# Modelling the Interplay of Eye-Tracking Temporal Dynamics and Personality for Emotion Detection in Face-to-Face Settings

Meisam J. Seikavandi[1,3]
meis@itu.dk
Jostein Fimland[1]
jostein.fimland@gmail.com
Fabricio Batista Narcizo[2,3]
fabn@itu.dk
Maria Barrett[2]
mria.barrett@gmail.com
Ted Vucurevich[3]
tvucurevich@gn.com
Jesper Bünsow Boldt[3]
jboldt@gn.com
Andrew Burke Dittberner[3]
adittberner@gn.com
Paolo Burelli[1]
pabu@itu.dk

[1] brAIn Lab
IT University of Copenhagen
Copenhagen, Denmark

[2] IT University of Copenhagen
Copenhagen, Denmark

[3] GN Advanced Science
GN Group
Ballerup, Denmark

**Abstract**

Accurate recognition of human emotions is critical for adaptive humancomputer interaction, yet remains challenging in dynamic, conversation-like settings. This work presents a personality-aware multimodal framework that integrates **eye-tracking sequences**, **Big Five personality traits**, and **contextual stimulus cues** to predict both *perceived* and *felt* emotions. Seventy-three participants viewed speech-containing clips from the CREMA-D dataset while providing eye-tracking signals, personality assessments, and emotion ratings. Our neural models captured temporal gaze dynamics and fused them with trait and stimulus information, yielding consistent gains over SVM and literature baselines. Results show that (i) stimulus cues strongly enhance perceived-emotion predictions (macro F1 up to **0.77**), while (ii) personality traits provide the largest improvements for *felt* emotion recognition (macro F1 up to **0.58**). These findings highlight the benefit of combining physiological, trait-level, and contextual information to address the inherent subjectivity of emotion. By distinguishing between perceived and felt responses, our approach advances multimodal affective computing and points toward more personalized and ecologically valid emotion-aware systems.

# 1 Introduction

Emotion recognition is a central challenge in affective computing, with applications in adaptive HCI, virtual agents, education, and teleconferencing. Accurate recognition enables systems to respond more personally and context-sensitively. Yet most models still rely on simplified representationsstatic snapshots or coarse labelsthat fail to capture the richness of real interactions. Genuine perception unfolds dynamically, shaped by subtle attentional cues and modulated by stable traits such as personality [20, 22, 33].

Theoretically, this lies at the intersection of two perspectives. *Basic Emotion Theory* (BET) [13] treats emotions as discrete, biologically hard-wired categories, while constructionist accounts such as the *Theory of Constructed Emotion* (TCE) [3] emphasize interactions between core affect and conceptual knowledge. A layered view suggests both matter: observers may recognize expressed cues (BET) while constructing distinct internal experiences (TCE).

This distinction is salient in the *listeners perspective*. In meetings, video calls, or virtual-agent interactions, people interpret a talking face without full turn-taking dialogue. Such contexts approximate dialogue while retaining one-sided attention. Our study targets this scenario, where participants watched CREMA-D clips [5], allowing us to probe convergence and divergence between perceived and felt emotions.

Key challenges remain: neglected temporal dynamics, under-modeled individual differences, and divergence between perceived and felt states [3]. To address them, we adopt a framework distinguishing *Expressed Emotions* ($E_e$), *Perceived Emotions* ($E_p$), and *Felt Emotions* ($E_f$). Encountering a stimulus involves recognizing $E_e$, forming a perception $E_p$, and potentially experiencing a distinct $E_f$. Modeling both $E_p$ and $E_f$ in parallel is therefore essential.

We introduce a multimodal approach that integrates **eye-tracking data**, **temporal modeling**, and **personality traits** to predict $E_p$ and $E_f$ in speech-based settings. Seventy-three *non-actor participants* contributed (1) detailed eye movements (fixations, pupil size), (2) self-reported Big Five profiles [15, 21], and (3) perceived and felt ratings per trial. Our contributions are threefold:

1. **Integration:** Combining eye-tracking, personality, and temporal dynamics improves recognition in talking-face scenarios.
2. **Personality effects:** Traits modulate both how participants *perceive* others emotions and how they *feel*.
3. **Modeling:** A multimodal neural architecture achieves strong predictive performance for both perceived and felt states, relevant for adaptive, user-centered affective computing.

By jointly modeling dynamic cues, stable traits, and the divergence between perceived and felt states, our work moves beyond static or acted benchmarks. Although not fully interactive, it reflects a critical real-world pattern: emotional decoding by the listener. This perspective is essential for developing next-generation recognition systems that capture the layered nature of human emotion.

