# Multi-Scale Spatiotemporal EEG and Self-Supervised Audio Fusion: A Mixture-of-Experts Approach to Continuous Affect

Ali Amini
Centre of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania
ali.amini@ktu.edu

Sarmad Maqsood
Centre of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania
sarmad.maqsood@ktu.lt

Irfan Abbas
Centre of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania
irfan.abbas@ktu.edu

Muhammad Abdullah Sarwar
Centre of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania
m.sarwar@ktu.edu

Rytis Maskeliūnas
Centre of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania
rytis.maskeliunas@ktu.lt

Robertas Damaševičius
Centre of Real Time Computer Systems
Kaunas University of Technology
Kaunas, Lithuania
robertas.damasevicius@ktu.lt

## Abstract

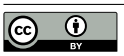
Emotion-aware human–computer interaction increasingly relies on continuous emotion recognition (CER) to track affective states over time. This paper investigates continuous valence prediction on the MAHNOB-HCI database using a multimodal EEG+audio framework. The proposed model combines (i) a hierarchical spatiotemporal EEG encoder that captures per-channel temporal dynamics and inter-channel dependencies and (ii) self-supervised audio embeddings extracted from SSLAM in a feature-based transfer setting. To integrate modalities robustly, we introduce a modality-sensitive mixture-of-experts (MoE) fusion module with modality-specific experts and a learned gating network that dynamically weights EEG and audio contributions. Under session-preserving partitioning with 5-fold cross-validation, the proposed approach achieves strong agreement with ground-truth valence trajectories (PCC=0.778, CCC=0.742, RMSE=0.034), comparing favorably with representative prior results reported under broadly related MAHNOB-HCI protocols. Finally, we provide interpretability evidence by linking the EEG encoder's alpha-band saliency asymmetry to classical alpha-band power asymmetry patterns, thereby supporting the neurophysiological plausibility of the learned representations.

## CCS Concepts

• **Computing methodologies** → **Neural networks**; • **Human-centered computing** → *Human computer interaction (HCI).*

## Keywords

Electroencephalography (EEG), Continuous emotion recognition, Multimodal fusion, Mixture of experts, Transfer learning.



This work is licensed under a Creative Commons Attribution 4.0 International License.
*ICMI '26, Napoli, Italy*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2318-6/26/10
https://doi.org/10.1145/3776574.3831113

**ACM Reference Format:**
Ali Amini, Sarmad Maqsood, Irfan Abbas, Muhammad Abdullah Sarwar, Rytis Maskeliūnas, and Robertas Damaševičius. 2026. Multi-Scale Spatiotemporal EEG and Self-Supervised Audio Fusion: A Mixture-of-Experts Approach to Continuous Affect. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3776574.3831113

## 1 Introduction

Emotion-aware human-computer interaction (HCI), affective multimedia retrieval, and mental health monitoring all depend on the ability to automatically infer the user's emotional state from observable and physiological signals. Recent survey results suggest that the recognition of emotions is now at the forefront of biomedical engineering, intelligent interfaces, as well as BCI/brain-computer interfaces because of the increase in application opportunities such as driver monitoring, adaptive learning, and personalized media recommendations [19, 22]. In contemporary affective computing, emotions are typically modeled either as discrete categories or as a low-dimensional continuous space, most commonly along valence and arousal axes, with some frameworks also including dominance [16, 19]. Continuous emotion recognition (CER) attempts to provide the evolving affective state of an individual over time, which is typically expressed as a real-valued trajectory in valence that is synchronized with the physiological and/or audiovisual data streams [11, 13]. CER provides a framework for HCI in the real world by identifying both gradual mood shifts and moment-to-moment emotional states in a dynamic multimedia environment, rather than assigning a single label to each stimulus [11, 13].

Electroencephalography (EEG) is one of the best methods to measure emotion, as it quantifies brain electrical activity with high timing accuracy, and is less likely than either facial expression or voice to be purposely hidden from others [19]. In the last ten years, there has been considerable development in EEG techniques for emotion recognition, including traditional approaches that use time-frequency features with machine learning and more modern techniques that use deep neural networks (DNNs) [19, 22]. For example, [8] showed that deep-learning-derived features from EEG

rhythms can achieve strong emotion classification performance, emphasizing the importance of learned representations over handcrafted features.

Although EEG measures the brain's internal stimulus-related response, it does not reflect the context in which this may be communicated [14]. Emotions are expressed not only through physiological responses but also acoustically, through both speech and non-speech sounds, including musical and environmental sources [23]. Today, most speech emotion recognition (SER) systems use DNNs trained on spectro-temporal representations of spoken emotion; however, these systems remain limited by the availability and uneven distribution of emotional training data, and variations influence their performance in speaker and recording conditions [21, 25].

To address the limitations of individual modalities, research in EEG-based BCIs and affective computing has focused on multimodal emotion recognition by combining EEG with peripheral physiological signals (e.g., ECG and EDA), eye movements, and/or facial video to utilize complementary information and improve robustness to noise [1]. Although there are usually improvements made with multimodal systems when compared to unimodal baselines, most multimodal systems concentrate on discrete classification, and often rely primarily on simple fusion methodologies (e.g., feature concatenation, late decision fusion) and not on an autonomous dynamic weighting system that adjusts to the reliability of each modality [25]. Recent continuous-label research additionally emphasizes the importance of employing architectural designs that effectively capture long temporal dependencies [9]. However, most current CER studies continue to rely on EEG data as the primary physiological input for studying emotional states, without incorporating visual and/or audio data. Therefore, even with the progress made thus far, there continues to be a substantial gap at the intersection of CER, EEG, and audio-based modalities [10, 17].

In this work, we propose a Multimodal Audio and EEG architecture for continuous valence recognition on the MAHNOB-HCI database [24], which contains synchronized EEG and audio recordings along with continuous annotations of the valence dimension of people's responses to audio-visual stimuli.

