AVI Challenge 2026: Assessing True Personality Traits and Cognitive Ability from Asynchronous Video Interviews (AVIs)

Abstract

Asynchronous Video Interviews (AVIs) allow candidates to record responses to predefined questions using digital devices, offering flexibility and enabling remote assessments. Evaluating personality traits through AVIs provides organizations with valuable insights into candidates' profiles and helps predicting future job performance. Building on this, we successfully organized the *AVI Challenge 2025* at ACM Multimedia 2025, which established a high-fidelity benchmark by grounding assessments in Trait Activation Theory and employing Behaviorally Anchored Rating Scales (BARS). However, the AVI Challenge 2025 primarily focused on observer-reported personality, which captures "ascribed" traits susceptible to human perceptual bias and candidates' impression management. In addition, it bypass cognitive ability, a key predictor of job performance. To address these gaps, we introduce the "*AVI Challenge 2026*", featuring a novel dataset of mock AVIs (3,876 videos from 646 subjects). This iteration introduces two major shifts: (a) transitioning from social perception to "true" personality by utilizing validated self-reported inventories as ground truth, and (b) inaugurating a new track for cognitive ability estimation through semantic and paralinguistic analysis. By aligning multimedia analytics with the "gold standards" of psychology, this challenge aims to decode authentic professional potential. Additionally, ongoing data collection from Dutch and Mandarin speakers will make the dataset cross-cultural and multilingual, broadening its utility for global recruitment applications.

ACM Reference Format:

. 2026. AVI Challenge 2026: Assessing True Personality Traits and Cognitive Ability from Asynchronous Video Interviews (AVIs). In *Proceedings of ACM International Conference on Multimedia (ACM MM '26).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM MM '26, Rio de Janeiro, Brazil

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Interviews are the most common tool to select new employees [1]. The rapid evolution of Artificial Intelligence (AI) technology has fundamentally transformed traditional recruitment, a shift that has been further accelerated by the impact of the COVID-19 pandemic [2, 3]. A pivotal development in this landscape is the rise of Asynchronous Video Interviews (AVIs), one-way online assessments where candidates record video responses to standardized interview questions via personal devices[2]. When integrated with multimedia AI models, AVIs facilitate the automated assessment of personality traits and cognitive abilities, serving as critical predictors of future job performance[4, 5]. Consequently, AVIs offer a cost-effective alternative for candidate pre-screening[6], while enhancing the psychometric integrity through highly structured formats[2]. The industrial adoption of this technology has surged; for instance, major providers reported an increase in AVI volume from 26 million to over 40 million sessions between 2022 and 2024[7, 8].

Since 2011, a series of grand challenges, including *Vlog* (*AAAI*'11) [9], *WCPRST* (*ACM MM*'14) [10], *ChaLearn First Impressions V2* (*CVPR*'17) [11], have advanced state-of-the-art methods in multimedia personality and cognitive assessment. However, these pioneering efforts were often constrained by validity and methodological considerations. For instance, the influential *ChaLearn First Impressions* series [12, 13] relied on brief YouTube clips and crowdsourced annotations, which often failed to capture the nuanced psychological breadth of personality constructs as defined by Trait Activation Theory (TAT) [14].

To tackle this, we organized the first AVI Challenge[1] at ACM Multimedia 2025 [15], which established a rigorous framework for personality and interview performance assessment by replacing simplistic labeling with Behaviorally Anchored Rating Scales (BARS) and grounding the tasks in TAT [14]. This shift toward ecologically valid recruitment scenarios resonated strongly with the multimedia community: the AVI Challenge 2025 attracted 63 teams from both academia and industry, with 20 teams successfully achieving official rankings[2,3]. By bridging the gap between multimedia

[1] https://avichallenge.github.io/

[2] https://codalab.lisn.upsaclay.fr/competitions/23100

[3] https://codalab.lisn.upsaclay.fr/competitions/23101

research and real-world selection practices, the AVI Challenge 2025 [15] provided a high-fidelity benchmark reflecting the complexities of real staffing recruitment.

