

Cross-Protocol Personality Assessment with LLMs: Aligning Behavioral Evidence with Human Ratings

Wei Shan, Tianyi Zhang*, *Senior Member, IEEE*, Yuan Zong, *Member, IEEE*, and Wenming Zheng, *Senior Member, IEEE*,

Abstract—Personality assessment from asynchronous video interviews (AVIs) plays a crucial role in recruitment by enabling standardized and behaviorally grounded evaluation for job applicants. However, most of the previous works are confined to a single interview protocol. Independently collected protocols differ in question sets and question-trait mapping. Thus, the human-rated evidence from one protocol cannot be used directly under another. To address this limitation, we propose the Cross-protocol Evidence-grounded Personality Assessment (CEPA) framework, which aligns behavioral representations across interview protocols. CEPA addresses protocol heterogeneity by projecting source behavioral records into the target protocol's construct-oriented representation space. It retrieves behaviorally relevant and personality-complementary human-rated exemplars to ground an ensemble of LLM raters, which enables behavior-to-personality associations without target-specific supervision. In cross-protocol experiments, CEPA reduces average mean squared error (MSE) by 43.8% (Forward) and 17.0% (Reverse) relative to the strongest baseline. Our empirical experiments show that human-rated evidence from one protocol can support personality assessment under another without target-specific supervision or retraining.

Index Terms—Personality assessment; Large language models; Asynchronous video interviews

I. INTRODUCTION

Asynchronous video interviews (AVIs) are now a widely used tool in high-volume personnel screening. In AVIs, candidates record answers to predefined and structural questions for interviewer screening [1], [2]. The structured format scales to large candidate pools at low cost. Many companies use AVIs to assess personality traits in job recruitment as personality traits are considered as important indicators of how well someone might fit the job and workplace culture [3], [4]. Thus, plenty of multimodal algorithms have been developed to assess personality traits from AVIs [5]–[7].

The uniqueness of AVI-based personality assessment is that behavior cues related to personality traits are activated by structured and protocol-specific AVI questions. According to Trait Activation Theory [8], situational cues (i.e., questions in AVIs) trigger the expression of relevant latent traits through observable behaviors. Unlike personality assessment from daily conversations or vlog [9]–[11] (i.e., first impression personality), the models for AVI-based personality assessment

need to capture the question-trait mapping and the correspondence between each question and the traits it activated.

Although previous works have developed plenty of methods to capture the question-trait mappings, they are typically restricted to a single AVI protocol. Independently collected protocols are heterogeneous in their questions and question-trait mappings even when they target the same traits. A question may target one trait or several, and multiple questions can target the same trait. Thus, the same observed behavior can carry different trait interpretations under different protocols. For example, a candidate's description of organizing a team project might be interpreted as evidence of Extraversion under one protocol and Conscientiousness under another. The cross-protocol problem can lead to the cross-protocol problem for AVI-based personality assessment: behavioral evidence gathered under one protocol cannot be used directly under another.

To address the cross-protocol problem, two kinds of methods are developed. Some of the methods use zero-shot LLM to assess personality from behavioral language without target-specific supervision. Its judgments agree with self-report ratings and are valid in AVI-based assessment [12]–[16]. However, this kind of method ignores the fact that the human judgments already collected under other protocols. Other methods use supervised fine-tuning, including parameter-efficient methods such as LoRA, adapts a model to a target protocol [17], [18]. This kind of method needs target-specific supervision and optimization for every new protocol. Neither approach therefore exploits evidence already collected under other protocols.

To address this limitation, we develop **CEPA**, Cross-protocol Evidence-grounded Personality Assessment, a retrieval-augmented framework for multimodal personality assessment across protocols. CEPA can reuse human-rated behavioral evidence from existing protocols to assess personality under new protocols without requiring labeled target data or model retraining. CEPA has two modules: the Representation Alignment module and the personality-space Max-Min Evidence Selection module. **Representation Alignment** maps multimodal behavioral records from different protocols into a shared representation that preserves each record's personality profile. **Personality-space Max-Min Evidence Selection (PMES)** then constructs a compact evidence pool that balances behavioral relevance with coverage of the personality space. Finally, multiple LLM raters reason over the retrieved records and their judgments are aggregated into a continuous personality score. Together, these components enable human-rated records from one

Wei Shan, Tianyi Zhang, Yuan Zong, and Wenming Zheng are with the Key Laboratory of Child Development and Learning Science (Ministry of Education), School of Biological Sciences and Medical Engineering, Southeast University, Nanjing, China.

Corresponding author: Tianyi Zhang; t.zhang@seu.edu.cn

protocol to provide empirical evidence for personality assessment under another protocol. In general, our work make below contributions to the multimedia community:

- 1) We propose **CEPA**, a protocol-agnostic framework that reuses human-rated records from one assessment protocol to score personality under another. It relies on previously collected human judgments at **assessment** time and requires no target-specific supervision or optimization. This avoids the evidence loss of zero-shot assessment and the per-protocol supervision that fine-tuning requires.
- 2) In bidirectional transfer between two independently collected AVI datasets, CEPA reduces average MSE by 43.8% (Forward) and 17.0% (Reverse) over the strongest zero-shot baseline. Ablations attribute most of the gain to retrieved human-rated evidence and Representation Alignment.
- 3) We show that the empirical relationship between observed behavior and human personality judgment transfers across protocols. By contrast, the source personality label does not transfer. Using retrieved records as evidence consistently outperforms using their attached personality profiles directly. A multi-rater reliability analysis separates prediction error from the stability of the scoring process.

II. RELATED WORK

A. Personality Assessment from Asynchronous Video Interviews

Early work predicted the Sixteen Personality Factor (16PF) dimensions from facial features [19]. Deep bimodal regression predicted apparent personality from short video sequences [9]. Later models combined channels through trait-specific fusion and joint classification-regression objectives [10], [20], [21]. Multimodal fusion now dominates because personality evidence is spread across verbal and non-verbal behavior [6], [22]–[24]. These fusion models are tied to the protocol they were trained on. Their behavior-to-trait mappings do not transfer to new question sets or trait definitions. Large language models instead score personality directly from behavioral language.

A first line of LLM-based work applies pre-trained models in a zero-shot setting. Peters and Matz [12] showed that GPT-3.5 and GPT-4 infer Big Five traits from Facebook updates. Ganesan et al. [14] found that zero-shot GPT-3 matched a supervised state-of-the-art model on broad Big Five classification once trait knowledge was added to the prompt, but fell to a most-frequent-class baseline on fine-grained classification. Zhang et al. [16] provide a systematic validity and fairness evaluation of LLM-based AVI assessment. These methods score a candidate from observed behavior alone. They never reuse human-rated records collected under other protocols. The cross-protocol problem is left unaddressed.