On the audio side, we utilize the SSLAM encoder architecture [2], a self-supervised learning transformer trained on massive volumes of audio mixtures to produce general-purpose audio embeddings, which we will use only as a frozen feature extractor in a feature-based transfer learning setting. The EEG branch has employed a Hierarchical spatio-temporal Encoder to encode both channel-specific temporal dynamics and global spatial interactions, projecting these spatio-temporal feature sets into a compact latent embedding format for fusion across modalities. To maximize the benefits of both audio and EEG data while reducing the effects of noise or incomplete information, an innovative fusion of the two data sources was created. The fusion of these modes occurs in a mixture-of-experts (MoE)-type module, which draws on concepts from recent multimodal fusion models that use MoE architectures [15]. The MoE module is designed so that each modality is represented by expert networks attached to a gating network that determines the weight assigned to each feature representation based on the current input. Concurrent work such as BiMoE has also explored brain-inspired expert routing for EEG-dominant affect recognition, but focuses on EEG and peripheral physiological signals for affective-state classification, whereas our work studies continuous valence regression with EEG and self-supervised audio representations [26].

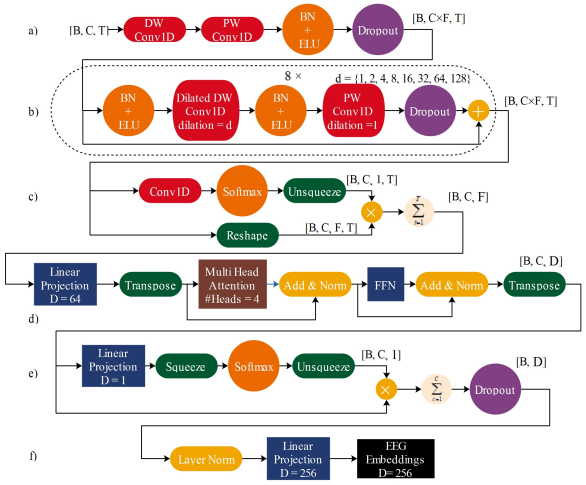



**Figure 1: EEG encoder used in the audio–EEG model. The network first extracts channel-local temporal features, expands the temporal context with a dilated residual stack, and then performs saliency-weighted temporal pooling to retain informative sub-intervals. Cross-electrode structure is introduced via self-attention, followed by learned electrode weighting and a final projection to a $D$=256 embedding for multimodal fusion. Subpanels illustrate: (a) channel-local stem, (b) dilated temporal stack, (c) temporal saliency pooling, (d) inter-channel self-attention, (e) channel attention pooling, and (f) projection. Notation: $C$ channels, temporal length $T$, expansion factor $F$.**

The work presented here has three contributions:

(1) We present a novel framework for continuous valence recognition that jointly leverages self-supervised audio representations and a hierarchical spatiotemporal EEG encoder.
(2) We introduce a mixture-of-experts-inspired fusion architecture tailored to audio–EEG integration. This mechanism enables the model to dynamically emphasize the most reliable modality at any given moment.
(3) We validate the biological plausibility of our model by analyzing how the network's internal "alpha-saliency" relates to frontal alpha asymmetry, a classical neuro-physiological marker of valence [6, 7]. The results of this analysis supported the conclusion that the representations generated by the proposed solution are predictive and consistent with findings from previous investigations of changes in alpha-band hemispheric activity access.

## 2 Methodology

### 2.1 Dataset

Experiments are conducted on MAHNOB-HCI, which provides EEG recordings synchronized with the audio track of emotion-eliciting

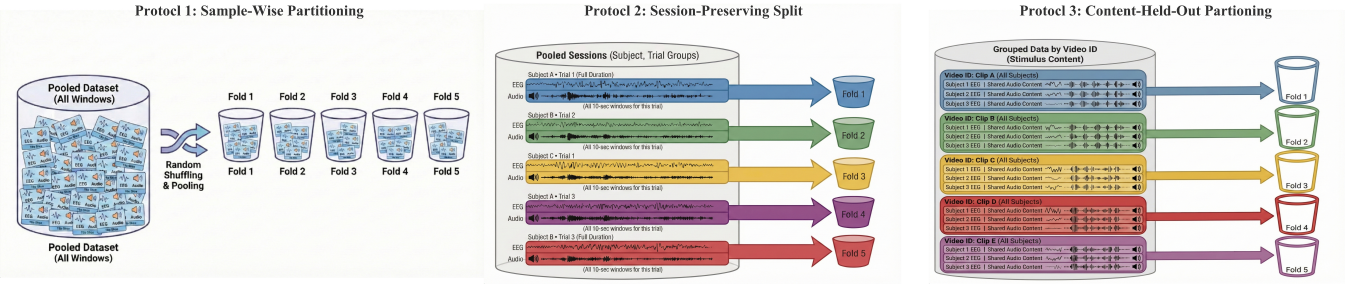



**Figure 2: Illustration of train–test partitioning strategies for audio–EEG 5-fold cross-validation. Sample-Wise Partitioning (left): individual 10-s multimodal segments are shuffled and distributed across folds without regard to the session structure. Session-Preserving Partitioning (center): all Segments derived from a single (Subject, Recording) session are assigned to one fold, preventing temporal overlap between training and evaluation. Content-Held-Out Partitioning (right): folds are constructed by stimulus identity, ensuring that the acoustic content encountered during training is absent from the test set across all participants.**

video stimuli [24]. We use the expert-annotated subset and study continuous valence regression. Valence annotations are available at 4 Hz, and we form the target sequence by averaging the annotators' traces at each time index.

Training samples are generated by synchronizing EEG and audio to the annotation grid and extracting sliding windows of length $W$ seconds (here, $W = 10$). Let $\Delta = 0.25$ s denote the label period. For each time step $t_k$, we extract the multimodal segment spanning the preceding window $[t_k - W, t_k)$ and predict the valence value at $t_k$. This formulation ensures a causal prediction setting, where the affect at the current moment is inferred from the immediate history of neural and acoustic context.