# 2 Background

A wide range of studies has explored different modalities and methodologies, often achieving impressive accuracy levels. However, many of these works simplify emotion detection by focusing on limited arousal and valence scores or by relying on highly controlled datasets. In practice, real-world emotion perception is shaped by multiple, often interdependent factors such as personality traits, gaze patterns, and contextual cues. Consequently, several

important challenges remain insufficiently addressed, especially regarding the use of non-actor participants, individual differences, and temporal dynamics. These unresolved issues echo long-standing theoretical debates, e.g. whether emotions are discrete biological kinds or context-dependent constructions, highlighting the need for frameworks that link physiology, attention, and conceptual labeling in a single pipeline.

## 2.1 Emotion Models

Two primary frameworks characterize how emotions are commonly modeled. **Discrete models** rooted in *Basic Emotion Theory* (BET) [13] group emotions into basic categories such as anger, disgust, fear, joy, sadness, and surprise. These labels are intuitive and map neatly onto facial action patterns, but they often struggle to capture subtle or mixed states in realistic settings [37]. **Dimensional models**, by contrast, represent affect along continuous axes such as arousal and valence [29, 34], occasionally adding dominance as a third dimension. They allow nuanced representations, yet many engineering studies discretize them into low/medium/high bins, obscuring fine-grained shifts.

**Trade-offs and hybrid views.** Discrete categories can over-simplify overlapping expressions, whereas purely dimensional schemes cannot easily separate qualitatively distinct emotions that share similar core affect (e.g. anger vs. fear). Recent evidence shows that subjective reports cluster into at least 27 categories connected by smooth gradients [12], motivating *hybrid* pipelines that first estimate core affect and then map it onto discrete concepts, an approach compatible with *Theory of Constructed Emotion* (TCE) [3]. Following Van Heijst et al. [41], we view BET and TCE as complementary layers: BET explains why evolution endowed us with affect programs; TCE explains how any given episode is constructed from core affect plus conceptual knowledge. This layered stance underpins our decision to model both continuous ($E_f$) and categorical ($E_p$) labels in later sections.

## 2.2 Multimodal Emotion Recognition Approaches

Leveraging multiple modalities facial expressions, vocal prosody, and physiological signals has yielded robust gains. For example, Kollias et al. [24] used multi-task audio-visual learning to detect valence, arousal, expressions, and action units. Reviews by Li et al. [26] and Zhang et al. [44] report strong scores in controlled labs. Yet most pipelines conflate low-level arousal cues with high-level categorical labels, ignoring the layered distinction between core affect and conceptual emotion. Moreover, many still rely on *actors* or compress ratings to binaries, limiting ecological validity.

Practical constraints persist: collecting EEG or GSR requires specialized hardware, so researchers increasingly explore more accessible channels such as eye tracking or basic speech while still capturing real-world complexity.

## 2.3 Eye-Tracking-Based Emotion Recognition

Among visual modalities, **eye tracking** uniquely captures both attentional focus (fixations) and arousal (pupil dilation) [30]. Eye movements correlate with emotional states [36] and reveal which facial regions observers deem salient [38]. Lu et al. [28] combined eye tracking with EEG to boost accuracy. Attachment style and personality dimensions modulate pupillary responses e.g. avoidant individuals show blunted dilation to happy faces, underscoring the value of trait-aware models [40]. Nevertheless, lighting, eyewear, or calibration drift can degrade data quality.

### 2.4 Gaze Strategies

Eye-gaze strategies how observers allocate fixations across a face offer insight into emotion-specific cues. Patterns vary by age [6] and gender [11]. Different regions (eyes vs. mouth) carry diagnostic weight for particular emotions [35], and gaze direction modulates perception [27]. Systems focusing solely on the eyes may thus miss critical mouth cues, and vice versa.

### 2.5 Stimuli and Participant Considerations

A recurring critique is reliance on *actors* displaying prototypical expressions [9, 31]. Such datasets inflate accuracy yet transfer poorly to spontaneous contexts. Recruiting **non-actors** and using more naturalistic stimuli improves ecological validity but increases variability. Our talking face paradigm with non-actor participants aims to balance realism and control.

### 2.6 Temporal Dynamics in Emotion Recognition

Most pipelines still treat emotions as static snapshots, ignoring their evolution. Wang et al. [42] emphasize the need for sequence modeling to capture rapid affective transitions; without temporal context, fleeting cues may be misinterpreted.

### 2.7 Personality and Emotion Recognition

Personality, typically the Big Five shapes both expression and perception [23, 43]. Neurotic individuals fixate on negative content; extraverts seek positive cues [10]. Eye-tracking studies now infer personality traits themselves [1, 7, 33], paving the way for adaptive systems. These advances raise privacy concerns and demand careful trait inference.

### 2.8 Personality-Inspired Eye-Tracking-Based Emotion Recognition

Incorporating personality scores can improve gaze-based emotion recognition. High-neuroticism observers linger on negative features [8]; extraverts scan positive cues [18]. Real-world validations remain scarce, and challenges include maintaining calibration and safeguarding privacy. Nevertheless, personality-aware pipelines represent a stride toward user-centric, empathetic computing.