Despite these advancements, there remains significant scope to enhance the psychometric rigor and comprehensiveness of the assessment framework. Specifically, the previous iteration's reliance on observer-reported personality essentially measured "ascribed" traits. Such annotations are inherently susceptible to human perceptual biases and candidates' tactical impression management, creating a divergence from self-reported (true) personality traits and compromises the models' construct validity [16]. Furthermore, the AVI Challenge 2025 bypassed cognitive ability, arguably the most robust and consistent predictor of job performance across diverse occupational domains [17]. Without accounting for these internal psychological constructs, multimedia models risk merely mimicking superficial social perceptions rather than decoding a candidate's authentic professional potential.

To overcome these limitations, we propose the "AVI Challenge 2026". This iteration introduces two major shifts:

- **From Perception to Reality**: We shift the ground truth for personality assessment from observer ratings to validated self-reported inventories as "true" personality traits, tasking participants with developing models capable of distinguishing enduring traits from situational acting.
- **Cognitive Ability Estimation**: We inaugurate a new track dedicated to inferring problem solving capacities. This requires extracting high-level semantic and paralinguistic features, such as linguistic complexity, logical coherence, and response latency.

By integrating these dimensions, the AVI Challenge 2026 provides a benchmark that aligns state-of-the-art multimedia analytics with the "gold standards" of Industrial and Organizational (I-O) psychology, fostering more equitable and predictive AI-driven hiring systems. Notably, while the current dataset features English responses, we are actively expanding to include Dutch and Mandarin speakers, paving the way for future cross-cultural and multilingual challenges.

2 Dataset

The dataset is available on Google Drive[4]. Below is the details of how the dataset is constructed.

2.1 Subjects

The dataset consisted of video interview data from 646 subjects. Subjects ($n = 793$) were recruited through the online platform Prolific [18]. After excluding subjects (a) with incomplete responses ($n = 40$), (b) who did not consent for their data to be shared ($n = 10$), (c) who did not pass the attention checks ($n = 7$), (d) whose variation in personality (HEXACO [19]) items was either too large or too small [20] ($n = 6$), (e) who self-reported that they did not take the study seriously ($n = 12$), (f) whose videos contained corrupted audio ($n = 51$), (g) and who were flagged by personality raters as non-compliant ($n = 21$), the final sample size consisted of $n = 646$ subjects.

Subjects were evenly distributed among men and women (309 men, 309 women, 26 non-binary) and were mainly White ($n = 467$), Black or African American ($n = 73$), Hispanic or Latino ($n = 46$), Asian ($n = 30$), 'other' ethnicities ($n = 25$), or did not disclose ethnicity ($n = 5$). The average age of subjects was 36.69 ($SD = 11.92$), and they had on average 15.96 years of working experience ($SD = 11.36$). Among subjects, 6 had not finished high-school, 76 were high-school graduates, 220 were college graduates, 233 had a Bachelor's degree, 89 had a Master's degree, 16 had a doctorate, and 6 did not disclose education level.

2.2 Procedure

Subjects applied to a fictitious *management traineeship position*. Part of the application procedure was to complete an AVI using a platform we developed for the purpose of the study. During the AVI, subjects responded to six interview questions. Two of them were ***generic questions***, frequently asked in selection interviews. The other four questions were related to the personality traits of **Honesty-Humility, Extraversion, Agreeableness, and Conscientiousness**, as described by the HEXACO model of personality [21] (i.e, the ***personality questions***). Subjects were instructed to reply within 1-2 minutes to the interview questions.

2.3 AVI question development

The interview followed a structured format, since previous literature suggests that structured (vs. non- or semi-structured) interviews have stronger reliability and validity [22, 23]. Subjects always started with the generic questions and proceeded to the personality questions. Table 1 shows the content, order and type of the six questions and their corresponding personality traits.

2.3.1 Generic questions. For the development of generic interview questions, we created an initial pool of 86 job interview questions taken from previous literature [24–26] and a list of frequently asked interview questions provided by a Dutch consultancy company. As a first step, we screened those questions for eligibility. Questions were included if they were (a) open-ended, (b) conveyed personality information to some extent, and (c) could apply to multiple jobs. Questions were excluded if they described specific behaviors, specific jobs, or knowledge, values, and motives. This procedure ended up in retaining 61 questions.