The expert-annotated MAHNOB-HCI subset used here contains 24 subjects and 239 trials spanning 20 video stimuli. Using $W$=10 s windows and $\Delta$=0.25 s stride yields 68,136 multimodal windows in total.

## 2.2 Validation protocols

Figure 2 illustrates the three evaluation protocols considered in this work. All use 5-fold cross-validation but differ in how windows are assigned to folds.

*Sample-wise partitioning:* Individual 10-s windows are shuffled and distributed across folds without regard to session structure. This protocol treats each window as an independent sample and serves as a reference point for comparison with session-aware splitting strategies.

*Session-preserving partitioning:* All windows from the same (subject, trial) session are assigned to a single fold, preventing temporal overlap between training and evaluation. This is our primary protocol; it reflects a realistic deployment scenario where new recordings from known participants are processed without access to the specific trial during training.

*Content-held-out partitioning:* Folds are constructed by stimulus/video identity: all windows associated with a given video clip, across all subjects, are assigned to the same fold. This ensures that acoustic content encountered during training is entirely absent from the test set, probing whether the model generalizes to unseen stimuli rather than memorizing clip-specific valence trajectories.

## 2.3 EEG and Audio Preprocessing

The 32 EEG channels are mapped to a 10–20 montage and cleaned using a standard filter–artifact–segmentation sequence. We apply a 1–45 Hz band-pass filter together with a 50 Hz notch filter, then re-reference to the common average when the montage assignment is valid. Artifact attenuation is performed using ICA; ICLabel is used to tag non-neural components, and components attributed to ocular, muscle, cardiac, or channel noise are removed before reconstructing the signal. Finally, stimulus timing is determined from the Status channel, the recording is cropped to the presentation interval, and the result is downsampled to 128 Hz for window extraction.

The stimulus audio is extracted per trial, mixed down to mono, and resampled to 16 kHz to match the SSLAM frontend.

## 2.4 EEG Model Explanation

Our EEG branch maps each $W$-second window to a compact representation that can be fused with the audio embedding while preserving temporally localized evidence. Let $\mathbf{X} \in \mathbb{R}^{B\times C\times T}$ denote the windowed EEG (batch size $B$, $C$=32 channels, $T$=128 $\cdot$ $W$ samples; $W$ = 10 s in our setup). The encoder outputs an embedding $\mathbf{z}_{\text{eeg}} \in \mathbb{R}^{B\times 256}$. The architecture proceeds in stages: channel-wise temporal filtering and multi-scale integration build per-electrode descriptors; an attention-based temporal pooling step selects salient moments; cross-channel self-attention captures inter-electrode interactions; and a learned channel aggregation followed by a projection head produces the final embedding. Fig. 1 provides an overview.

*Channel-Local Temporal Filtering (Fig. 1a):* The encoder begins by extracting short-range temporal structure within each EEG channel. We apply a depthwise Conv1D along the time dimension (one filter group per electrode) to capture local waveform patterns without mixing information across electrodes. The resulting features are then expanded with a pointwise ($1 \times 1$) convolution to obtain $F$ features per channel (we use $F$ = 16). Batch normalization, ELU,

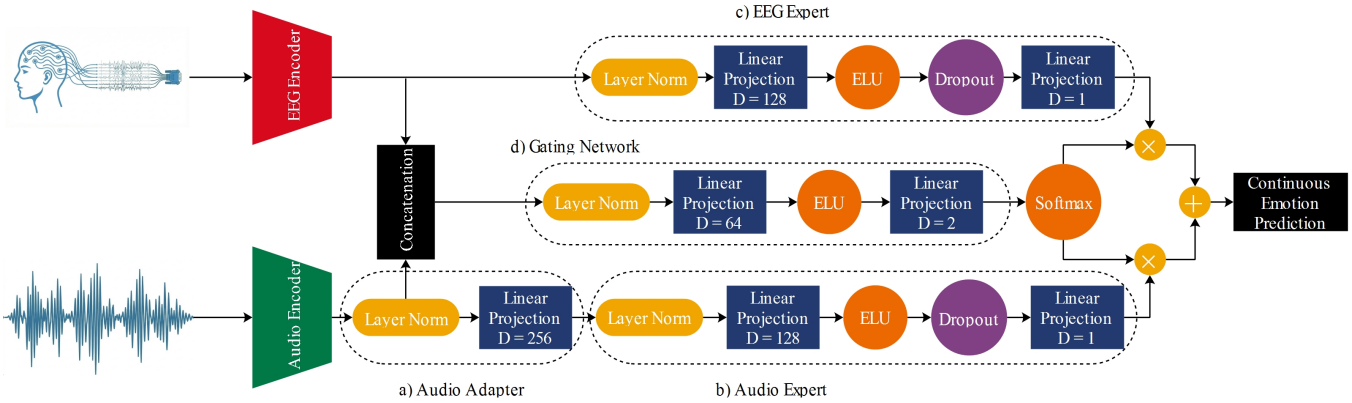



**Figure 3: End-to-end audio–EEG model with gated mixture-of-experts fusion. A frozen SSLAM encoder provides a clip-level audio embedding that is adapted to the EEG latent dimension, while the EEG branch produces a 256-D window representation. Two lightweight experts regress valence from their respective modality embeddings, and a learned gate routes between them on a per-window basis, yielding a single fused prediction.**

and dropout are used for stability, and the stride provides the first reduction of the temporal resolution.

*Multi-Rate Dilated Residual Temporal Stack (Fig. 1b):* Affective cues in EEG can unfold at different rates within a $W$-second window. To model both fast fluctuations and slower trends without resorting to very deep networks, we employ a residual stack of depthwise dilated Conv1D layers. Dilation is increased across successive blocks to progressively enlarge the temporal context. Each block applies normalization and ELU, performs a dilated depthwise convolution over time, mixes features with a pointwise ($1 \times 1$) convolution within each channel group, and uses dropout for regularization before adding the residual. This design achieves wide temporal coverage with modest computational cost.