In sum, emotion-recognition pipelines still over-rely on acted stimuli, neglect individual differences, and ignore layered temporal dynamicsgaps our study addresses by combining naturalistic talking-face stimuli with eye-tracking and personality traits in a BETTCE framework.

## 3 Dataset Collection and Preprocessing

### 3.1 Participants

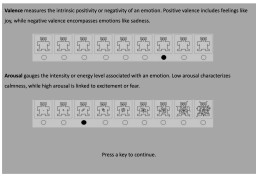
We recruited 73 participants (52 males, 21 females; mean age $27.4 \pm 6$ years). All participants reported normal or corrected-to-normal vision and no neurological disorders. Participants came from diverse educational backgrounds (see Table 1). The participants agreed and signed the informed consent following the university's ethical guidelines. Although this sample provides educational diversity, we note a moderate gender imbalance and a relatively young average age, which may limit broader generalizability.

### 3.2 Experimental Design and Procedure

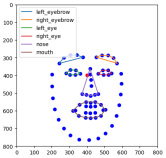
We simulated the listening aspect of a conversational setting where participants engaged with dynamic, emotionally expressive stimuli. Each participant completed 88 trials (4 practice

(a) Experimental setup

(b) Stimulus frame




(c) Rating scales




(d) Facial landmarks and ROIs

Figure 1: (a) experimental setup with eye tracker, (b) example stimulus frame from CREMA-D, (c) 9-point arousal/valence rating scales, and (d) facial landmarks extracted by OpenFace [2] partitioned into multiple ROIs.

Table 1: Participant demographics (N=73).

| Gender (M/F) | Avg. Age | Glasses | College | Bachelor | Master | PhD |
|---|---|---|---|---|---|---|
| 72% / 28% | $27.4 \pm 6$ | 33% | 28% | 43% | 19% | 10% |

and 84 main) in random order. Stimuli were 84 video clips from the CREMA-D dataset [5], featuring 91 actors (48 male, 43 female) aged 20–74, each portraying one of six basic emotions (Anger, Disgust, Fear, Happy, Neutral, Sad) at varying intensities. The selected clips balanced emotions and actor demographics to enhance expressiveness and generalizability.

To approximate face-to-face interaction, a short written scenario was displayed before each video, prompting participants to imagine conversing with the individual shown. This contextual priming aimed to increase engagement and emotional alignment, despite the lack of true turn-taking. The use of short textual prompts was intended not only to simulate real conversational framing but also to standardize participants cognitive approach, ensuring consistent engagement across trials rather than replicating spontaneous dialogues.

Eye-tracking data were recorded using a GP3 HD eye tracker at 150 Hz. The eye tracker was calibrated for each participant with a standard 9-point procedure. We synchronized data collection with stimulus presentation via the Lab Streaming Layer (LSL) to ensure precise alignment between eye-tracking data and stimulus onset.

Figure 1(a) illustrates the experimental setup, with the participant seated in front of the monitor and wearing sensors.

Before starting the trials, participants completed the BFI-44 questionnaire [15] to assess openness, conscientiousness, extraversion, agreeableness, and neuroticism. After each video, participants rated their *perceived* and *felt* emotions on 9-point Likert scales for valence (1 = very negative, 9 = very positive) and arousal (1 = very calm, 9 = very excited). We chose a 9-point scale for its higher resolution, capturing more subtle affective nuances [4, 25] compared to smaller scales. These self-reported ratings form the ground truth labels for our emotion recognition models.

## 3.3 Data Preprocessing and Feature Extraction

Preprocessing involved quality filtering (blinks/tracking loss), normalization of gaze coordinates, and baseline correction of pupil size. For each trial we extracted: **fixation metrics** (duration/dispersion), **pupil metrics** (mean, min, max, variance), **saccadic metrics** (amplitude, duration, peak velocity, acceleration), **gaze regions** (eyes, eyebrows, nose, mouth, outside, from OpenFace [2]), **environmental variables** (ambient light, temperature, stimulus brightness), and **personality traits** (BFI-44, scaled to [0,1]). These features span dynamic gaze signals, static trait/context variables, and categorical stimulus labels (Table 3).