A second line adapts LLMs with target-specific supervision. Representative methods fine-tune the LLM [17] or train a predictor over its embeddings [25]. Zhang et al. [18] proposed PersonalityLLM, a coarse-to-fine LoRA framework that

regresses a continuous score within an empirical trait range, improving prediction by 16.6%/37.9% on two benchmarks. MoA-Personality [26] pools multiple LLM annotators with an aggregator to curb positivity bias, outperforming medium-sized open-source LLMs and approaching closed-source GPT-3.5 and GPT-4.

These methods learn a protocol-specific behavior-to-trait mapping from annotated data. Neither line reuses evidence from other protocols. The zero-shot line discards prior human judgments. The fine-tuning line needs target-specific supervision for each new protocol. Retrieval-augmented generation avoids per-protocol training by conditioning on retrieved evidence.

B. Retrieval-Augmented Generation and Evidence Transfer

Retrieval-augmented generation (RAG) combines a parametric language model with a non-parametric retrieval index. Karpukhin et al. [27] showed that neural dual-encoders can replace sparse retrievers, and Lewis et al. [28] built a general framework around a dense retriever and a parametric generator. Retrieval has since been integrated into language-model pre-training [29], [30], with the RAG design space now covered by dedicated surveys [31]. Effective evidence selection requires relevant items and a diverse retrieved set [32]–[35].

Personality assessment imposes a different evidence requirement. RAG in open-domain QA retrieves factual passages that contain an answer. Personality assessment instead needs records that tie observed behavior to a human judgment of the trait. Conventional RAG assumes a general or task-specific knowledge base. No such base links multimodal behavior to traits across different question-trait mappings. Independently collected datasets fill this role. Because they already pair behavioral observations with human personality judgments. These knowledgeable-other judgments have well-documented validity [36]. One dataset can serve as an evidence base for another.

This setting connects to cross-dataset transfer and domain adaptation. Classic methods learn domain-invariant features through gradient reversal [37] or RKHS mean matching [38]. Adaptation has also been applied to personality analysis across short-video platforms [39] and to RAG retrieval for open-domain QA [40]. These methods share one assumption: each transfers a model or a feature representation. Retrieval-based methods adjust a retrieval policy. They usually require target-domain data or parameter updates. They do not address whether the behavior-to-trait mappings embedded in human-rated records can themselves be transferred at assessment time. CEPA targets this gap directly. It retrieves human-rated records from one protocol as an evidence base and aggregates the judgments of multiple LLM raters. This enables behavior-to-trait mappings to transfer across protocols without target-specific supervision or parameter updates.

III. METHOD

A. Problem Setting and Framework Overview

We consider cross-protocol transfer of multimodal behavioral evidence for personality assessment. The independently

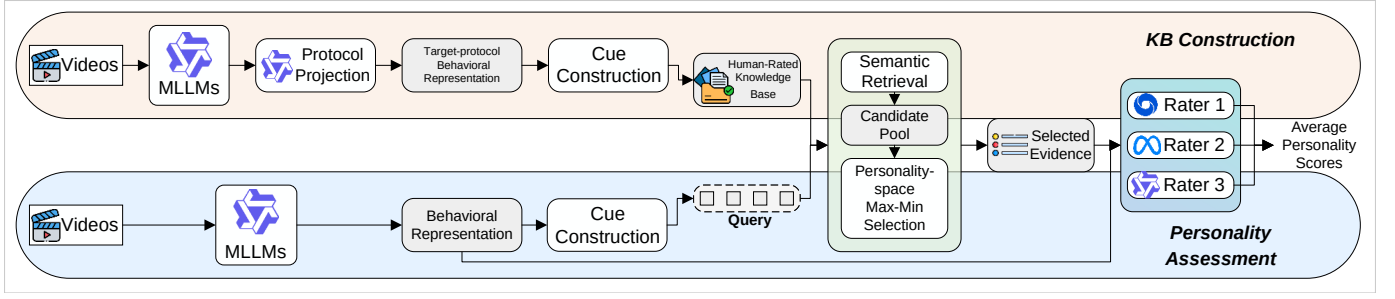


Fig. 1. Overview of the proposed **CEPA** framework. Multimodal interview videos are first converted into textual behavioral representations. Source evidence is aligned to the target protocol and encoded into a shared construct-oriented retrieval space. For a target subject, semantic retrieval forms a candidate pool, PMES selects complementary human-rated exemplars, and independent LLM raters produce continuous personality scores.

collected interview datasets may differ in their question-trait mappings. Let $\mathcal{D}_s$ and $\mathcal{D}_t$ denote the source and target datasets, respectively. The source provides human-rated behavioral exemplars for knowledge-base (KB) construction. The target is held out from personality-rating supervision and from retrieval-configuration selection. Candidate-pool size, evidence budget, and retrieval encoder are not tuned on the target. At assessment, only the target subject's behavioral observations are available as a query.

Let $\mathcal{F}_t$ denote the construct space induced by the target assessment protocol and used for evidence construction, selection, and scoring. $\mathcal{F}_t$ is defined by the target protocol, so its granularity may differ from the source dataset's native annotation granularity. During KB construction, each source profile is mapped into $\mathcal{F}_t$. Every admitted evidence card then carries a complete target-space profile.

We aim to transfer human-rated behavioral evidence from $\mathcal{D}_s$ to $\mathcal{D}_t$ and predict continuous personality scores without target-dataset personality supervision. Fig. 1 shows the proposed framework, which we call **CEPA** (Cross-protocol Evidence-grounded Personality Assessment). It has three stages: **Representation Alignment**, **Personality-space Max-Min Evidence Selection (PMES)**, and **Evidence-Grounded Continuous Scoring**. Representation Alignment conditions source behavior on the target protocol and converts it into a construct-oriented retrieval representation. PMES retrieves a relevant candidate region by semantic similarity and then selects complementary human-rated exemplars. Evidence-Grounded Continuous Scoring uses independent LLM raters to produce continuous personality scores conditioned on the selected evidence.

B. Multimodal Behavioral Preprocessing

Each interview video is first converted into a textual multimodal behavioral representation before evidence construction or assessment. The preprocessing stage obtains the interview transcription together with textual descriptions of audio and visual behaviors. These descriptions form a modality-serialized representation of the observable interview behavior and serve as input to the Protocol Projection and Cue Construction modules.

We denote the resulting behavioral representations of a source subject i and of the target subject by x_i and x_t , respectively. The preprocessing procedure is applied independently

to the source and target data and does not use personality ratings from the target dataset.

C. Representation Alignment

Representation Alignment transforms source behavioral evidence into a target-protocol-compatible retrieval representation and preserves the source human-rated profile. It has two sequential operations. *Protocol Projection* aligns source behavior with the target protocol. *Cue Construction* converts the projected behavior into a construct-oriented semantic representation.