*Temporal Saliency Pooling (Fig. 1c):* To obtain a fixed-length descriptor per electrode while retaining temporally localized evidence, we replace uniform averaging with a saliency-weighted temporal aggregation. A lightweight grouped Conv1D produces a score sequence for each channel; after softmax normalization along the temporal axis, these scores act as pooling weights. Applying the weighted sum collapses the time dimension, yielding $\mathbf{H} \in \mathbb{R}^{B\times C\times F}$ (with $F = 16$), thereby prioritizing the most informative sub-intervals within the $W$-second window.

*Inter-Channel Self-Attention (Fig. 1d):* After temporal summarization, we model cross-electrode interactions to capture distributed patterns that are not visible in any single channel. We treat the $C$ electrodes as a set of tokens and apply multi-head self-attention over the channel axis. Each electrode descriptor is projected to an attention space ($D = 64$) and processed by a transformer-style block (multi-head attention followed by a position-wise feed-forward subnetwork), with residual connections and layer normalization. The output $\tilde{\mathbf{H}} \in \mathbb{R}^{B\times C\times D}$ encodes spatially informed electrode features that reflect learned inter-channel coupling.

*Channel Attention Pooling (Fig. 1e):* We collapse the electrode set into a single vector using a learned topographic weighting. A small linear scoring layer produces one logit per channel; after softmax normalization across channels, these weights define an attention distribution over electrodes. The global descriptor is obtained as the corresponding weighted sum, yielding $\mathbf{u} \in \mathbb{R}^{B\times D}$. This step produces a compact, window-level EEG summary that is well-suited for downstream audio fusion.

*Projection Head (Fig. 1f):* The pooled vector $\mathbf{u}$ is converted into the final EEG embedding through a lightweight normalization–projection step. We apply LayerNorm and a linear map to obtain $\mathbf{z}_{\text{eeg}} \in \mathbb{R}^{B\times 256}$. This representation defines the EEG-side interface to the multimodal model and is used by the downstream fusion and regression components.

Unless stated otherwise, the EEG encoder uses a depthwise temporal stem (kernel 41, stride 4) followed by an 8-block dilated residual stack with dilations 1–128; dropout is 0.05 in the stem and 0.1 in the temporal stack, and inter-channel attention uses 4 heads with $D$=64.

## 2.5 Audio Model Explanation

For each aligned 10-s audio window, we extract a fixed-length representation using SSLAM, a self-supervised audio representation model designed to improve robustness on *polyphonic* soundscapes by learning from partially mixed audio during pre-training [2]. SSLAM is pre-trained on AS-2M (an AudioSet-derived large-scale corpus) without using label information, making it well-suited for general-purpose audio encoding. This pre-training domain is closely related to the audio in the MAHNOB-HCI stimuli (speech, music, and environmental sounds), motivating its use as our audio feature extractor.

Following SSLAM's pre-training recipe, input waveforms are resampled to 16 kHz and converted to 128-bin mel features using a 25 ms Hann window and a 10 ms hop. We then feed the resulting time–frequency representation to SSLAM's ViT-Base encoder and obtain a clip-level embedding by mean-pooling over time tokens.

The resulting audio embedding is 768-dimensional and serves as the audio modality input to our multimodal model.

## 2.6 Modality-Sensitive MoE Fusion for Valence Regression

For each $W$=10 s window, we fuse neural activity with the concurrently presented stimulus audio to predict a scalar valence value. The EEG encoder yields $z_{\text{eeg}} \in \mathbb{R}^{256}$, while the audio branch provides a clip-level representation from SSLAM, $z_{\text{aud}}^{\text{SSLAM}} \in \mathbb{R}^{768}$, extracted in a feature-based transfer setting. The proposed fusion head combines these embeddings using modality-specific experts and an input-conditioned gate, allowing the contribution of EEG and audio to vary across windows.

*Audio adapter (Fig. 3a):* SSLAM produces a high-dimensional clip descriptor, so we introduce a small projection layer to obtain a 256-D audio vector compatible with the EEG embedding. We first normalize the SSLAM embedding and then apply a linear map:

$$\tilde{z}_{\text{aud}} = \mathbf{W}_{\text{aud}}\,\text{LN}\Big(z_{\text{aud}}^{\text{SSLAM}}\Big) + \mathbf{b}_{\text{aud}} \;\in \mathbb{R}^{256}, \tag{1}$$

where $\text{LN}(\cdot)$ denotes LayerNorm. This adapter provides a compact audio interface for the fusion head while keeping the SSLAM backbone frozen.

*Modality experts (Fig. 3b–c):* We use two lightweight expert regressors, one per modality, to convert embeddings into scalar valence estimates. Each expert is a shallow MLP operating on a LayerNorm-normalized input:

$$\begin{aligned}\hat{y}_m &= f_m(z_m)\\ &= \mathbf{w}_m^{\top}\,\text{Drop}\Big(\text{ELU}\big(\mathbf{W}_m\,\text{LN}(z_m) + \mathbf{b}_m\big)\Big) + b_m^{\text{out}},\\ &\qquad\qquad m \in \{\text{eeg}, \text{aud}\},\end{aligned} \tag{2}$$

where the hidden width is 128. In particular, $z_{\text{aud}} = \tilde{z}_{\text{aud}}$ is the adapted SSLAM embedding, producing $\hat{y}_{\text{aud}}$, and $z_{\text{eeg}}$ produces $\hat{y}_{\text{eeg}}$. Each modality expert is a 1-hidden-layer MLP with width 128, ELU activation, and dropout 0.3; the gate MLP uses a 64-unit hidden layer.

