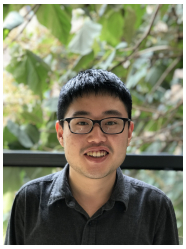
[28] H.-C. Yang and C.-C. Lee, "A media-guided attentive graphical network for personality recognition using physiology," *IEEE Transactions on Affective Computing*, 2023.

[29] Y. Shoda, S. LeeTiernan, and W. Mischel, "Personality as a dynamical system: Emergence of stability and distinctiveness from intra and interpersonal interactions," *Personality and Social Psychology Review*, 2002.

[30] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychological Bulletin*, 1992.
[31] M. K. Tellamekala, T. Giesbrecht, and M. Valstar, "Dimensional affect uncertainty modelling for apparent personality recognition," *IEEE Transactions on Affective Computing*, 2022.
[32] F. Celli, A. Kartelj, M. Đorđević, D. Suhartono, V. Filipović, V. Milutinović, G. Spathoulas, A. Vinciarelli, M. Kosinski, and B. Lepri, "Twenty years of personality computing: Threats, challenges and future directions," *arXiv preprint arXiv:2503.02082*, 2025.
[33] G. Mohammadi and P. Vuilleumier, "A multi-componential approach to emotion recognition and the effect of personality," *IEEE Transactions on Affective Computing*, 2022.
[34] R. Liao, S. Song, and H. Gunes, "An open-source benchmark of deep learning models for audio-visual apparent and self-reported personality recognition," *IEEE Transactions on Affective Computing*, 2024.
[35] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE transactions on affective computing*, 2018.
[36] M. Jammot, B. Braun, P. Streli, R. Wampfler, and C. Holz, "egoemotion: Egocentric vision and physiological signals for emotion and personality recognition in real-world tasks," in *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025.
[37] X. Ning, J. Wang, Z. Feng, T. Xin, S. Zhang, S. Zhang, Z. Lian, Y. Ding, Y. Lin, and Z. Jia, "REFED: A subject real-time dynamic labeled EEG-fNIRS synchronized recorded emotion dataset," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.
[38] T. Zhang, A. Koutsoumpis, J. K. Oostrom, D. Holtrop, S. Ghassemi, and R. E. de Vries, "Can large language models assess personality from asynchronous video interviews? a comprehensive evaluation of validity, reliability, fairness, and rating patterns," *IEEE Transactions on Affective Computing*, 2024.
[39] R. D. P. Principi, C. Palmero, J. C. S. J. Junior, and S. Escalera, "On the effect of observed subject biases in apparent personality analysis from audio-visual signals," *IEEE Transactions on Affective Computing*, 2021.
[40] W. TA and C. C, "The five factor model of personality structure: an update," *World Psychiatry*, 2019.
[41] H. Lin, C. Wang, and Y. Sun, "How big five personality traits influence information sharing on social media: A meta analysis," *Plos one*, 2024.
[42] J. Jiang, V. Manoranjan, H. Salam, and O. Celiktutan, "Towards generalised and incremental bias mitigation in personality computing," *IEEE Transactions on Affective Computing*, 2024.
[43] H.-J. Yang, G.-S. Lee, S.-H. Kim *et al.*, "End-to-end learning for multimodal emotion recognition in video with adaptive loss," *IEEE MultiMedia*, 2021.
[44] J. W. Kim and T. M. Chock, "Personality traits and psychological motivations predicting selfie posting behaviors on social networking sites," *Telematics and Informatics*, 2017.
[45] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.
[46] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.
[47] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.
[48] N. Ambady, F. J. Bernieri, and J. A. Richeson, "Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream," *Advances in Experimental Social Psychology*, 2000.
[49] Y. Güçlütürk, U. Güçlü, X. Baro, H. J. Escalante, I. Guyon, S. Escalera, M. A. Van Gerven, and R. Van Lier, "Multimodal first impression analysis with deep residual networks," *IEEE Transactions on Affective Computing*, 2017.
[50] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, and C. Chen, "Hetemotionnet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.
[51] R. Wang, X. Zhao, X. Xu, and Y. Hao, "A multimodal personality prediction framework based on adaptive graph transformer network and multi-task learning," *Computer Graphics Forum*, 2025.
[52] X. Zhao, Y. Liao, Z. Tang, Y. Xu, X. Tao, D. Wang, G. Wang, and H. Lu, "Integrating audio and visual modalities for multimodal personality trait recognition via hybrid deep learning," *Frontiers in Neuroscience*, 2023.
[53] Z. Jia, T. Du, Z. Tian, H. Li, Y. Zhang, and C. Liu, "A multimodal bimamba network with test-time adaptation for emotion recognition based on physiological signals," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
[54] D. Guo, K. Li, B. Hu, Y. Zhang, and M. Wang, "Benchmarking micro-action recognition: Dataset, methods, and applications," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[55] G. Pedersen and S. Karterud, "Using measures from the scl-90-r to screen for personality disorders," *Personality and Mental Health*, 2010.
[56] R. Sapkota and M. Karkee, "Ultralytics yolo evolution: An overview of yolo26, yolo11, yolov8 and yolov5 object detectors for computer vision and pattern recognition," *arXiv preprint arXiv:2510.09653*, 2025.
[57] K. A. Lyon, R. Elliott, K. Ware, G. Juhasz, and L. J. E. Brown, "Associations between facets and aspects of big five personality and affective disorders: A systematic review and best evidence synthesis," *Journal of Affective Disorders*, 2021.
[58] J. Luo, B. Zhang, M. Cao, and B. W. Roberts, "The stressful personality: A meta-analytical review of the relation between personality and stress," *Personality and social psychology review*, 2023.
[59] H. R. Slobodskaya, "Personality development from early childhood through adolescence," *Personality and Individual Differences*, 2021.
[60] W. J. Chopik and S. Kitayama, "Personality change across the life span: Insights from a cross-cultural, longitudinal study," *Journal of personality*, 2018.
[61] M. B. Donnellan and R. E. Lucas, "Age differences in the big five across the life span: Evidence from two national samples," *Psychology and Aging*, 2008.
[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2016.
[63] S. Tang, C. Li, P. Zhang, and R. Tang, "Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.
[64] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2018.
[65] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2020.
[66] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *The Thirty-Sixth Annual Conference on Neural Information Processing Systems*, 2022.
[67] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *International Conference on Machine Learning*, 2021.