1) *Protocol Projection*: Let P_s and P_t denote the source and target assessment protocols. For source behavioral evidence x_i with human-rated profile y_i^s , Protocol Projection constructs the target-protocol-compatible representation $\tilde{x}_{i \rightarrow t}$. It also maps the source profile into the target construct space as $y_{i \rightarrow t} = (s_f^i)_{f \in \mathcal{F}_t}$, where s_f^i is the human-rated score of subject i on construct f . The projection rewrites only the behavioral representation. The human-rated profile comes from source annotations through a predefined construct mapping and is not inferred from the projected behavior.

The target construct space is $\mathcal{F}_t = \{f_1, \dots, f_{|\mathcal{F}_t|}\}$, and $\mathcal{C}_f$ denotes the official definition and assessment description of construct f . The source-to-target profile mapping provides a human-rated score s_f^i for each $f \in \mathcal{F}_t$ required by the transfer setting. Source records without a complete mapped profile are excluded from the KB rather than having missing construct scores imputed. All evidence cards admitted to PMES carry a complete human-rated profile in the target construct space.

Projection is applied during KB construction. Projected records are admitted only if they pass structural-completeness and score-leakage checks. All $|\mathcal{F}_t|$ construct slots must be populated, and the projected representation must not contain the numeric score value. Records that fail either check are discarded rather than repaired.

2) *Cue Construction and Semantic Representation*: Each projected behavioral representation is converted into a construct-oriented textual cue:

$$c_i^f = G_{\text{cue}}(\tilde{x}_{i \rightarrow t}, \mathcal{C}_f), \quad (1)$$

where $G_{\text{cue}}(\cdot)$ denotes the cue-generation template and c_i^f is the resulting textual cue of subject i on construct f .

Each cue organizes observable evidence into five semantic slots: HIGH-POLE EVIDENCE, LOW-POLE EVIDENCE, TRANSCRIPT EVIDENCE, AUDIO EVIDENCE, and VIDEO EVIDENCE. The cue does not emit numerical personality scores or diagnostic trait labels. Interview questions are not reproduced, and slot labels are removed before embedding. Prompt specifications and validation constraints are provided in the Supplementary Material.

For KB construction, source cues are generated from projected source representations. Protocol Projection maps source multimodal observations into target-protocol answer fields, so the AUDIO EVIDENCE and VIDEO EVIDENCE slots are set to *none*. Target cues are generated directly from target transcript, audio, and video descriptions. Both sides use the same construct definitions, score-free output format, semantic encoder, and per-construct aggregation.

Let $E_{\text{cue}}(\cdot)$ denote the semantic embedding operation. The subject-level representations are

$$\begin{aligned} z_i &= \frac{1}{|\mathcal{F}_t|} \sum_{f \in \mathcal{F}_t} E_{\text{cue}}(G_{\text{cue}}(\tilde{x}_{i \rightarrow t}, \mathcal{C}_f)), \\ z_t &= \frac{1}{|\mathcal{F}_t|} \sum_{f \in \mathcal{F}_t} E_{\text{cue}}(G_{\text{cue}}(x_t, \mathcal{C}_f)). \end{aligned} \quad (2)$$

where z_i and z_t are the construct-mean pooled embeddings of the per-construct cues $\mathcal{C}_f^i$ in (1), for the source and target subjects, respectively.

We refer to each human-rated source instance admitted to the KB as a source *exemplar*. Each exemplar is stored as an *evidence card* $(z_i, x_i^{\text{display}}, y_{i \rightarrow t})$ in the knowledge base. Here z_i is the retrieval representation, x_i^{display} is the projected textual description shown to the raters (without slot labels), and $y_{i \rightarrow t}$ is the source human-rated profile in the target construct space. Representation Alignment is applied only to the source during KB construction. Target observations are directly converted into the same cue representation.

D. Personality-space Max-Min Evidence Selection

For a target subject, evidence selection balances behavioral relevance with complementary coverage of the human-rated personality space. The target construct space $\mathcal{F}_t$ induces the corresponding source-side evidence space after source-profile mapping. The scores in this space are fixed before target evaluation rather than inferred from target subjects.

Given the target representation z_t in (2), Stage 1 retrieves a candidate pool C_t of K_{pre} source evidence cards by semantic similarity. Here $\text{sim}(z_t, z_i)$ is the cosine similarity between the target representation and the retrieval representation z_i of card e_i . Stage 2 applies the max-min selection within this pool subject to an evidence budget $K \leq K_{\text{pre}}$. Candidates with scores for all $f \in \mathcal{F}_t$ form $C_t^{\text{comp}} \subseteq C_t$. For these candidates, the standardized personality-space distance is

$$d_\sigma(e_i, e_j) = \sqrt{\sum_{f \in \mathcal{F}_t} \left(\frac{s_f^i - s_f^j}{\sigma_f} \right)^2}, \quad (3)$$

Algorithm 1 Personality-space Max-Min Evidence Selection (PMES)

Require: Candidate pool C_t ; target representation z_t ; evidence budget K

Ensure: Selected evidence set E

- 1: $e^{(1)} \leftarrow \arg \max_{e_i \in C_t^{\text{comp}}} \text{sim}(z_t, z_i)$
 - 2: $E \leftarrow \{e^{(1)}\}$
 - 3: **while** $|E| < K$ and $C_t^{\text{comp}} \setminus E \neq \emptyset$ **do**
 - 4: $e^{(|E|+1)} \leftarrow \arg \max_{e_i \in C_t^{\text{comp}} \setminus E} \left[\min_{e_j \in E} d_\sigma(e_i, e_j) \right]$
 - 5: $E \leftarrow E \cup \{e^{(|E|+1)}\}$
 - 6: **end while**
 - 7: Append remaining candidates in decreasing semantic similarity order if $|E| < K$
 - 8: **return** E
-

where s_f^i and s_f^j denote the mapped source scores of cards e_i and e_j on construct f , and σ_f is the standard deviation of construct f computed from complete source-KB profiles.

The first selected card is the semantically nearest complete-profile candidate. Each subsequent card greedily maximizes its minimum personality-space distance to the selected set:

$$e^{(k)} = \arg \max_{e_i \in C_t^{\text{comp}} \setminus E} \min_{e_j \in E} d_\sigma(e_i, e_j), \quad (4)$$

where $e^{(k)}$ is the k -th selected card and E is the set of cards already selected in earlier iterations.

Semantic similarity defines the relevant behavioral region. Personality-space distance then spreads the human-rated profiles across the target construct space within that region. If fewer than K complete-profile candidates are available, remaining positions are filled by candidates in decreasing semantic-similarity order.

The standard deviations in (3) are computed solely from complete source-KB profiles. The greedy selection in (4) also operates only on complete-profile candidates. PMES does not need target personality labels or target-side personality statistics.