To assess those 61 questions, we asked 17 professional recruiters from a Dutch consultancy company to assess how

[4] https://drive.google.com/drive/folders/1KzUwsBhDYjR_zFHiUOQlamVFqYoKF6pT?usp=sharing

Table 1. The content, order and type of the interview questions and their corresponding personality traits.

Order	Interview question	Question Type	Personality Trait
1	What would you consider among your greatest strengths and weaknesses as an employee?	Generic	\
2	How would your best friend describe you?	Generic	\
3	Think of situations when you made professional decisions that could affect your status or how much money you make. How do you usually behave in such situations? Why do you think that is?	Personality	Honesty-Humility
4	Think of situations when you joined a new team of people. How do you usually behave when you enter a new team? Why do you think that is?	Personality	Extraversion
5	Think of situations when someone annoyed you. How do you usually react in such situations? Why do you think that is?	Personality	Agreeableness
6	Think of situations when your work or workspace were not very organized. How typical is that of you? Why do you think that is?	Personality	Conscientiousness

frequently they use each of those questions in practice. Responses were given on a 3-point scale. The inter-rater agreement between the recruiters was *ICC(2,17) = 0.88*. We then asked four personality experts to assess each interview question on a 7-point scale using three criteria. Namely, whether the questions applied to limited or multiple jobs, whether they activated one or more personality traits, as well as provide a general assessment (*ICC(2,4) = 0.66*). Then, we calculated the average score per criterion (professional recruiters, personality experts) and excluded all questions that scored below the average (per criterion). This process returned 16 questions which (a) were frequently used by practitioners, (b) were not specific to a particular job, and (c) activated more than one personality traits. Of those 16 questions, we slightly edited and selected two questions that received the highest ratings from recruiters and personality experts.

2.3.2 Personality questions. For the development of personality interview questions, we created an initial pool of 25 past behavior interview questions for the personality traits of Honesty-Humility ($n = 6$), Extraversion ($n = 8$), Agreeableness ($n = 5$), and Conscientiousness ($n = 6$). The questions were developed to target the core facets of each personality trait. Questions were developed in a past-behavior formal (e.g., "Think of situations when…") since this type of format are more suited to elicit personality-relevant information according to previous research [27]. Four personality experts independently selected one question per personality trait and later discussed any disagreements between them until a consensus was reached, retaining one question per personality trait. After some further editing, we ended up with four personality-related questions (Table 1).

2.4 Annotations

2.4.1 "True" personality traits. Self-reported personality traits were collected directly from the participants using validated inventories. Prior to the interview, each subject completed a comprehensive self-assessment based on the HEXACO model, following a standardized protocol similar to the one described by [28]. Responses were provided using a Behavioral Anchored Response Scale (BARS) [29], which was adapted for self-perception. The BARS contained four items per personality domain (one item per facet), and participants recorded their responses on a 5-point scale (1 = Very low; 5 = Very high), with the precision of one decimal point. Personality domain scores were calculated by averaging the four constituent facet scores.

2.4.2 Cognitive ability. Cognitive ability was assessed through a multi-dimensional psychometric test administered to all candidates. The assessment comprised 28 items categorized into four distinct tasks:

- ***Verbal reasoning (16 items)***: These items evaluate the ability to comprehend, analyze, and manipulate complex verbal and numerical information. For instance, questions such as "*What number is one-fifth of one-fourth of one-ninth of 900?*" assess deductive reasoning and the capacity to process sequential logical constraints.
- ***Letter-number series (4 items)***: This task focuses on inductive reasoning and pattern recognition. Candidates are required to identify the underlying rule in an alphanumeric sequence, such as "K, N, P, S, U, ...", reflecting their ability to discern abstract relationships within structured data.