Table 2: Feature inventory.

| Modality | Variables | Dim. | Temp. | Notes |
|---|---|---|---|---|
| Eye-tracking | FixDur, Pupil, Sacc., Regions | 15×12 | ✓ | Sequential |
| Personality | O,C,E,A,N | 5 | ✗ | Static |
| Stimulus emo | One-hot | 6 | ✗ | Contextual |
| Environment | Lux, Temp | 2 | ✗ | Control |

Table 3: Feature inventory across modalities.

| Modality | Variables | Dim. | Temporal | Notes |
|---|---|---|---|---|
| Eye-tracking | FixDur, PupilMean, Saccades, Regions | 15 | ✓ | Sequential |
| Personality | Big-Five (O,C,E,A,N) | 5 | ✗ | Static |
| Stimulus emo | One-hot (6) | 6 | ✗ | Contextual |
| Environment | Lux, Temp | 2 | ✗ | Control |

### 3.4 Features for Modeling

Because fixations and saccades vary in frequency and timing, we standardized the temporal dimension via **interpolation** into 15 equally spaced time steps per trial. This uniform representation accommodates 24 second videos and facilitates sequential modeling with architectures such as Long Short-Term Memory (LSTM) networks [17]. Interpolating to equal-length sequences allows us to treat each trial as a short emotion episode in the layered-affect sense, i.e., a window where core-affect fluctuations (pupil, arousal) can be mapped to conceptual labels.

Each time step includes gaze-region allocations, pupil size, and saccadic measures, capturing how attention and arousal evolve over time. By integrating these time-series features with static context variables (environment and personality), our models capitalize on both dynamic and trait-level information to boost emotion recognition accuracy.

## 4 Machine Learning Modeling

### 4.1 Emotion Labeling and Data Preparation

We aimed to predict four emotion labels: **felt valence**, **perceived valence**, **felt arousal**, and **perceived arousal**. Given the imprecision in self-reported data, each label was grouped into three classes: low/negative (1–3), medium/neutral (4–6), and high/positive (7–9). We divided the dataset into training (64%), validation (16%), and testing (20%) subsets using stratified splits to maintain class distribution. These splits were not strictly subject-independent. To reduce the risk of personality vectors being memorized across folds, we injected small random Gaussian noise into personality scores during training for each trial, which acted as a regularization strategy.

While this binning approach improves model stability, it may mask finer affective distinctions, so future research could explore regression-based or ordinal classification. Following prior studies [14, 16, 19], binning continuous ratings also mitigates subjective variability in self-reported emotions. Nonetheless, some granularity is inevitably lost.

### 4.2 Feature Engineering and Preprocessing

#### Normalization and Scaling

To ensure fair feature contribution and reduce bias, we applied consistent preprocessing. Personality trait scores (0–50) were scaled by dividing by 50 [32]. Highly skewed features, like saccade amplitude/duration, were transformed with `MinMaxScaler`. Other continuous features (pupil sizes, environment variables) were standardized using `StandardScaler`, subtracting the training-set mean and dividing by its standard deviation to avoid leakage.

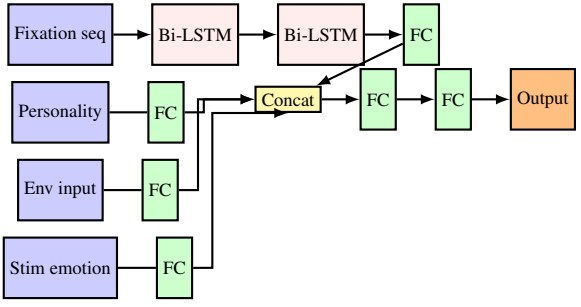



Figure 2: Architecture integrating eye, personality, environment, and stimulus emotion.

### One-Hot Encoding

We encoded stimulus emotion (happy, sad, neutral, angry, disgust, fear) as a 6-dimensional one-hot vector, allowing the model to distinguish each emotion category independently.

## 4.3 Neural Network Architecture

Our neural network (NN) integrates multiple input streams (Figure 2). Each stream undergoes separate preprocessing before feature fusion. This design ensures that temporal features (e.g., eye-tracking data) and static features (e.g., personality traits, stimulus emotion) receive tailored treatment for their respective roles in predicting emotional states.

**Eye-Tracking Data**: Processed through LSTM layers to capture temporal dependencies.
**Personality Traits**: Processed through fully connected layers.
**Stimulus Emotion**: One-hot encoded and passed through fully connected layers.
**Environmental Variables**: Processed through fully connected layers.

To mitigate overfitting on personality or environmental variables, we injected small Gaussian noise into these inputs during training as a form of data augmentation. This noise helps the model generalize across participants and conditions.

## 4.4 Classification Approach

We framed emotion prediction as a three-class classification task, applying softmax activation for class probabilities. We used categorical cross-entropy loss with inversely proportional class weights to handle imbalance.

## 4.5 Model Training and Evaluation

We performed manual and grid search hyperparameter tuning for our neural networks. The search space covered learning rates $\{10^{-3}, 10^{-4}, 10^{-5}\}$, dropout rates $\{0.2, 0.3, 0.5\}$, and weight decay values. We selected the best configurations based on macro F1 scores on the validation set. Training employed early stopping, halting if validation did not improve within 10 epochs.