E. Evidence-Grounded Continuous Scoring

We reuse the construct space $\mathcal{F}_t$ defined in the problem setting. The multi-rater scorer, hereafter **Cont3**, produces a continuous score vector $\hat{\mathbf{y}}_t \in [1, 5]^{|\mathcal{F}_t|}$ over the constructs in $\mathcal{F}_t$. When the target measures are coarser, a predefined aggregation of the corresponding construct scores is applied consistently to the predictions and the ground-truth targets.

Each independent LLM rater receives the target construct definitions and assessment scenarios, the selected evidence cards with their behavioral descriptions and human-rated profiles, and the target subject's multimodal behavioral description. For construct f , rater m produces a continuous score $y_f^{(m)} \in [1, 5]$.

We instantiate Cont3 with three independent raters, each based on a different LLM. A rater is valid only if it returns a complete score vector over all constructs in $\mathcal{F}_t$. At least two valid raters are required for a subject to be retained. Subjects

TABLE I
BIDIRECTIONAL CROSS-PROTOCOL TRANSFER SETTINGS. *Facet-mean*
DENOTES FACET-LEVEL TARGETS AGGREGATED TO THE FACTOR LEVEL.

Direction	Source	Target	Evaluation space	n
Forward	OPVA	AVI-6	H/X/A/C	644
Reverse	AVI-6	OPVA	X/C (facet-mean)	677

with fewer valid raters are excluded from evaluation. The final construct-level prediction averages the valid judgments:

$$\hat{y}_f = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} y_f^{(m)}, \quad (5)$$

where $\mathcal{M}$ is the set of valid raters. Raters output only the requested continuous scores. No intermediate chain-of-thought or other rater reasoning is used in aggregation.

IV. EXPERIMENTAL SETUP

A. Datasets

We evaluate the framework on two separately collected asynchronous video interview (AVI) datasets, **OPVA** [18] and **AVI-6** [7]. The two datasets differ in question design and in annotation granularity, the two axes that jointly define the cross-protocol transfer setting. AVI-6 provides factor-level annotations over Honesty-Humility (H), Extraversion (X), Agreeableness (A), and Conscientiousness (C). OPVA carries facet-level annotations that are aggregated to the factor level through the predefined target annotation mapping. Both datasets retain the construct annotations needed to map a source profile into the target construct space. The transfer directions differ only in their evaluation space (Table I). The two datasets therefore allow evaluation of the same evidence construction, retrieval, and scoring framework in both transfer directions. For each direction, source-side personality profiles are mapped into the target construct space before knowledge-base indexing. Sample-level filtering is applied as described in the Supplementary Material.

B. Implementation Details

All experiments use the method described in Section III. Retrieval configurations are fixed before target evaluation.

a) Preprocessing: Visual descriptions are generated with Qwen2.5-VL-7B-Instruct [41], and audio descriptions with Fun-Audio-Chat-8B [42]. Dataset transcripts are used directly. Perception prompts restrict outputs to observable behavior. Decoding, sampling, and prompt specifications are provided in the Supplementary Material.

b) Representation Alignment: Cue texts are generated with Qwen3-Max-2026-01-23 and encoded with Octen-Embedding-0.6B [43]. Subject representations average the corresponding construct-level cue vectors. Protocol Projection uses Qwen3.5-Plus, distinct from the cue-generation and scoring models. The closed-source models are used only for representation construction. Personality scoring uses the open-source raters described below. Qwen3.5-Plus also appears as a zero-shot textual baseline in Table II. Temperature,

decoding, and embedding-dimension settings are provided in the Supplementary Material.

c) Retrieval: Retrieval is fixed before target evaluation at $(K_{\text{pre}}, K) = (50, 5)$. The configuration is selected based on the evidence budget, personality-space diversity, and context length (see Section VI for the retrieval-budget analysis). The same configuration is used in both transfer directions without target-side retuning.

d) Multi-rater scoring: We use three open-source LLMs as independent continuous raters: Qwen3-14B [44], Llama-3.1-8B-Instruct [45], and Gemma-3-12B-it [46]. All are served with vLLM at temperature 0 and produce scores in $[1.0, 5.0]$ for all constructs required by the target assessment protocol. Subjects with fewer than two complete rater outputs are excluded, and valid ratings are averaged as defined in Section III-E. The same three models are used individually in the Textualized LLM zero-shot baselines (Table II). The evidence-free ablation applies the Cont3 ensemble without retrieved evidence.

C. Evaluation Metrics

We report mean squared error (MSE) in the original $[1, 5]$ score space (lower is better). MSE is the primary metric because the task is continuous score prediction and MSE measures squared prediction error directly. Mean absolute error (MAE) is reported as a complementary magnitude-oriented measure. Its definition and complete tables are provided in the Supplementary Material. Let N denote the number of target subjects, $\mathcal{E}$ the primary evaluation space of the transfer setting, and y_{if} and $\hat{y}_{if}$ the ground-truth and predicted scores of subject i on evaluation construct f , respectively. For construct f ,

$$\text{MSE}_f = \frac{1}{N} \sum_{i=1}^N (\hat{y}_{if} - y_{if})^2. \quad (6)$$

The primary metric reported in the tables is the average across the evaluation constructs:

$$\overline{\text{MSE}} = \frac{1}{|\mathcal{E}|} \sum_{f \in \mathcal{E}} \text{MSE}_f, \quad (7)$$

where MSE is computed per construct and then averaged across constructs, so that every construct contributes equally. The primary evaluation space is $\mathcal{E} = \{H, X, A, C\}$ for Forward transfer and the aggregated $\mathcal{E} = \{X, C\}$ targets for Reverse transfer. Here $\mathcal{E}$ is the aggregated evaluation view of the target construct space $\mathcal{F}_t$ defined in Section III.

To quantify uncertainty in paired comparisons, we use a paired subject-level bootstrap. For each target subject i , we retain the paired predictions produced by Ours and the comparison condition and compute their subject-level squared-error difference

$$d_i = \frac{1}{|\mathcal{E}|} \sum_{f \in \mathcal{E}} \left[(\hat{y}_{if}^{\text{Ours}} - y_{if})^2 - (\hat{y}_{if}^{\text{Comp}} - y_{if})^2 \right]. \quad (8)$$

The reported ΔMSE is the mean of d_i over all target subjects. We resample target subjects with replacement and keep

the pairing between the two conditions. We then recompute ΔMSE for each bootstrap sample and repeat $B = 10,000$ times. The 95% confidence interval is obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution. This interval captures uncertainty in the paired performance difference rather than assuming the two MSE estimates are independent. A negative ΔMSE favors Ours. An interval entirely below zero means the improvement persists across resamples.