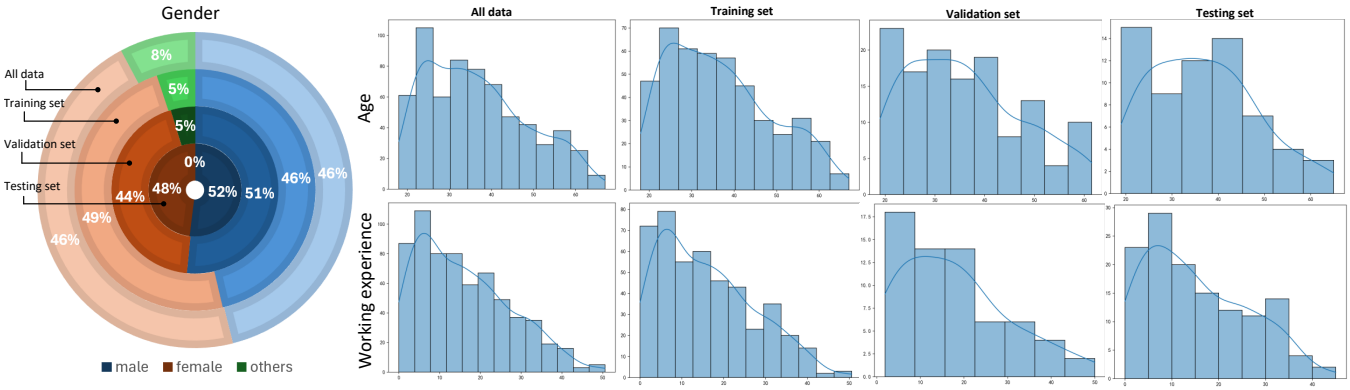

Figure 1. The gender, age, working experience distributions of the training, validation and testing set

- ***Matrix reasoning (4 items)***: Utilizing non-verbal, visuospatial stimuli, this section assesses fluid intelligence. Candidates must complete a logical pattern (e.g., "*Select the best answer to complete the figure below*"), measuring their ability to solve novel problems without relying on prior linguistic or cultural knowledge.
- ***Three-dimensional rotation (4 items)***: This task specifically targets visuospatial processing and mental manipulation. By asking candidates to identify the correct rotation of a labeled cube (e.g., "*Select the choice that could represent a rotation of the cube X*"), the test evaluates the ability to form and transform mental representations of complex objects.

Following the assessment, professional psychologists conducted a comprehensive evaluation of the candidates' responses. To ensure practical utility for recruitment modeling, scores were synthesized and categorized into a three-tier ordinal scale: *Low*, *Medium*, and *High* cognitive ability. This expert-led classification ensures that the ground truth reflects not just raw accuracy, but a holistic psychometric interpretation of the candidate's intellectual potential.

3 Tracks

AVI Challenge 2026 consists of two tracks: the personality (self-reported) assessment track and the cognitive ability assessment track. The former focuses on evaluating subjects' self-reported personality traits based on their responses to corresponding personality questions. The latter concentrates on using multimodal information from subjects' responses to all questions (i.e., including both generic and personality questions) to assess their cognitive ability. The detailed explanation of these two tracks are stated below.

3.1 Track 1: "True" personality assessment

In this track, participants will develop models and algorithms to assess the self-reported personality traits based on subjects' responses to the corresponding personality questions. For example, question 3 is the question to activate the trait of Honesty-Humility. Thus, participants will use the videos of subjects answering question 3 to assess the Honesty-Humility of these subjects. The task of this track is a single-input-single-label regression task. The purpose of the track is to encourage researchers in the multimedia community to advance methodologies for accurately predicting personality traits from standardized psychological data.

3.2 Track 2: Cognitive ability assessment

In this track, participants are tasked with developing models and algorithms to infer a candidate's cognitive ability from their asynchronous video interview responses. This track is formulated as an ordinal classification task, where models must categorize each candidate into one of three levels: Low, Medium, or High. This track serves as a high-fidelity simulation of cognitive screening in modern recruitment, aiming to encourage the multimedia community to move beyond surface-level behavioral analysis. Unlike simple performance metrics, cognitive ability assessment requires the extraction of deep latent features, such as linguistic complexity, logical structure in verbal responses, and paralinguistic indicators of mental processing speed. Such advancements are expected to enhance the predictive validity of AI hiring systems, ensuring that automated recruitment processes are grounded in the most critical predictors of long-term job success.