Hyperparameters like learning rate, dropout, and weight decay were chosen per validation performance. We used the F1-score for evaluation due to dataset imbalance, as it considers both precision and recall.

We compared our NN models with support vector machines (SVMs) as baselines. The SVMs used stimulus emotion alone or combined with personality data. We chose SVM because it is simple, effective for non-temporal data, and reveals the benefit of adding sequential modeling in the NN.

Table 4: Model performance (F1-scores) and hyperparameters (Learning Rate and Dropout) for different input features.

| | Low | Medium | High | Macro F1 | Learning Rate | Dropout |
|---|---|---|---|---|---|---|
| **NN with Eye-Tracking Data (No Env)** | | | | | | |
| Perceived Arousal | 0.32 | 0.54 | 0.17 | 0.34 | 0.0002 | 0.3 |
| Perceived Valence | 0.35 | 0.22 | 0.28 | 0.28 | 0.0002 | 0.3 |
| Felt Arousal | 0.45 | 0.37 | 0.07 | 0.30 | 0.0003 | 0.2 |
| Felt Valence | 0.29 | 0.49 | 0.24 | 0.34 | 0.0002 | 0.3 |
| **NN with Eye-Tracking Data** | | | | | | |
| Perceived Arousal | 0.18 | 0.57 | 0.24 | 0.33 | 0.00035 | 0.3 |
| Perceived Valence | 0.58 | 0.17 | 0.29 | 0.34 | 0.00035 | 0.3 |
| Felt Arousal | 0.45 | 0.41 | 0.23 | 0.36 | 0.0003 | 0.2 |
| Felt Valence | 0.32 | 0.46 | 0.25 | 0.34 | 0.0003 | 0.2 |
| **NN with Eye-Tracking + Personality** | | | | | | |
| Perceived Arousal | 0.46 | 0.49 | 0.40 | 0.45 | 0.0003 | 0.2 |
| Perceived Valence | 0.57 | 0.33 | 0.29 | 0.40 | 0.0002 | 0.2 |
| Felt Arousal | 0.58 | 0.58 | 0.40 | **0.52** | 0.0002 | 0.2 |
| Felt Valence | 0.38 | 0.57 | 0.28 | 0.41 | 0.0002 | 0.2 |
| **NN with Eye-Tracking + Stimuli Emotion** | | | | | | |
| Perceived Arousal | 0.56 | 0.33 | 0.58 | 0.49 | 0.0002 | 0.3 |
| Perceived Valence | 0.77 | 0.56 | 0.91 | 0.75 | 0.0002 | 0.3 |
| Felt Arousal | 0.47 | 0.45 | 0.29 | 0.40 | 0.0002 | 0.2 |
| Felt Valence | 0.50 | 0.51 | 0.54 | **0.52** | 0.0003 | 0.3 |
| **NN with Eye-Tracking + Personality + Stimuli** | | | | | | |
| Perceived Arousal | 0.63 | 0.48 | 0.65 | **0.59** | 0.0007 | 0.3 |
| Perceived Valence | 0.77 | 0.63 | 0.90 | **0.77** | 0.0007 | 0.3 |
| Felt Arousal | 0.61 | 0.53 | 0.48 | **0.54** | 0.0004 | 0.3 |
| Felt Valence | 0.53 | 0.62 | 0.60 | **0.58** | 0.0007 | 0.3 |
| **SVM with Stimuli Emotion** | | | | | | |
| Perceived Arousal | 0.57 | 0.26 | 0.61 | 0.48 | N/A | N/A |
| Perceived Valence | 0.76 | 0.52 | 0.92 | 0.73 | N/A | N/A |
| Felt Arousal | 0.48 | 0.28 | 0.32 | 0.36 | N/A | N/A |
| Felt Valence | 0.50 | 0.36 | 0.59 | 0.48 | N/A | N/A |

**Baseline limitation.** Our SVM baseline used aggregated features without temporal gaze dynamics. A stricter comparison with identical feature sets would further strengthen causal attribution of performance gains, but was beyond the scope of this paper.

Although this comparison demonstrates the added value of temporal modeling, a full ablationtraining both NN and SVM on identical feature setswas beyond the scope of this paper but remains a clear next step for strengthening causal attribution of performance gains.

### 4.6 Results

Table 4 shows the F1-scores for different models and emotion labels, with the best results for each label **bolded**.

Integrating personality traits, temporal eye-tracking data, and stimulus emotion notably boosted performance, especially for **felt emotions**. This finding suggests that subjective felt experiences profit most from incorporating high-level personality data. The SVM baselines performed well on perceived emotions, possibly reflecting direct stimulus influence, but they could not model sequential dependencies.