V. RESULTS

A. Comparison with Baselines

We first evaluate whether human-rated behavioral evidence can be transferred across independently collected assessment protocols without target-dataset personality supervision. Throughout the section, we call this supervision-free setting *target-label-free* and refer to a method that additionally receives no retrieved evidence as *zero-shot* (ZS).

The baseline set covers two complementary zero-shot settings. The *Textualized LLM* baselines read the same behavioral descriptions as our framework but score personality directly, without retrieved evidence. Because the input text is held fixed, they isolate the contribution of the retrieved-evidence block. This group includes the three open-source raters used by our scorer, plus their evidence-free three-rater ensemble (Ours w/o evidence, the Cont3 configuration). It also includes two closed-source text models (Qwen3.5-Plus and GPT-5). The *Native AV* baselines operate directly on the raw audio or video stream with the perception models used for description generation in preprocessing (Qwen2.5-VL and Fun-Audio-Chat). They test whether skipping textualization and retrieval is sufficient. Our method is evaluated in both directions using the fixed configuration described in Section IV. Table II summarizes the primary comparison.

We intentionally exclude target-supervised baselines, such as supervised fine-tuning or target-domain label training. These methods require target personality labels and retraining for each new protocol, so they operate at a higher supervision level than the target-label-free setting studied here. The strongest evidence-free configuration (Ours w/o evidence) is therefore the correct reference for isolating the contribution of retrieved human-rated evidence.

In Forward transfer, Ours achieves a primary MSE of 0.2279, a 43.8% error reduction relative to the strongest textualized zero-shot baseline (Ours w/o evidence, 0.4052). In Reverse transfer, it achieves a primary MSE of 0.1651, a 17.0% reduction relative to the strongest evidence-free baseline (Ours w/o evidence, 0.1990). This is also the best MSE among target-label-free methods. The complementary MAE results follow the same ranking and are reported in the Supplementary Material.

At the construct level, the improvement is selective rather than uniform. In Forward transfer, Ours attains the lowest MSE on Honesty-Humility, Agreeableness, and Conscientiousness. Qwen3.5-Plus is slightly better on Extraversion (MSE 0.1937 vs. 0.2249), which is the only Forward construct on which a text-only zero-shot baseline leads the full framework. In Reverse transfer, Ours is lowest on both OPVA constructs under MSE.

Because Ours w/o evidence is the evidence-free version of the scorer used by the full method, we also verify significance against Qwen3.5-Plus, the strongest zero-shot textual baseline outside the Cont3 rater set. All four paired bootstrap intervals lie entirely below zero, so Ours outperforms both the evidence-free ensemble and this external baseline with consistent subject-level significance in both transfer directions (Table III). Relative to Qwen3.5-Plus, the Forward improvement corresponds to a 49.9% relative error reduction.

B. Component Ablation

We next isolate the contributions of retrieved human-rated evidence, personality-space selection, and Representation Alignment using matched configurations. All comparisons use the same target subjects, target annotations, scoring models, and assessment configuration. Only the component under investigation is changed.

We consider three controlled ablations. **Ours w/o evidence** removes the retrieved-evidence block while keeping the construct definitions, assessment scenarios, target behavioral descriptions, and the Cont3 multi-rater scorer unchanged. **Ours w/o PMES** replaces personality-space Max-Min selection with similarity-only Top-5 selection from the same Stage-1 candidate pool, which isolates the effect of personality-space complementarity. **Ours w/o Alignment** removes the entire Representation Alignment pipeline, including Protocol Projection and Cue Construction, from retrieval. Source evidence and target queries then use the corresponding raw behavioral text, while PMES and multi-rater scoring remain unchanged. Because Cue Construction operates on the protocol-projected source representation, Protocol Projection and Cue Construction are ablated jointly rather than evaluated as separate components. All paired comparisons reuse the bootstrap procedure defined in Section IV-C.

Ours w/o evidence increases MSE from 0.2279 to 0.4052 in Forward transfer and from 0.1651 to 0.1990 in Reverse transfer. The paired bootstrap confirms a consistent, subject-level improvement from incorporating retrieved human-rated evidence in both transfer directions (Table III, Ours w/o evidence rows).

Ours w/o PMES increases MSE to 0.2838 in Forward transfer and 0.1750 in Reverse transfer. The paired MSE differences are -0.0559 (95% CI $[-0.0695, -0.0424]$) and -0.0099 (95% CI $[-0.0197, -0.0001]$), respectively. PMES therefore provides a clearer additional gain in Forward transfer and a marginal gain in Reverse transfer, where the CI upper bound is near zero.

Ours w/o Alignment degrades performance substantially. MSE rises to 0.3636 in Forward transfer and 0.4235 in Reverse transfer. The paired differences are -0.1357 (95% CI $[-0.1553, -0.1159]$) and -0.2584 (95% CI $[-0.2835, -0.2333]$). In Reverse transfer this is the largest single-component degradation, whereas in Forward transfer removing retrieved evidence is the largest single-component loss (0.4052).

Taken together, these results show distinct contributions from the three components. Retrieved human-rated evidence

TABLE II

EXTERNAL COMPARISON OF THE FORWARD AND REVERSE TRANSFER SETTINGS (MSE, LOWER IS BETTER). ALL BASELINES ARE ZERO-SHOT (ZS) AND RECEIVE THE SAME BEHAVIORAL DESCRIPTIONS WITHOUT RETRIEVED EVIDENCE. BOLD DENOTES THE BEST (LOWEST) RESULT WITHIN EACH COLUMN.

Mean Squared Error (MSE)									
Dataset		AVI-6					OPVA		
Category	Approach	H	X	A	C	Avg	X	C	Avg
Textualized LLM ZS	Qwen3-14B	0.5282	0.3919	0.5603	0.5817	0.5155	0.2267	0.1788	0.2027
	Llama-3.1-8B	0.4315	0.3652	0.5210	0.4934	0.4528	0.3360	0.2612	0.2986
	Gemma-3-12B	0.5273	0.3156	0.5773	0.6215	0.5104	0.2170	0.1984	0.2077
	Ours w/o evidence	0.4380	0.2971	0.3991	0.4867	0.4052	0.2217	0.1764	0.1990
	Qwen3.5-Plus	0.6212	0.1937	0.3679	0.6371	0.4550	0.2035	0.2396	0.2216
	GPT-5 [47]	0.7759	0.3014	0.6961	0.8095	0.6457	0.2243	0.2786	0.2515
Native AV ZS	Qwen2.5-VL video	1.0503	0.2567	0.6467	0.5655	0.6298	1.3123	0.5871	0.9497
	Fun-Audio-Chat	2.0098	0.4191	1.5715	1.5931	1.3984	0.5265	0.3671	0.4468
Ours	Ours (full)	0.1998	0.2249	0.2032	0.2837	0.2279	0.1726	0.1576	0.1651

TABLE III

PAIRED SUBJECT-LEVEL BOOTSTRAP COMPARISON AGAINST THE EVIDENCE-FREE SCORER (OURS W/O EVIDENCE) AND THE STRONGEST ZERO-SHOT TEXTUAL BASELINE OUTSIDE THE CONT3 RATER SET (QWEN3.5-PLUS).