4 Challenge guidelines

4.1 Dataset split

The dataset used for the challenge is split into training (70%, $n = 452$), validation (10%, $n = 64$), and testing (20%, $n = 130$) sets. The training and validation sets will be made available to participants for algorithm development. The dataset is divided at the subject level, ensuring that videos from a single subject are assigned exclusively to one of the training, validation, or testing sets. Although the input videos differ between the two tracks (i.e., track 1 uses only videos for answering personality questions, while track 2 uses videos for answering all questions), the split remains consistent across

both tracks of our challenge. In splitting the dataset, we consider the distribution of gender, age, and working experience of the subjects. Specifically, we employ joint sampling to ensure that these three sets maintain similar distributions of these demographic and experiential variables. The gender, age, working experience distribution of the training, validation and testing sets are shown in Fig 1.

4.2 Evaluation metrics

For the AVI Challenge 2026, we employ two distinct metrics tailored to the specific nature of each track:

- **Track 1 ("True" personality traits assessment)**: The primary evaluation metric is **Mean Squared Error (MSE)**. As Track 1 is formulated as a regression task, MSE provides a rigorous measure of model precision by calculating the average squared difference between the predicted and ground-truth values for each personality trait. This metric is particularly effective for penalizing larger deviations, ensuring that the models achieve high granularity in capturing continuous psychological constructs.
- **Track 2 (Cognitive ability assessment)**: The performance will be evaluated using **Accuracy (Acc)**. Given that the cognitive ability task is defined as a three-tier classification (Low, Medium, and High), Accuracy serves as a straightforward and intuitive measure of the model's ability to correctly categorize candidates according to the psychologists' expert ratings.

4.3 Schedule

- Data, baseline paper & code available 20 April, 2026
- Results submission start 20 June, 2026
- Results submission deadline 1 July, 2026
- Deadline for paper submission 12 July, 2026
- Paper acceptance notification 5 August, 2026
- Deadline for camera-ready papers 19 August, 2026

5 Challenge organizers

5.1 Organizers

Tianyi Zhang is an associate professor at the School of Biological Science and Medical Engineering, Southeast University. Before that, he worked as a postdoc researcher at Vrije Universiteit Amsterdam. He got his PhD degree in Delft University of Technology. He was also associated with the Distributed & Interactive Systems (DIS) group at the national research institute for mathematics and computer science in the Netherlands (CWI). His research interests lie in human-computer interaction, affective computing, and personality recognition.

Reinout E. de Vries is Full Professor in Organizational Psychology with a chair in 'Personality at Work' at the Vrije Universiteit Amsterdam, the Netherlands. His main areas of interest are the theoretical background, structure, measurement, and effects of personality, leadership, communication styles, and situations. His current work focuses on the Situation-Trait-Outcome Activation (STOA) model and the automatic assessment of personality using algorithms. Reinout is on the editorial boards of The Leadership Quarterly, the European Journal of Personality, the International Journal of Selection and Assessment, and various other journals, and he has published more than 100 articles on personality and organizational topics in a wide range of scientific journals.

Wenming Zheng is currently a Professor with the Key Laboratory of Child Development and Learning Science, Ministry of Education, Southeast University. His research interests include affective computing, pattern recognition, machine learning, and computer vision. He has been elected as Institution of Engineering and Technology (IET) Fellow since 2022. He is a former associate editor of the IEEE Transactions on Affective Computing, the associate editor of the IEEE Transactions on Cognitive and Developmental Systems, and the editorial board member of The Visual Computer.

Janneke Oostrom is a Professor of Work & Organizational Psychology at Tilburg University's department of Social Psychology. Her research focuses on understanding and improving psychological assessments, with the goal to make them more predictive of future work behaviors and reducing discrimination against marginalized groups. She received her Master and Ph.D. in Work and Organizational Psychology from the Erasmus University Rotterdam.