*Gating network (Fig. 3d):* A small gating network determines how much to trust the EEG versus the audio for each window. We form a joint feature by concatenation, $z_f = [z_{\text{eeg}}; \tilde{z}_{\text{aud}}] \in \mathbb{R}^{512}$, and map it to two mixture weights with a two-layer MLP:

$$\begin{aligned}h &= \text{ELU}\Big(\mathbf{W}_g\,\text{LN}(z_f) + \mathbf{b}_g\Big) \in \mathbb{R}^{64},\\ g &= \text{softmax}(\mathbf{W}_o\,h + \mathbf{b}_o) \in \mathbb{R}^{2},\end{aligned} \tag{3}$$

where $g = [g_{\text{eeg}}, g_{\text{aud}}]$ and the hidden width is 64.

*Window-level output:* Given the expert scalars and the gate weights, the model returns a single valence estimate for the window by mixing the modality-specific predictions:

$$\hat{y} = g_{\text{eeg}}\,\hat{y}_{\text{eeg}} + g_{\text{aud}}\,\hat{y}_{\text{aud}}. \tag{4}$$

We evaluate the model at every label timestamp ($\Delta$=0.25 s), so sliding the $W$-second window along the sequence produces a dense 4 Hz prediction stream aligned to the ground-truth annotations.

## 2.7 Training Process

**Cross-validation and splits:** We evaluate with 5-fold cross-validation on multimodal windows aligned to the 4 Hz annotation grid. The three partitioning strategies described in Section 2.2 differ only in how windows are grouped for fold assignment: (i) sample-wise partitioning shuffles windows independently; (ii) session-preserving partitioning groups by (subject, trial) pairs; and (iii) content-held-out partitioning groups by video identity. Session-preserving is our primary protocol.

To reduce fold-to-fold label-distribution imbalance, we perform stratification using quantile bins derived from group-level valence summaries. For each outer fold, one fold is held out for testing; from the remaining folds, we create a validation subset via a grouped, stratified split (20% of training data). Model selection and early stopping use the validation objective only, and we report mean±std over the five test folds.

**Optimization:** Models are trained with AdamW (learning rate $3 \times 10^{-4}$, weight decay $10^{-4}$) and batch size 64. We apply gradient clipping with a maximum $\ell_2$ norm of 1.0 and an early stopping patience of 20. Dropout is used in the EEG encoder and throughout the MoE head to mitigate overfitting. The SSLAM encoder is used in a frozen feature-extraction setting; training updates only the audio adapter and the fusion/regression components.

**Objective and metrics:** We optimize the CCC loss on the fused prediction $\hat{y}$:

$$\rho_c \;=\; \frac{2\rho\,\sigma_{\hat{y}}\,\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2}, \tag{5}$$

and

$$\mathcal{L}_{\text{CCC}} \;=\; 1 - \rho_c. \tag{6}$$

We train for up to 1000 epochs with early stopping (patience 20), use no LR schedule, fix the random seed to 42, and run on a single RTX 4090 GPU. For evaluation, we compute PCC, CCC, and RMSE on each test fold and summarize results as mean±std.

# 3 Results

## 3.1 Comparison with Prior Work

Most MAHNOB-HCI continuous valence studies pair EEG with *visual* information (facial recordings or video descriptors) and rely on attention-style fusion to balance heterogeneous cues. Recent work has used residualized multimodal transformers to stabilize token-level integration across modalities [18], whereas earlier EEG–video pipelines employed attention to modulate each modality's contribution during fusion [4]. Semi-supervised learning has also been explored to exploit additional unlabeled video for continuous affect regression [5]. Complementary to attention-based fusion, approaches that explicitly separate modality-shared and modality-specific factors, for instance via graph-based knowledge distillation, aim to reduce redundancy and improve cross-modal transfer [12]. At the other end of the spectrum, feature-engineered designs (time–frequency and lateralization descriptors) combined with GA-optimized recurrent models emphasize handcrafted representations over end-to-end multimodal learning [20], and vision-only pipelines using ViT transfer learning with temporal modeling (TCN) show that strong pretrained backbones can be competitive even without EEG [3]. While these studies differ in modalities and protocols,

together they provide a useful performance context for MAHNOB-HCI continuous valence prediction and highlight that stimulus audio has been comparatively less emphasized than visual cues in prior CER work.

## 3.2 Evaluation of validation protocols

Table 1 reports performance under the three partitioning strategies illustrated in Fig. 2. Session-preserving partitioning, which groups windows by (subject, trial) pairs, serves as our primary benchmark. Sample-wise partitioning provides a reference point by treating each window independently. Content-held-out partitioning groups by video identity, ensuring that no acoustic content appears in both training and testing; this setting tests generalization to unseen stimuli.

**Table 1: Effect of evaluation protocol on performance (mean±std).**

| Protocol | Description | PCC↑ | CCC↑ | RMSE↓ |
|---|---|---|---|---|
| Session-Preserving | Groups = (subject, trial); prevents temporal overlap | 0.778 ± 0.056 | 0.742 ± 0.045 | 0.034 ± 0.004 |
| Sample-Wise (reference) | Random window shuffle (leaks context) | – | – | 0.003 ± 0.001 |
| Content-Held-Out | Groups = video ID; unseen acoustic content in test | 0.580 ± 0.064 | 0.575 ± 0.064 | 0.061 ± 0.003 |

The drop from the session-preserving setting to the content-held-out setting indicates that part of the predictive signal is stimulus-dependent: while the model generalizes across sessions, unseen video/audio content remains more challenging because affective trajectories can be strongly tied to the eliciting stimulus. We therefore treat the content-held-out result as a stricter estimate of stimulus-level generalization rather than as a failure case.

## 3.3 Performance results

Table 2 lists representative reported results under broadly related MAHNOB-HCI settings alongside our EEG+Audio model, which obtains PCC=0.778 ± 0.056, CCC=0.742 ± 0.045, and RMSE=0.034 ± 0.004. Because the compared studies differ in modality choices, preprocessing pipelines, metrics, and validation details, the table should be interpreted as contextual evidence rather than a strictly controlled head-to-head comparison.