Dir.	Baseline	MSE	Δ MSE	95% CI
Fwd.	Ours w/o evidence	0.4052	-0.1773	[-0.1990, -0.1563]
	Qwen3.5-Plus	0.4550	-0.2271	[-0.2508, -0.2036]
Rev.	Ours w/o evidence	0.1990	-0.0339	[-0.0481, -0.0201]
	Qwen3.5-Plus	0.2216	-0.0565	[-0.0734, -0.0386]

provides the primary gain over evidence-free scoring in Forward transfer. Representation Alignment provides the primary gain in Reverse transfer and is necessary for cross-protocol comparability in both directions. PMES provides a complementary gain by making the retrieved evidence set more diverse.

VI. DISCUSSION

The results raise two questions about cross-protocol evidence transfer. The first concerns what actually transfers across assessment protocols. A retrieved source record carries observed behavior plus the human personality judgment attached to that behavior. The two need not transfer in the same way. If the personality label transfers rather than the behavior-to-personality association, cross-protocol prediction reduces to nearest-neighbor label copying, where a behaviorally similar source subject is treated as though its personality profile were equally close to that of the target. If only the empirical association transfers, the retrieved record is evidence to interpret rather than a surrogate label. The second concerns how robustly the transferred evidence can be interpreted once it is identified. We examine the two questions with two sets of complementary analyses. The first set tests label transfer directly and diagnoses how the evidence set should be built. The second set decomposes the evidence-grounded scorer into individual models and measures inter-rater reliability.

A. What Transfers, and How Should the Evidence Set Be Built?

Whether a retrieved record should be copied as a label or interpreted as evidence determines how the assessment is designed. The label account is the simpler one, so we rule it out first by replacing evidence-grounded scoring with the simplest possible label transfer. We average the human-rated profiles attached to the retrieved records, which are the Protocol-Projected profiles mapped into the target construct space without any LLM reasoning or statistical rescaling. We compare two neighbor sets. The similarity-only Top-5 set holds the behaviorally most similar records, the set a nearest-neighbor label copier would use. The PMES set is the complementary evidence set the method actually uses. Table V compares Cont3 with averaging the same records' personality labels. All four subject-level Δ MSE intervals exclude zero, so on either retrieval set treating the records as evidence outperforms treating their scores as transferable labels. Diversifying the evidence has a direction-dependent effect on label averaging, improving the Forward estimate and worsening it in Reverse. Diversification does not reliably rescue label transfer. Behaviorally similar source subjects are therefore not reliable surrogate labels for the target. A personality rating is attached to an individual observed under a particular protocol. Behavioral similarity across protocols does not guarantee an equivalent underlying profile. The retrieved record is useful as empirical evidence about how human raters have interpreted a behavioral pattern. The final prediction remains a judgment about the target.

This interpretation also clarifies the role of PMES. If retrieved records are evidence rather than labels, retrieval should build a set whose members provide complementary information instead of finding the most behaviorally similar subjects. Table VI shows the trade-off. The reported minimum d_σ is the minimum pairwise standardized personality-space distance within the selected evidence set, averaged over queries. PMES maximizes this quantity, which therefore serves as the primary diversity statistic. Enlarging the candidate pool

TABLE IV
COMPONENT ABLATION RESULTS ACROSS FORWARD AND REVERSE TRANSFER (MSE, LOWER IS BETTER). ALL RETRIEVAL-BASED CONFIGURATIONS USE CONT3 WITH $(K_{\text{pre}}, K) = (50, 5)$. BOLD DENOTES THE BEST (LOWEST) RESULT WITHIN EACH COLUMN.

		Mean Squared Error (MSE)							
Dataset		AVI-6					OPVA		
Category	Configuration	H	X	A	C	Avg	X	C	Avg
Evidence retrieval	Ours w/o evidence	0.4380	0.2971	0.3991	0.4867	0.4052	0.2217	0.1764	0.1990
	Ours w/o PMES	0.2084	0.2728	0.1952	0.4586	0.2838	0.1766	0.1734	0.1750
Representation Alignment	Ours w/o Alignment	0.2517	0.3980	0.2847	0.5199	0.3636	0.3465	0.5005	0.4235
Full model	Ours (full)	0.1998	0.2249	0.2032	0.2837	0.2279	0.1726	0.1576	0.1651

TABLE V
EVIDENCE-GROUNDED CONT3 VS. NEIGHBOR-LABEL MEAN UNDER MATCHED RETRIEVAL SETS ($K=5$). NEGATIVE ΔMSE FAVORS CONT3.

Retrieval Predictor	Fwd.	Rev.
Cont3	0.2838	0.1750
Sim-5 Neighbor	0.3550	0.3341
ΔMSE	-0.0713	-0.1591
CI: [-.0839, -.0588]; [-.1783, -.1401]		
Cont3 (Ours)	0.2279	0.1651
PMES Neighbor	0.2905	0.3792
ΔMSE	-0.0626	-0.2141
CI: [-.0774, -.0481]; [-.2363, -.1923]		

TABLE VI
FORWARD OFFLINE RETRIEVAL-BUDGET DIAGNOSTICS ($n=644$).

(a) Candidate-pool size (fixed $K=5$)			
K_{pre}	20	50	100
Minimum d_σ	2.08	2.54	2.85
(b) Evidence budget (fixed $K_{\text{pre}}=50$)			
K	3	5	10
Minimum d_σ	3.19	2.54	1.82
Context (tokens)	~8.4k	~9.6k	~12.6k

raises the minimum pairwise d_σ within the selected records, while enlarging the evidence budget lowers it, because more records are packed into a bounded selection rather than adding redundant information. We therefore fix $(K_{\text{pre}}, K) = (50, 5)$ on engineering grounds. This choice is not tuned to downstream prediction error, as doing so would compromise the zero-shot setting. It is also not a claim that a larger evidence budget is always better. Because the budget is a single global setting applied identically in both directions, we report the diagnostic on the Forward queries. The trade-off concerns the construction of the evidence set rather than its interpretation.

Cross-protocol retrieval should therefore be formulated as evidence construction rather than nearest-neighbor label transfer. The retrieved records matter because they jointly carry complementary information about how observed behavior is interpreted, not because any one of them is behaviorally similar to the target.