Djurre Holtrop is an assistant professor at Tilburg University's department of Social Psychology. He has worked in consulting for psychometric assessments, leading large scale online assessment projects. He completed his PhD at the VU Amsterdam studying the refinement of personality and vocational interest questionnaires. Subsequently, he worked for the University of Western Australia and Curtin University to study the recruitment, motivation, and retention of volunteers. Currently, his research focusses on personnel recruitment and selection and volunteer attraction and engagement.

Yuan Zong received the Ph.D. degree in biomedical engineering from the Southeast University, Nanjing, China, in 2018. He is currently an associate professor with the Key Laboratory of Child Development and Learning Science (Ministry of Education), School of Biological Science and Medical Engineering, Southeast University. He has authored or coauthored more than 30 articles in mainstream journals and conferences, such as the IEEE T-IP, T-CYB, T-AFFC, AAAI, IJCAI, and ACM MM. His research interests include affective computing, pattern recognition, and computer vision.

5.2 Data chair

Antonis Koutsoumpis works as a postdoc researcher at department of social psychology, at Tilburg University the Netherlands. He received his PhD candidate at the Organizational Section of the Experimental and Applied Psychology Department, at the Vrije Universiteit Amsterdam, the Netherlands. He has a background in psychology and his research interests include automatic personality assessment from asynchronous video interviews as well as the verbal and non-verbal behaviors that individuals exhibit depending on their personality traits.

Tianhua Qi is currently working toward the PhD degree with the School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China, and also with the Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, China. He is the general co-chair of The 12th ISCA-SAC Doctoral Consortium. His research interests include affective computing, deep learning, and speech signal processing.

6 Commitment

If our proposal is lucky enough to be accepted, we commit to publish and maintain a website related to the Grand Challenge containing the information, datasets, tasks for the Grand Challenge at least the next 3 years. For any question regarding this challenge, please connect with Tianyi Zhang (t.zhang@seu.edu.cn).

References

[1] Stephen D. Risavy, Patricia A. Fisher, Chet Robie, and Cornelius J. König. Selection tool use: A focus on personality testing in canada, the united states, and germany. *Personnel Assessment and Decisions*, 5(1):4, 2019.

[2] Eden-Raye Lukacik, Joshua S Bourdage, and Nicolas Roulin. Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, 32(1):100789, 2022.

[3] Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, pages 1–27, 2019.

[4] Louis Hickman, Nigel Bosch, Vincent Ng, Rachel Saef, Louis Tay, and Sang Eun Woo. Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107(8):1323, 2022.

[5] Jon Liff, Nicholas Mondragon, Caitlin Gardner, Christopher J Hartwell, and Angela Bradshaw. Psychometric properties of automated video interview competency assessments. *Journal of Applied Psychology*, 109(6):921–948, 2024.

[6] Edwin N Torres and Cynthia Mejia. Asynchronous video interviews in the hospitality industry: Considerations for virtual employee selection. *International Journal of Hospitality Management*, 61:4–13, 2017.

[7] HireVue. Explainability statement (white paper)., 2022.

[8] HireVue. Explainability statement (white paper)., 2024.

[9] Joan-Isaac Biel, Oya Aran, and Daniel Gatica-Perez. You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 446–449, 2011.

[10] Joan-Isaac Biel and Daniel Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2012.

[11] Julio CS Jacques Junior, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel AJ Van Gerven, Rob Van Lier, et al. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing*, 13(1):75–95, 2019.

[12] Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 400–418. Springer, 2016.

[13] Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Junior Jacques, Meysam Madadi, Xavier Baró, and Rob Van Lier. Design of an explainable machine learning challenge for video interviews. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3688–3695. IEEE, 2017.

[14] Robert P Tett, Margaret J Toich, and S Burak Ozkum. Trait activation theory: A review of the literature and applications to five lines of personality dynamics research. *Annual Review of Organizational Psychology and Organizational Behavior*, 8:199–233, 2021.

[15] Tianyi Zhang, Tianhua Qi, Antonis Koutsoumpis, Yuan Zong, Wenming Zheng, Janneke K. Oostrom, Djurre Holtrop, Zhaojie Luo, and Reinout E. de Vries. Assessing personality traits and interview performance from asynchronous video interviews. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 13895–13900, New York, NY, USA, 2025. Association for Computing Machinery.