**Table 2: Continuous valence on MAHNOB-HCI (session-preserving 5-fold CV). "–" = not reported.**

| Study | PCC↑ | CCC↑ | Error↓ |
|---|---|---|---|
| **Ours (EEG+Audio)** | **0.778±0.056** | **0.742±0.045** | **0.034±0.004 (RMSE)** |
| [18] | 0.755±0.082 | 0.736±0.078 | 0.051±0.007 (RMSE) |
| [12] | 0.737 | – | 0.049 (MAE) |
| [3] | – | 0.628 | 0.053 (RMSE) |
| [5] | 0.530±0.070 | – | 0.038±0.003 (RMSE) |
| [4] | 0.530±0.070 | – | 0.036±0.003 (RMSE) |

*Sample-wise partitioning.* As a reference point, we also evaluated under sample-wise partitioning. Under this setting, our EEG+Audio model attains RMSE=0.0030±0.0007, while [20] reports RMSE=0.0156±0.0110. Because this protocol differs from our primary session-preserving evaluation, these values are provided for continuity with earlier work.

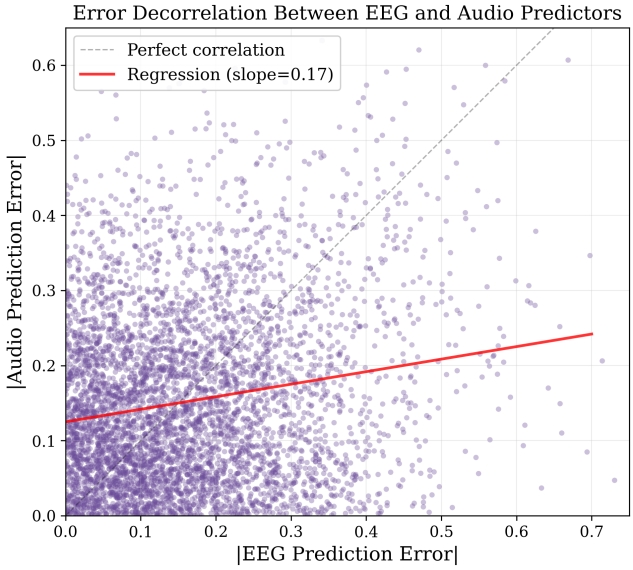



**Figure 4: Per-window absolute prediction error correlation between EEG-only and Audio-only models ($\rho$ = 0.16, $R^2$ = 0.026).**

*EEG-only ablation:* To quantify what the EEG branch can achieve on its own, we evaluate an EEG-only ablation under the same session-preserving partitioning used throughout this section. Table 3 contrasts our EEG-only results with representative EEG-based baselines reported for MAHNOB-HCI. Using the same encoder backbone but removing the audio pathway, the model still achieves strong agreement with the valence trajectory, supporting the view that the proposed temporal hierarchy and cross-electrode modeling provide an effective EEG representation for continuous regression.

**Table 3: EEG-only valence regression on MAHNOB-HCI under the session-preserving partitioning (5-fold CV; mean±std).**

| Study | PCC↑ | CCC↑ | RMSE↓ |
|---|---|---|---|
| **Ours (EEG-only)** | **0.511±0.023** | **0.502±0.026** | **0.061±0.005** |
| [5] | 0.290±0.080 | – | – |
| [18] | 0.492±0.135 | 0.450±0.127 | 0.077±0.014 |

## 3.4 Architecture ablation

Table 4 presents the main ablation study under the session-preserving partitioning 5-fold protocol to isolate the contribution of each design choice in the proposed multimodal system. It compares unimodal baselines (EEG-only and audio-only) against progressively stronger fusion variants, including simple feature concatenation with an MLP, equal-weight late fusion, and the proposed gated two-expert MoE head, while also probing the impact of key EEG encoder components such as temporal saliency pooling and inter-channel self-attention. By reporting PCC, CCC, and error metrics for each configuration, the table clarifies whether the observed gains stem primarily from the self-supervised audio features, the

**Table 4: Ablation results under the session-preserving 5-fold protocol (mean±std).**

| Model variant | PCC↑ | CCC↑ | RMSE↓ |
|---|---|---|---|
| EEG-only (proposed) | 0.511±0.023 | 0.502±0.026 | 0.061±0.005 |
| Audio-only (SSLAM) | 0.557±0.031 | 0.533±0.028 | 0.063±0.005 |
| EEG+Audio concat | 0.697±0.051 | 0.671±0.046 | 0.044±0.006 |
| Late fusion (equal) | 0.738±0.039 | 0.709±0.036 | 0.040±0.004 |
| **Gated (MoE)** | **0.778±0.056** | **0.742±0.045** | **0.034±0.004** |
| MoE w/o saliency | 0.724±0.052 | 0.695±0.047 | 0.044±0.006 |
| MoE w/o self-attn | 0.733±0.044 | 0.701±0.039 | 0.045±0.005 |

hierarchical spatiotemporal EEG representation, or the dynamic gating mechanism that adaptively reweights the contributions of modalities across windows.

To characterize the relationship between modality-specific predictions, we computed the Pearson correlation between per-window absolute errors from the EEG-only and Audio-only models across all held-out test windows. The resulting correlation is low ($\rho = 0.16$, 95% CI: [0.15, 0.17]; Fig. 4), with a coefficient of determination $R^2 = 0.026$, indicating that only 2.6% of the error variance is shared between modalities. The regression slope ($\beta = 0.17$) is substantially flatter than the identity line, confirming that large errors in one modality do not reliably predict large errors in the other. This low shared variance implies that when EEG-based prediction is inaccurate, due to artifacts, inter-subject variability, or ambiguous neural patterns, the audio-based prediction often remains informative, and vice versa. For near-independent predictors, the optimal combination yields a fused error variance lower than either constituent, consistent with classical results on predictor aggregation. The gated fusion mechanism approximates this behavior by adaptively emphasizing the more reliable modality on a per-window basis.