B. How Robustly Can Transferred Evidence Be Interpreted?

Even with a useful evidence set, the final assessment depends on the model that interprets it. Model-specific variation is a distinct failure mode from evidence-selection error. We therefore separate two quantities that are easy to conflate: prediction error (MSE) and inter-rater reliability (ICC). Reliability asks whether the scoring process is stable across independent models, and error asks whether the prediction is close to the target. Aggregation is not a uniformly stronger predictor. In Forward transfer, the mean of the three individual raters, denoted Cont3, is clearly better than the best single model. In Reverse transfer, Cont3 is essentially tied with the strongest individual rater (Qwen3-14B). The direction-dependent benefit of averaging motivates measuring reliability rather than error alone.

Reliability is quantified with the Shrout-Fleiss intraclass correlation under a two-way random-effects model with absolute agreement. It is computed separately for each construct (facet in Reverse) from the three scoring models' per-subject predictions. The single-rater reliability is

$$\text{ICC}(2, 1) = \frac{\text{MS}_B - \text{MS}_E}{\text{MS}_B + (k - 1) \text{MS}_E + \frac{k}{n} (\text{MS}_J - \text{MS}_E)}, \quad (9)$$

where MS_B is the between-subjects mean square, MS_J the between-raters mean square, MS_E the residual mean square, $k = 3$ the number of raters, and n the number of subjects. The reliability of the three-rater mean follows from the Spearman-Brown formula

$$\text{ICC}(2, 3) = \frac{3 \text{ICC}(2, 1)}{1 + 2 \text{ICC}(2, 1)}. \quad (10)$$

In Forward transfer, $\text{ICC}(2, 1)$ ranges from 0.272 to 0.678 and $\text{ICC}(2, 3)$ from 0.529 to 0.863. In Reverse the corresponding ranges are 0.500 to 0.834 and 0.750 to 0.938. Median $\text{ICC}(2, 1)$ is higher in Reverse than in Forward, but the construct-wise ranges still overlap. We treat these as descriptive only because Forward and Reverse use different construct granularities. The low end of the Forward range indicates that for some constructs the three models agree only weakly, so agreement is not uniformly high in either direction.

Pairwise Pearson correlations among the three raters range from 0.532 to 0.743 in Forward and 0.746 to 0.803 in Reverse, separating the two directions more cleanly than the overlapping ICC ranges. Both sets of statistics are point estimates

without confidence intervals or a formal between-direction test. In Forward, raters disagree more strongly and averaging has more variation to attenuate, which match the larger MSE gain. In Reverse, raters already agree more and the additional gain is small. Reliability and prediction error are therefore complementary: MSE evaluates closeness to the target rating, while ICC characterizes the stability of the scoring process.

Taken together, the two sets of analyses support cross-protocol personality assessment as empirical evidence transfer rather than direct label transfer, with retrieval framed as evidence construction and multi-rater aggregation as a robustness mechanism.

VII. LIMITATIONS

Two limitations bound the current results.

First, building the source knowledge base introduces an information bottleneck. Protocol Projection and Cue Construction convert multimodal behavioral observations into target-protocol-compatible semantic text. This enables cross-protocol retrieval but trades off modality-specific signals and fine-grained temporal structure. Retaining raw video and audio would reduce this loss but sharply increase retrieval cost. Future work should explore adaptive multimodal representations that use compact semantic representations for efficient retrieval while retaining raw or temporally localized evidence.

Second, the current evaluation covers two independently collected interview datasets and one set of scoring models. Bidirectional transfer provides evidence beyond a single direction, but broader evaluation across protocols, participant populations, and model families is needed to establish generality. A supervised upper bound is not reported, so the performance cost of forgoing target supervision remains unquantified.

VIII. CONCLUSION

This work addressed cross-protocol personality assessment: reusing human-rated behavioral evidence collected under one protocol to score personality under another without target-dataset supervision. We proposed CEPA (Cross-protocol Evidence-grounded Personality Assessment), a retrieval-augmented framework that aligns multimodal records to the target protocol, builds a compact evidence pool, and interprets that evidence through multi-rater scoring.

Across bidirectional transfer between OPVA and AVI-6, CEPA outperforms direct label transfer, confirming that cross-protocol assessment is feasible without target labels.

Two principles emerge. First, cross-protocol transfer requires behavioral comparability rather than direct personality-label transfer. Second, the value of retrieved evidence depends on complementary organization and joint interpretation under a limited budget. What transfers is the observed behavior-to-judgment association, not the source label itself. Extending CEPA to broader protocols, participant populations, and model families is a natural next step.

REFERENCES

- [1] F. S. Brenner, T. M. Ortner, and D. Fay, "Asynchronous video interviewing as a new technology in personnel selection: The applicant's point of view," *Frontiers in Psychology*, vol. 7, p. 863, 2016.
- [2] H.-Y. Suen, K.-E. Hung, and C.-L. Lin, "Tensorflow-based automatic personality recognition used in asynchronous video interviews," 2019. [Online]. Available: <https://ieeexplore.ieee.org/document/8660507/>
- [3] L. Hickman, N. Bosch, V. Ng, R. Saef, L. Tay, and S. E. Woo, "Automated video interview personality assessments: Reliability, validity, and generalizability investigations," *Journal of Applied Psychology*, vol. 107, no. 8, p. 1323, 2022.
- [4] J. Liff, N. Mondragon, C. Gardner, C. J. Hartwell, and A. Bradshaw, "Psychometric properties of automated video interview competency assessments," *Journal of Applied Psychology*, vol. 109, no. 6, pp. 921–948, 2024.
- [5] A. Koutsoumpis *et al.*, "Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning," 2024. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S074756322300479X>
- [6] J. Lv, C. Chen, and Z. Liang, "Automated scoring of asynchronous interview videos based on multi-modal window-consistency fusion," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 799–814, 2024.
- [7] T. Zhang *et al.*, "Assessing personality traits and interview performance from asynchronous video interviews," in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 13895–13900. [Online]. Available: <https://doi.org/10.1145/3746027.3762016>
- [8] R. P. Tett and D. D. Burnett, "A personality trait-based interactionist model of job performance," 2003.
- [9] X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," 2018. [Online]. Available: <https://ieeexplore.ieee.org/document/8066355/>
- [10] Y. Li *et al.*, "Cr-net: A deep classification-regression network for multimodal apparent personality analysis," 2020. [Online]. Available: <http://link.springer.com/10.1007/s11263-020-01309-y>
- [11] Y. Wang *et al.*, "Inside-out personality assessment: From latent self-report traits to apparent multimodal behaviors," 2018.
- [12] H. Peters and S. C. Matz, "Large language models can infer psychological dispositions of social media users," *PNAS Nexus*, vol. 3, no. 6, p. pgae231, 06 2024. [Online]. Available: <https://doi.org/10.1093/pnasnexus/pgae231>
- [13] A. G. C. Wright *et al.*, "Assessing personality using zero-shot generative ai scoring of brief open-ended text," 2026. [Online]. Available: <https://www.nature.com/articles/s41562-025-02389-x>
- [14] A. V. Ganesan, Y. K. Lal, A. Nilsson, and H. A. Schwartz, "Systematic evaluation of GPT-3 for zero-shot personality estimation," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, J. Barnes, O. De Clercq, and R. Klinger, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 390–400. [Online]. Available: <https://aclanthology.org/2023.wassa-1.34/>
- [15] Y. Ji, W. Wu, H. Zheng, Y. Hu, X. Chen, and L. He, "Is chatgpt a good personality recognizer? a preliminary study," 2023. [Online]. Available: <http://arxiv.org/abs/2307.03952>
- [16] 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*, vol. 15, no. 3, pp. 1769–1785, Sep. 2024.
- [17] L. Shen, Y. Long, X. Cai, G. Chen, I. Razzak, and S. Jameel, "Less but better: Parameter-efficient fine-tuning of large language models for personality detection," 2025. [Online]. Available: <http://arxiv.org/abs/2504.05411>
- [18] T. Zhang *et al.*, "Personalityllm: Fine-tuning large language models for personality assessment from asynchronous video interviews," 2026. [Online]. Available: <https://ieeexplore.ieee.org/document/11481797/>
- [19] M. Gavrilescu and N. Vizireanu, "Predicting the sixteen personality factors (16pf) of an individual by analyzing facial features," 2017. [Online]. Available: <https://jivp-eurasipjournals.springeropen.com/articles/10.1186/s13640-017-0211-4>
- [20] J. Li *et al.*, "Traits run deeper: Trait-specific asymmetric fusion for personality assessment," 2018.
- [21] X. Zhao *et al.*, "Integrating audio and visual modalities for multimodal personality trait recognition via hybrid deep learning," 2023. [Online]. Available: <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.1107284/full>
- [22] 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," 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/9373959/>