[16] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[17] Frank L Schmidt and John E Hunter. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, 124(2):262, 1998.

[18] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of experimental social psychology*, 70:153–163, 2017.

[19] Michael C Ashton and Kibeom Lee. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and social psychology review*, 11(2):150–166, 2007.

[20] Ard J Barends and Reinout E De Vries. Noncompliant responding: Comparing exclusion criteria in mturk personality research to improve data quality. *Personality and individual differences*, 143:84–89, 2019.

[21] Kibeom Lee and Michael C Ashton. Psychometric properties of the hexaco-100. *Assessment*, 25(5):543–556, 2018.

[22] Allen I Huffcutt, Chad H Van Iddekinge, and Philip L Roth. Understanding applicant behavior in employment interviews: A theoretical model of interviewee performance. *Human Resource Management Review*, 21(4):353–367, 2011.

[23] Paul R Sackett, Charlene Zhang, Christopher M Berry, and Filip Lievens. Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, 2021.

[24] Mohammed Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 697–706, 2013.

[25] Iftekhar Naim, Md Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, 9(2):191–204, 2016.

[26] Hung-Yue Suen, Mavis Yi-Ching Chen, and Shih-Hao Lu. Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior*, 98:93–101, 2019.

[27] Julia Levashina, Christopher J Hartwell, Frederick P Morgeson, and Michael A Campion. The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, 67(1):241–293, 2014.

[28] Antonis Koutsoumpis, Sina Ghassemi, Janneke K Oostrom, Djurre Holtrop, Ward van Breda, Tianyi Zhang, and Reinout E de Vries. Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning. *Computers in Human Behavior*, page 108128, 2024.

[29] George S Rotter and Vera Tinkleman. Anchor effects in the development of behavior rating scales. *Educational and Psychological Measurement*, 30(2):311–318, 1970.

A Baseline method

To establish a benchmark for the challenge, we provide a multimodal baseline framework based on established deep learning architectures[5]. This framework aims to assist challenge participants in gaining a clear understanding of the expected performance standards and provide a reference point for evaluating their own models. As illustrated in Fig. 2, the system processes single-speaker video clips to simultaneously address Track 1 ("True" Personality Assessment) and Track 2 (Cognitive Ability Assessment). The pipeline of the system can be divided into three modalities.

- **Visual modality**: Captures non-verbal cues such as facial micro-expressions, gestures, and body language, which are pivotal for personality expression and reflecting cognitive effort or stress.
- **Audio modality**: Analyzes prosodic features, pitch, and vocal tone, which often correlate with emotional stability in personality and processing speed or fluency in cognitive tasks.
- **Verbal modality**: Analyzes the semantic content and linguistic structure of speech, reflecting both the lexical tradition of personality models (e.g., HEXACO) and the logical coherence indicative of cognitive ability.

The workflow consists of three stages: Feature Extraction, where modality-specific encoders compute local descriptors; Temporal Aggregation and Fusion, which compresses the temporal sequence into a unified multimodal vector; and Regression, where specialized heads predict the specific scores for both tracks.

A.1 Feature extraction

A.1.1 Visual modality. Visual behavior, particularly facial dynamics and posture, serves as a window into both an individual's trait-based tendencies and their real-time cognitive state. We utilize CLIP[6] (Contrastive Language-Image Pre-Training) to extract high-level semantic visual features. By leveraging its vision-transformer backbone, the model captures subtle cues in appearance and "visual affect" that are essential for predicting personality traits and detecting signs of cognitive load or engagement.