Importantly, the improvement of the gated MoE over both equal late fusion and feature concatenation suggests that the learned gate does not merely rescale a dominant modality, but exploits window-specific disagreement between the EEG and audio experts to arbitrate their contributions.

## 3.5 Neurophysiological Interpretability via Alpha Asymmetry

We assess whether the proposed EEG encoder relies on neurophysiologically plausible *alpha asymmetry* patterns by comparing model-derived alpha-band saliency with classical alpha power lateralization, a well-established marker linked to valence and approach/withdrawal tendencies [6, 7]. For each test window $n$ and channel $c$, we compute a gradient×input saliency signal

$$s_c(n,t) \;=\; x_c(n,t)\,\frac{\partial \hat{y}(n)}{\partial x_c(n,t)}, \tag{7}$$

where $x_c(n,t)$ is the EEG input and $\hat{y}(n)$ is the predicted valence.

To isolate alpha-band contributions, we project saliency onto 8–13 Hz using the magnitude spectrum

$$S_c^{\alpha}(n) \;=\; \sum_{f\in\mathcal{F}_\alpha} \left|\mathcal{F}\{s_c(n,t)\}_f\right|, \tag{8}$$

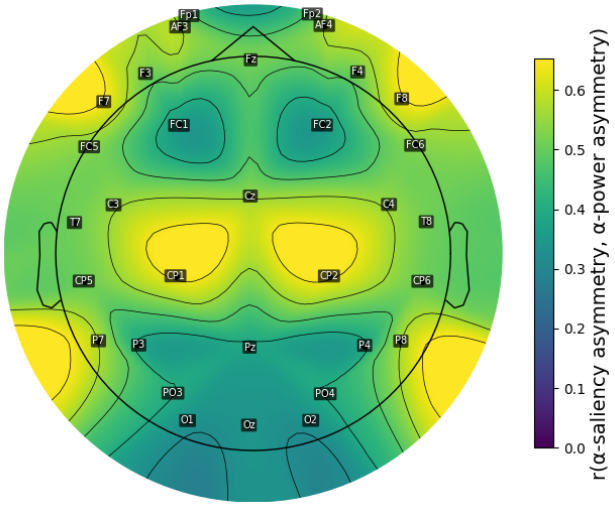



**Figure 5: Topographic summary of alpha-band interpretability. The EEG encoder's 8–13 Hz saliency asymmetry shows a structured left–right lateralization pattern and covaries with conventional alpha power asymmetry across homologous electrode pairs.**

and compute classical alpha power from the raw EEG via Welch PSD,

$$P_c^{\alpha}(n) \;=\; \sum_{f\in\mathcal{F}_\alpha} \mathrm{PSD}(x_c(n,t))_f\,. \tag{9}$$

Here, $\mathcal{F}\{\cdot\}$ denotes the discrete Fourier transform (DFT) over $t$, evaluated on discrete frequency bins $f$, while $\mathrm{PSD}(\cdot)$ denotes the power spectral density estimated via Welch's method on the same frequency axis.

For each left–right homologous pair $(L,R)$, we define normalized asymmetry indices

$$\begin{aligned} A^{\alpha}_{\mathrm{sal}}(n;L,R) \;&=\; \frac{S_L^{\alpha}(n)-S_R^{\alpha}(n)}{S_L^{\alpha}(n)+S_R^{\alpha}(n)+\varepsilon}, \\ A^{\alpha}_{\mathrm{pow}}(n;L,R) \;&=\; \frac{P_L^{\alpha}(n)-P_R^{\alpha}(n)}{P_L^{\alpha}(n)+P_R^{\alpha}(n)+\varepsilon}, \end{aligned} \tag{10}$$

with $\varepsilon > 0$ preventing division by zero. Although $S_c^{\alpha}$ and $P_c^{\alpha}$ are obtained from different spectral estimators, we compare their *normalized left–right asymmetry indices*, which reduces sensitivity to absolute scaling.

Table 5 reports electrode-pair–level effect sizes for alpha-band interpretability by listing, for each left–right homologous pair, the mean saliency asymmetry, mean alpha power asymmetry, and their correlation, thereby quantifying how the model's 8–13 Hz sensitivity lateralization covaries with classical alpha asymmetry patterns.

Across all windows, frontal and fronto-central pairs exhibit a consistent *leftward* saliency asymmetry (e.g., F7/F8 and FC5/FC6 show positive mean $A^{\alpha}_{\mathrm{sal}}$), whereas posterior pairs tend to show *rightward* asymmetry, yielding a coherent lateralization topography. Model-based asymmetry tracks the signal statistics of alpha power:

**Table 5: Alpha-band asymmetry correspondence between model saliency and signal alpha power across homologous electrode pairs.**

| Pair (L/R) | $\mathbb{E}[A^{\alpha}_{\text{sal}}]$ | $\mathbb{E}[A^{\alpha}_{\text{pow}}]$ | $\rho$ |
|---|---|---|---|
| F7/F8 | +0.156 | −0.016 | +0.634 |
| FC5/FC6 | +0.156 | −0.020 | +0.531 |
| T7/T8 | +0.150 | −0.042 | +0.514 |
| P7/P8 | +0.021 | +0.076 | +0.558 |
| O1/O2 | −0.090 | +0.025 | +0.341 |

for each pair, we compute

$$\rho_{L,R} \;=\; \operatorname{corr}\Big(A^{\alpha}_{\text{sal}}(\cdot;L,R),\; A^{\alpha}_{\text{pow}}(\cdot;L,R)\Big), \tag{11}$$

and observe moderate-to-strong positive correlations for several frontal pairs (up to $\rho \approx 0.63$), indicating that when alpha power becomes more left- (or right-) dominant, the model's alpha-band saliency tends to shift in the same direction. Since the analysis spans all samples, corresponding hypothesis tests yield extremely small $p$-values; therefore, we emphasize effect sizes (correlations and mean asymmetry) rather than reporting significance levels.