- [23] Y. Bao, X. Liu, Y. Qi, R. Liu, and H. Li, "Adaptive information fusion network for multi-modal personality recognition," 2024. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cav.2268>
- [24] X. Duan, H. Li, F. Yang, B. Chen, J. Dong, and Y. Wang, "Multimodal automatic personality perception using vit, bilstm and vggish," 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10604109/?arnumber=10604109>
- [25] T. Zhang, W. Shan, Y. Zong, T. Qi, and W. Zheng, "Llm-based multimodal personality recognition via facial action unit-text semantic fusion," *IEEE Transactions on Affective Computing*, 2026. [Online]. Available: <https://ieeexplore.ieee.org/document/11610589/>
- [26] T. Zhang, S. Liang, W. Zheng, A. Koutsoumpis, J. K. Oostrom, and R. E. de Vries, "Mixture-of-Expert Large Language Models for Text-Based Personality Assessment From Asynchronous Video Interviews," *IEEE Transactions on Affective Computing*, vol. 17, no. 02, pp. 2171–2187, Apr. 2026. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/TAFFC.2026.3667846>
- [27] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," 2020. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.550/>
- [28] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.
- [29] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Realm: Retrieval-augmented language model pre-training," 2020. [Online]. Available: <http://arxiv.org/abs/2002.08909>
- [30] S. Borgeaud et al., "Improving language models by retrieving from trillions of tokens," 2022. [Online]. Available: <https://proceedings.mlr.press/v162/borgeaud22a.html>
- [31] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," 2023. [Online]. Available: <https://www.semanticscholar.org/paper/46f9f7b8f88f72e12cbdb21e3311f995eb6e65c5>
- [32] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," 1998. [Online]. Available: <https://dl.acm.org/doi/10.1145/290941.291025>
- [33] Z. Yang et al., "Representative demonstration selection for in-context learning with two-stage determinantal point process," 2023. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.331/>
- [34] S. H. Khan et al., "Df-rag: Query-aware diversity for retrieval-augmented generation," 2026. [Online]. Available: <https://aclanthology.org/2026.findings-eacl.150/>
- [35] I. Levy, B. Bogin, and J. Berant, "Diverse demonstrations improve in-context compositional generalization," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1401–1422. [Online]. Available: <https://aclanthology.org/2023.acl-long.78/>
- [36] D. W. Kolar, D. C. Funder, and C. R. Colvin, "Comparing the accuracy of personality judgments by the self and knowledgeable others," 1996. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1111/j.1467-6494.1996.tb00513.x>
- [37] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1180–1189. [Online]. Available: <https://proceedings.mlr.press/v37/ganin15.html>
- [38] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 97–105. [Online]. Available: <https://proceedings.mlr.press/v37/long15.html>
- [39] S. An, X. Sun, Y. Li, Y. Yang, and G. Xu, "Personality analysis from online short video platforms with multi-domain adaptation," 2024. [Online]. Available: <https://arxiv.org/abs/2411.00813>
- [40] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," 2023. [Online]. Available: <https://aclanthology.org/2023.tacl-1.1/>
- [41] S. Bai et al., "Qwen2.5-vl technical report," 2025. [Online]. Available: <http://arxiv.org/abs/2502.13923>
- [42] T. F. Team et al., "Fun-audio-chat technical report," 2026. [Online]. Available: <http://arxiv.org/abs/2512.20156>
- [43] O. Team, "Octen series: Optimizing embedding models to #1 on rteb leaderboard," 2025. [Online]. Available: https://octen-team.github.io/octen_blog/posts/octen-rteb-first-place/
- [44] A. Yang et al., "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
- [45] A. Grattafiori et al., "The llama 3 herd of models," 2024. [Online]. Available: <https://arxiv.org/abs/2407.21783>
- [46] G. Team et al., "Gemma 3 technical report," 2025. [Online]. Available: <https://arxiv.org/abs/2503.19786>
- [47] A. Singh et al., "Openai gpt-5 system card," 2026. [Online]. Available: <https://arxiv.org/abs/2601.03267>



Wei Shan received the bachelor's degree Nanjing Forestry University in 2024. He is currently pursuing the master's degree with the Southeast University-Monash University Joint Graduate School. His current research interests include large language models, data mining, and personality analysis.



Associate Editor of *IEEE Transactions on Affective Computing*.

Tianyi Zhang (Senior Member, IEEE) is an associate professor at the School of Biological Sciences 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 affective computing and personality recognition. He is an



Yuan Zong (Member, IEEE) 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 of Ministry of Education, Southeast University. From 2016 to 2017, he was a Visiting Student with the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland. His research interests include affective computing, pattern recognition, computer vision, and speech signal processing.



Wenming Zheng (Senior Member, IEEE) received the Ph.D. degree in signal processing from Southeast University. Since 2004, he has been with the Research Center for Learning Science, Southeast University. He is currently a Full 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 an IET Fellow since 2022. He is an Associate Editor of *IEEE Transactions on Affective Computing* and the Editorial Board Member of the *Visual Computer*.