## 5 Discussion

We combined eye-tracking data, personality traits, and stimulus-emotion labels to enhance emotion recognition in short speech-containing clips. Although not a fully interactive setup, emphasizing the listeners perspective under controlled conditions allowed us to isolate key predictors of perceived and felt emotions.

## 5.1 Complexity and Agreement

Emotions emerge from the interplay of stimuli, individual traits, and context, making them challenging to model. Table 5 shows user agreement ranging from 56.0% (felt arousal) to 77.7% (perceived valence). Personalized calibration could help address subjective variability. Moreover, as gaze patterns vary by age and gender, our demographic imbalance may introduce bias and limit generalizability.

## 5.2 Model Performance and Stimulus Emotion

The inclusion of stimulus emotion as an input is motivated by its role as a contextual prior: it represents the actors intended expression ($E_e$), which observers can perceive ($E_p$) and relate to their own felt states ($E_f$). This is not equivalent to ground truth but serves as a contextual feature that participants explicitly rated against. We deliberately did not use raw audiovisual features from the video, as our goal was to focus on observer-centric signals (gaze, personality) rather than re-training an audiovisual recognition system on acted data already well-studied in prior work.

Our best model attained a macro F1 of **0.77** for perceived valence (Table 4). Stimulus emotion greatly aided perceived-emotion prediction; an SVM with only stimulus emotion was already strong. However, the NN outperformed it on felt emotions by integrating physiological cues from eye tracking. The fact that perceived valence is easier mirrors findings that observers rely on learned emotion concepts (BET-like), whereas felt states reflect constructionist variability [3, 13].

## 5.3 Multimodality and Individual Differences

Combining stimulus emotion, personality, and eye-tracking gave the highest macro F1 scores (0.77 for perceived valence, 0.58 for felt valence). Personality clarified individual tendencies, eye tracking offered real-time physiological measures [39], and stimulus emotion contextualized perception. This pattern aligns with a layered framework of affect: physiological arousal signals (e.g., pupil dilation) guide attentional deployment (gaze patterns), which in turn feed into conceptual emotion labeling in the observers mind [41].

## 5.4 Comparison and Future Directions

While our NN surpassed the SVM baseline for felt emotionslargely due to sequential informationwe intentionally kept the LSTM architecture compact to reduce overfitting risk given the dataset size and short sequence lengths. Future work could evaluate transformer-based multimodal architectures or cross-modal attention mechanisms, which may better exploit multimodal dependencies. This decision was intentional: the dataset size and the relatively short temporal sequences favored compact architectures with fewer parameters, reducing the risk of overfitting. Future work could evaluate transformer-based multimodal architectures or cross-modal attention mechanisms, which may exploit richer interdependencies. A thorough ablation would align features for both models. Additional future work includes improving sample representativeness, investigating true two-way interactions, and refining interpretability via attention weighting or feature ablation. Real-world applications (e.g., telehealth or adaptive tutoring) must also respect data privacy and informed consent. The reliance on the acted CREMA-D dataset limits ecological validity, and future research should

Table 5: User agreement (%).

| | Felt Arousal | Felt Valence | Perceived Arousal | Perceived Valence |
|---|---|---|---|---|
| Agreement | 56.0 | 65.9 | 60.6 | 77.7 |

pursue more spontaneous, diverse stimuli.

# 6 Conclusion

Grounded in a layered affect framework bridging Basic Emotion Theory and constructionist accounts, this study shows that integrating *temporal eye-tracking*, *personality traits*, and *stimulus emotion* improves recognition of both *perceived* and *felt* emotions in speech-based clips. By emphasizing the *listeners perspective*, we found that personality traits enhanced felt predictions (e.g., felt arousal rose from 0.36 to 0.52), while stimulus emotion strongly supported perceived performance (perceived valence from 0.34 to 0.77). Separating core-affect dynamics from conceptual labeling proved valuable for modeling both.

The implications are twofold. First, unifying physiological signals (pupil, gaze), attentional strategies (fixations), and contextual traits (personality, stimulus cues) yields a richer, more individualized account of emotion, with applications in adaptive agents, teleconferencing, and mental health. Second, the findings support a layered theoretical view in which physiological fluctuations, attentional deployment, and conceptual knowledge jointly shape emotional construction.

Limitations remain: reliance on acted CREMA-D clips constrains ecological validity; the young, male-skewed sample limits generalizability; and discretizing continuous ratings into three bins stabilized training but reduced nuance. Addressing these issues will require larger, more diverse samples, spontaneous dialogue data, and models handling continuous or ordinal ratings. Expanding beyond eye tracking to prosody or micro-expressions could further strengthen ecological validity.

In outlook, effective recognition systems must explicitly separateand then reintegratecore-affect signals, attention strategies, and conceptual constructs. Such systems can move beyond static, actor-driven benchmarks toward interactive, real-time, and ethically responsible applications. We envision layered approaches enabling adaptive, privacy-conscious affective computing that better captures the complexity and subjectivity of human emotion.