[5]Code: https://github.com/AVIChallenge/AVI2026_baseline
[6]https://huggingface.co/openai/clip-vit-base-patch32

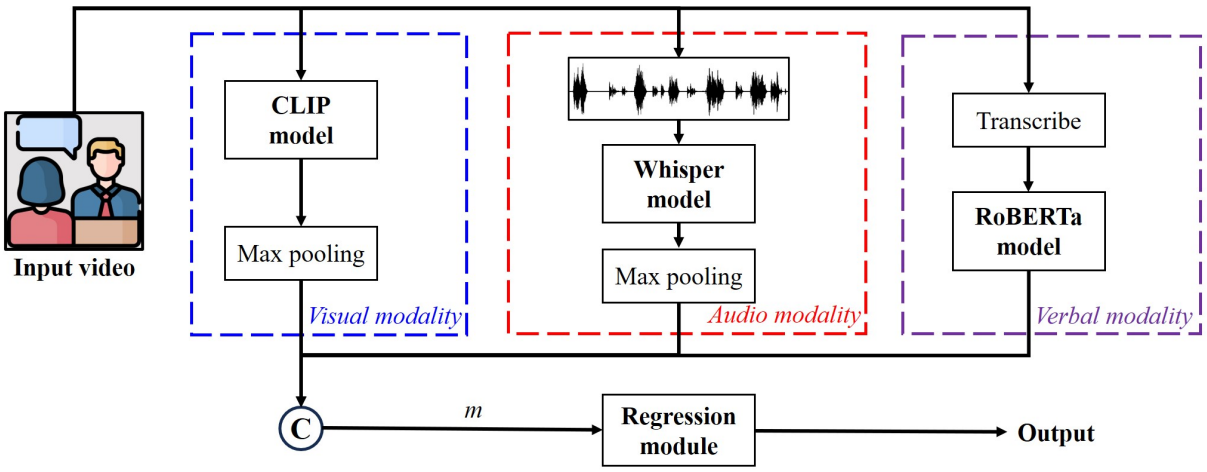

Figure 2. Pipeline overview: visual (blue), audio (red), verbal (purple) modalities, and 'C' stands for concatenation. During training, only the regression module requires personality annotations.

Table 2. Baseline performance for "True" Personality Assessment (*Track 1*) and Cognitive Ability Assessment (*Track 2*). Results for Track 1 are reported in Mean Squared Error (MSE ↓), and Track 2 in Accuracy (Acc ↑).

Method	Track 1: "True" Personality (MSE)					Track 2: Cognitive Ability
	H	E	A	C	*Avg.*	*Score (Acc)*
Proposed Baseline	0.2591	0.4347	0.3873	0.2524	0.3334	0.4062

A.1.2 Audio modality. Acoustic characteristics provide vital clues regarding an individual's internal state. For Track 1, voice quality helps distinguish traits like *Extraversion* or *Emotionality*; for Track 2, it captures speech latency and phonation patterns related to cognitive processing. We employ whisper[7], a robust speech-processing model, to extract deep acoustic features. Unlike traditional handcrafted features, these embeddings capture a rich representation of the audio signal suited for diverse downstream regression tasks.

A.1.3 Verbal modality. The linguistic choices made by a speaker are perhaps the most direct indicators of both personality (through word choice and sentiment) and cognitive ability (through vocabulary diversity and syntactic complexity). We use RoBERTa[8] to process speech transcripts. This transformer-based model generates contextualized representations that capture the subtle nuances of dialogue, enabling the system to map language patterns to HEXACO personality dimensions and cognitive performance scores.

A.2 Modality fusion

To handle the asynchronous nature of the modalities, we apply temporal global average pooling to each stream. The resulting vectors are concatenated into a unified multimodal representation m in Fig. 2.

A.3 Regression

This joint embedding is then fed into the Regression Module. To ensure robustness and provide a measure of predictive uncertainty, we employ a Deep Ensemble approach. The module consists of 32 independent Multi-Layer Perceptrons (MLPs), each featuring hidden layers of 32 and 8 units. Separate regression heads are trained for Track 1 (predicting HEXACO traits) and Track 2 (predicting cognitive scores).

A.4 Results

Table 2 summarizes the performance of the baseline system across both tracks. For Track 1, we report the MSE for four key HEXACO personality traits. For Track 2, we provide the baseline accuracy for Cognitive Ability Assessment, establishing the difficulty level of the task.

[7] https://github.com/openai/whisper
[8] https://huggingface.co/FacebookAI/roberta-base