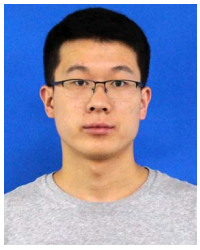


Feng-Qi Cui (Student Member, IEEE) is currently pursuing the Ph.D. degree with the MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, School of Information Science and Technology, University of Science and Technology of China, Hefei, China. He is also affiliated with Anhui Provincial Key Laboratory of Affective Computing and Advanced Intelligent Machines, Institute of Artificial Intelligence, Hefei Comprehensive National Science Center. His research interests include computer vision and brain-inspired intelligence.



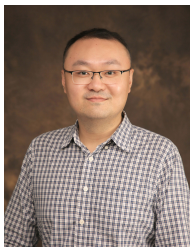
Jinyang Huang (Member, IEEE) is an Associate Professor at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT). His research interests include Multimodal Perception, Human-computer Interaction, Wireless Security, and Signal Processing. In this area, he has published 62 papers in international peer-reviewed journals and conferences, including ToN, TMC, TIFS, TDSC, MobiCom, IEEE S&P, USENIX Security, Ubicomp, NeurIPS, INFOCOM, and ACM MM. He has served as a TPC member for conferences, including ACM MM, IEEE ICME, and Globecom, and has the honor of becoming ACM MM 2024 Outstanding Reviewers. He is a Guest Editor for the Technical Committee of ICME 25 Special Session. He is the recipient of the Young Scientist of Anhui Computer Federation and IEEE HITC Distinguished PhD Dissertation Award.

Sirui Zhao received the PhD degree with the Department of Computer Science and Technology from University of Science and Technology of China (USTC). He is also a faculty member with the USTC. His research interests include multi-modal analysis, human-computer interaction (HCI) and affect computing. He has published 30+ papers in refereed conferences and journals, including ACM MM, KDD, ICME, IEEE TAFFC, ACM TOMM, etc.