# 7 Ethical Impact Statement

This research investigates emotion detection in dialogues by integrating eye-tracking data, temporal dynamics, and personality traits. As the study involves human participants, it was conducted with oversight from an ethical review board. Informed consent was obtained from all participants, clarifying data collection, usage, and analysis. All data were anonymized, and participants were informed of their right to withdraw at any time without consequence.

Potential risks include privacy concerns related to emotion recognition, especially for metrics like pupil size that participants cannot consciously control. Unlike facial or vocal expressions, eye metrics such as pupil dilation have minimal cultural awareness, raising the risk of unintentionally revealing emotional states. We addressed these concerns by anonymizing data, restricting data access to authorized personnel, and clearly explaining data usage to participants.

The anonymization of data, explicit communication of its purpose, and careful ethical handling are key mitigation strategies. Our findings may enable more sensitive, context-aware affective computing applications that respect user privacy while advancing the field of emotion recognition in a safe, ethically responsible manner.

# References


[1] Tagduda Ait Challal and Ouriel Grynszpan. What gaze tells us about personality. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 129–137, 2018.

[2] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6 (2):20, 2016.

[3] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

[4] C. Felipe Benitez-Quiroz, Ronnie B. Wilbur, and Aleix M. Martinez. Improving the measurement of emotional responses with fine-grained likert scales. *Emotion Review*, 14(1):26–36, 2022.

[5] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[6] Laurence Chaby, Isabelle Hupont, Marie Avril, Viviane Luherne-du Boullay, and Mohamed Chetouani. Gaze behavior consistency among older and younger adults when looking at emotional faces. *Frontiers in Psychology*, 8:548, 2017.

[7] Li Chen, Wanling Cai, Dongning Yan, and Shlomo Berkovsky. Eye-tracking-based personality prediction with recommendation interfaces. *User Modeling and User-Adapted Interaction*, 33 (1):121–157, March 2023. ISSN 0924-1868, 1573-1391. doi: 10.1007/s11257-022-09336-9. URL https://link.springer.com/10.1007/s11257-022-09336-9.

[8] Ling Chen, Xiqin Liu, Xiangrun Weng, Mingzhu Huang, Yuhan Weng, Haoran Zeng, Yifan Li, Danna Zheng, and Caiqi Chen. The emotion regulation mechanism in neurotic individuals: The potential role of mindfulness and cognitive bias. *International Journal of Environmental Research and Public Health*, 20(2):896, 2023.

[9] Rui Chen and Qing Liu. esee-d: Emotional state estimation based on eye-tracking dataset. *ArXiv Preprint*, arXiv:2403.11590, 2024.

[10] Paul T Costa and Robert R McCrae. Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people. *Journal of personality and social psychology*, 38(4):668, 1980.

[11] Antoine Coutrot, Nicola Binetti, Charlotte Harrison, Isabelle Mareschal, and Alan Johnston. Face exploration dynamics differentiate men and women. *Journal of vision*, 16(14):16–16, 2016.

[12] Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909, 2017.

[13] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[14] Linda Fiorini, Francesco Bossi, and Francesco Di Gruttola. Eeg-based emotional valence and emotion regulation classification: a data-centric and explainable approach. *Scientific Reports*, 14 (1):24046, 2024.

[15] Andrea Fossati, Serena Borroni, Donatella Marchione, and Cesare Maffei. The big five inventory (bfi). *European Journal of Psychological Assessment*, 2011.

[16] Nikhil Garg, Rohit Garg, Apoorv Anand, and Veeky Baths. Decoding the neural signatures of valence and arousal from portable eeg headset. *Frontiers in Human Neuroscience*, 16:1051463, 2022.

[17] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional lstm networks for improved phoneme classification and recognition. In *International conference on artificial neural networks*, pages 799–804. Springer, 2005.

[18] Brian W Haas, R Todd Constable, and Turhan Canli. Stop the sadness: Neuroticism is associated with sustained medial prefrontal cortex response to emotional facial expressions. *Neuroimage*, 42(1):385–392, 2008.

[19] Joseph Heffner and Oriel FeldmanHall. A probabilistic map of emotional experiences during competitive social interactions. *Nature communications*, 13(1):1718, 2022.

[20] David J Hughes, Ioannis K Kratsiotis, Karen Niven, and David Holman. Personality traits and emotion regulation: A targeted review and recommendations. *Emotion*, 20(1):63, 2020.

[21] Oliver P John, Eileen M Donahue, and Robert L Kentle. Big five inventory. *Journal of personality and social psychology*, 1991.

[22] Kai Kaspar and Peter König. Emotions and personality traits as high-level factors in visual attention: a review. *Frontiers in human neuroscience*, 6:321, 2012.