Kun Li is currently a Postdoctoral Fellow at United Arab Emirates University. He received his Ph.D. degree from Hefei University of Technology in 2023. His research interests include multimedia content analysis, computer vision, and video understanding. He regularly serves as a PC member for top-tier conferences in multimedia and artificial intelligence, like ACM Multimedia, IJCAI, AAAI, CVPR, ICCV, and ECCV.

Zhi Liu (S'11-M'14-SM'19) received the Ph.D. degree in informatics in National Institute of Informatics. He is currently an Associate Professor at The University of Electro-Communications. His research interest includes video network transmission and mobile edge computing. He is now an editorial board member of Springer wireless networks and IEEE Open Journal of the Computer Society. He is a senior member of IEEE.



Meng Li (Senior Member, IEEE) is a Professor and Personnel Secretary at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), China. He was a Post-Doc Researcher at the Department of Mathematics and HIT Center, University of Padua, Italy, where he is with the Security and PRIvacy Through Zeal (SPRITZ) research group led by Prof. Mauro Conti (IEEE Fellow). He obtained his Ph.D. in Computer Science and Technology from the School of Computer Science and Technology, Beijing Institute of Technology (BIT), China, in 2019. He was sponsored by ERCIM 'Alain Bensoussan' Fellowship Programme (from 2020.10 to 2021.3) to conduct Post-Doc research supervised by Prof. Fabio Martinelli at CNR, Italy. He was sponsored by China Scholarship Council (CSC) as a Joint Ph.D. student (from 2017.9 to 2018.8) supervised by Prof. Xiaodong Lin (IEEE Fellow) in the Broadband Communications Research (BBCR) Lab at University of Waterloo and Wilfrid Laurier University, Canada. He is supported by CSC as a Visiting Scholar (from 2025.3 to 2025.6) collaborating with Prof. Mauro Conti (IEEE Fellow) at the HIT Center, University of Padua, Italy. His research interests include security, privacy, applied cryptography, blockchain, TEE, and Internet of Vehicles. In this area, he has published 115 papers in topmost journals and conferences, including TIFS, TDSC, ToN, TMC, TKDE, TODS, TPDS, TSC, COMST, IEEE S&P, USENIX Security, ACM MobiCom, and ISSTA. He is a Senior Member of IEEE, CIE, CIC, and CCF. He is an Associate Editor for TIFS, TDSC, and TNSM. He has served as a TPC member for conferences, including ICDCS, Inscrypt, ICICS, and TrustCom. He is the recipient of 2024 IEEE HITC Award for Excellence (Early Career Researcher) and 2025 IEEE TCSVC Rising Star Award.

Ziyu Jia (Member, IEEE) is an Assistant Professor at the Institute of Automation, Chinese Academy of Sciences. His research focuses on time-series analysis methods and their applications in health and medicine, including multimodal affective computing, sleep stage classification, and brain-computer interfaces. He has published over 50 peer-reviewed papers in venues such as IEEE TAFFC, IEEE TMM, IEEE TNSRE, KDD, and ICLR. Dr. Jia currently serves as an Associate Editor or Editorial Board Member for prestigious journals including IEEE TAFFC and Information Fusion, and he is an Area Chair for major AI and machine learning conferences such as IJCAI and IJCNN. In addition to his academic contributions, Dr. Jia has extensive industry experience, having successfully led multiple R&D projects and secured several patents. He has received numerous honors, including the MSRA StarTrack Award, and the CIE Young Talent Award.

Dan Guo (Senior Member, IEEE) received the Ph.D. degree in system analysis and integration from Huazhong University of Science and Technology, Wuhan, China, in 2010. She is currently a Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China. Her research interests include computer vision and intelligent multimedia content analysis. She is an Associate Editor of the IEEE TMM.

Meng Wang (Fellow, IEEE) received the B.E. and Ph.D. degrees from the Special Class for the Gifted Young, Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Hefei University of Technology, Hefei. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He has authored or coauthored more than 200 book chapters and journal articles and conference papers in these areas. Dr. Wang was a recipient of the ACM SIGMM Rising Star Award in 2014. He is an Associate Editor of the IEEE TPAMI, the IEEE TKDE, the IEEE TCSVT, the IEEE TMM, and the IEEE TNNLS.