[23] Elizabeth G Kehoe, John M Toomey, Joshua H Balsters, and Arun LW Bokde. Personality modulates the effects of emotional arousal and valence on brain activation. *Social cognitive and affective neuroscience*, 7(7):858–870, 2012.

[24] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Andreas Papaioannou, Guoying Zhao, Björn Schuller, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 10790–10800, 2022.

[25] Peter J. Lang, Margaret M. Bradley, and Bruce N. Cuthbert. Affective norms for english words (anew): Affective ratings of words and instructions for use. *Behavior Research Methods*, 51(4): 1246–1265, 2019.

[26] Xuan Li and Yan Chen. Emotion recognition using different sensors, emotion models, methods, and datasets: A comprehensive review. *Frontiers in Neuroergonomics*, 5(1338243), 2024.

[27] Jing Liang, Yu-Qing Zou, Si-Yi Liang, Yu-Wei Wu, and Wen-Jing Yan. Emotional gaze: The effects of gaze direction on the perception of facial emotions. *Frontiers in psychology*, 12:684357, 2021.

[28] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining eye movements and eeg to enhance emotion recognition. In *IJCAI*, volume 15, pages 1170–1176. Buenos Aires, 2015.

[29] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14:261–292, 1996.

[30] Gelareh Mohammadi and Patrik Vuilleumier. A multi-componential approach to emotion recognition and the effect of personality. *IEEE Transactions on Affective Computing*, 13(3):11271139, July 2022. ISSN 2371-9850. doi: 10.1109/taffc.2020.3028109. URL http://dx.doi.org/10.1109/TAFFC.2020.3028109.

[31] Sungjoon Park and Jihoon Lee. Modality effects on emotion perception in english by chinese l2 english users: An eye-tracking study. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

[32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. *Scikit-learn: Machine learning in Python*, 2011.

[33] John F Rauthmann, Christian T Seubert, Pierre Sachse, and Marco R Furtner. Eyes as windows to the soul: Gazing behavior is related to personality. *Journal of Research in Personality*, 46(2): 147–156, 2012.

[34] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[35] MW Schurgin, J Nelson, S Iida, H Ohira, JY Chiao, and SL Franconeri. Eye movements during emotion recognition in faces. *Journal of vision*, 14(13):14–14, 2014.

[36] Meisam Jamshidi Seikavandi, Maria Jung Barrett, and Paolo Burelli. Modeling face emotion perception from naturalistic face viewing: Insights from fixational events and gaze strategies. In *Recent Advances in Deep Learning Applications: New Techniques and Practical Examples*. Taylor & Francis, 2025.

[37] Ingo Siegert, Ronald Böck, Bogdan Vlasenko, David Philippou-Hübner, and Andreas Wendemuth. Appropriate emotional labelling of non-acted speech using basic emotions, geneva emotion wheel and self assessment manikins. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2011.

[38] Vasileios Skaramagkas, Giorgos Giannakakis, Emmanouil Ktistakis, Dimitris Manousos, Ioannis Karatzanis, Nikolaos Tachos, Evanthia Tripoliti, Kostas Marias, Dimitrios I. Fotiadis, and Manolis Tsiknakis. Review of Eye Tracking Metrics Involved in Emotional and Cognitive Processes. *Ieee Reviews in Biomedical Engineering*, 16:260–277, 2023. ISSN 1937-3333, 1941-1189. doi: 10.1109/RBME.2021.3066072. URL https://ieeexplore.ieee.org/document/9380366/.

[39] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz Jan Rak. Eye-tracking analysis for emotion recognition. *Computational intelligence and neuroscience*, 2020, 2020.

[40] Stefania Victorita Vacaru, Theodore EA Waters, and Sabine Hunnius. Attachment is in the eye of the beholder: a pupillometry study on emotion processing. *Scientific reports*, 15(1):8015, 2025.

[41] Karlijn Van Heijst, Mariska E Kret, and Annemie Ploeger. Basic emotions or constructed emotions: Insights from taking an evolutionary perspective. *Perspectives on Psychological Science*, 20(3):377–391, 2025.

[42] Yu Wang and Shiyu Gao. Emotion recognition in adaptive virtual reality settings: Challenges and opportunities. *IEEE Transactions on Affective Computing*, 2023.

[43] Alex J Zautra, Glenn G Affleck, Howard Tennen, John W Reich, and Mary C Davis. Dynamic approaches to emotions and stress in everyday life: Bolger and zuckerman reloaded with positive as well as negative affects. *Journal of personality*, 73(6):1511–1538, 2005.

[44] Liang Zhang and Minghua Wang. Survey of deep emotion recognition in dynamic data using facial, speech, and textual cues. *Frontiers in Psychology*, 14:10978716, 2023.