Finally, stratifying windows into low/medium/high valence tertiles reveals that frontal $A^{\alpha}_{\text{sal}}$ tends to increase with valence (more leftward saliency at higher valence), consistent with classical frontal alpha asymmetry findings. Fig. 5 summarizes the spatial distribution of the alpha asymmetry effects. The correspondence is reported in terms of normalized asymmetry indices (rather than absolute power), providing evidence that the model's input sensitivities reflect physiologically meaningful hemispheric differences in the alpha band.

## 4 Conclusion & Future Work

We presented an audio–EEG approach for continuous valence regression on MAHNOB-HCI that pairs a hierarchical spatiotemporal EEG encoder with frozen self-supervised audio embeddings from SSLAM. The two streams are combined through a gated mixture-of-experts head, enabling reweighting of modality contributions when either EEG or audio is less informative. Under session-preserving partitioning with 5-fold cross-validation, the model achieves PCC=$0.778 \pm 0.056$, CCC=$0.742 \pm 0.045$, and RMSE=$0.034 \pm 0.004$; under content-held-out partitioning, which blocks stimulus identity across folds, the model attains PCC=$0.580 \pm 0.064$, CCC=$0.575 \pm 0.064$, and RMSE=$0.061 \pm 0.003$, providing a more conservative estimate of generalization to unseen stimulus content. In addition, we examined interpretability and found that alpha-band saliency exhibits structured hemispheric asymmetry consistent with classical frontal alpha asymmetry, supporting the neurophysiological plausibility of the learned EEG representation.

A remaining limitation is that the evaluation is restricted to the expert-annotated MAHNOB-HCI subset; consequently, any broader dataset-level generalization should be interpreted with caution until the approach is validated on additional affective EEG corpora.

Future work will expand this audio–EEG setting along two axes. First, we will broaden the input space by adding complementary signals, such as stimulus video and peripheral physiology, enabling tri-modal (EEG+audio+video) and multi-biosignal variants that can better separate stimulus-driven cues from internal responses. Second, we will explore richer routing mechanisms for fusion, including multiple experts per modality, hierarchical gates that operate at both window and sub-window scales, and uncertainty-aware weighting to handle modality dropouts and variable noise while retaining interpretability.

## Safe and Responsible Innovation Statement

This work uses an existing affective computing dataset and does not introduce any new human-subject data collection. The potential deployment of continuous emotion prediction systems should be approached with caution because affect estimates can be noisy, context-dependent, and sensitive. Responsible use should require informed consent, privacy-preserving data handling, transparency about model limitations, and safeguards against the use of physiological or affective predictions for surveillance, profiling, or high-stakes decisions without human oversight.

## Funding

Ali Amini is supported by a research grant received from Horizon Europe's Marie Sklodowska-Curie Action (Grant No. 101118964).

## References

[1] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends* 2, 01 (2021), 73–79.

[2] Tony Alex, Sara Ahmed, Armin Mustafa, Muhammad Awais, and Philip JB Jackson. 2025. SSLAM: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes. *arXiv preprint arXiv:2506.12222* (2025).

[3] Zhanghao Chen and Minmin Miao. 2024. Visual Modality Continuous Emotion Recognition Based on Vision Transformer and Transfer Learning. In *2024 16th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. IEEE, 34–38.

[4] Dong Yoon Choi, Deok-Hwan Kim, and Byung Cheol Song. 2020. Multimodal attention network for continuous-time emotion recognition using video and EEG signals. *IEEE Access* 8 (2020), 203814–203826.

[5] Dong Yoon Choi and Byung Cheol Song. 2020. Semi-supervised learning for facial expression-based emotion recognition in the continuous domain. *Multimedia Tools and Applications* 79, 37 (2020), 28169–28187.

[6] James A Coan and John JB Allen. 2004. Frontal EEG asymmetry as a moderator and mediator of emotion. *Biological psychology* 67, 1-2 (2004), 7–50.

[7] Richard J Davidson. 1992. Anterior cerebral asymmetry and the nature of emotion. *Brain and cognition* 20, 1 (1992), 125–151.

[8] Fatih Demir, Nebras Sobahi, Siuly Siuly, and Abdulkadir Sengur. 2021. Exploring deep learning features for automatic classification of human emotion using EEG rhythms. *IEEE Sensors Journal* 21, 13 (2021), 14923–14930.

[9] Yi Ding, Chengxuan Tong, Shuailei Zhang, Muyun Jiang, Yong Li, Kevin JunLiang Lim, and Cuntai Guan. 2025. EmT: A novel transformer for generalized cross-subject EEG emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2025).

[10] Yi Ding, Su Zhang, Chuangao Tang, and Cuntai Guan. 2024. Masa-tcn: Multi-anchor space-aware temporal convolutional neural networks for continuous and discrete eeg emotion recognition. *IEEE Journal of Biomedical and Health Informatics* 28, 7 (2024), 3953–3964.

[11] Kübra Erat, Elif Bilge Şahin, Furkan Doğan, Nur Merdanoğlu, Ahmet Akcakaya, and Pınar Onay Durdu. 2024. Emotion recognition with EEG-based brain-computer interfaces: a systematic literature review. *Multimedia tools and applications* 83, 33 (2024), 79647–79694.

[12] Hongxiang Gao, Zhipeng Cai, Xingyao Wang, Min Wu, and Chengyu Liu. 2025. Multimodal Fusion of Behavioral and Physiological Signals for Enhanced Emotion Recognition Via Feature Decoupling and Knowledge Transfer. *IEEE journal of biomedical and health informatics* (2025).

[13] Rong-Fei Gu, Yi-Dong Zhao, Li-Ming Zhao, Wei-Long Zheng, and Bao-Liang Lu. 2024. A Multi-task Emotion Recognition Model Based on Continuously Labeled EEG Signals. In *International Conference on Neural Information Processing*. Springer, 375–386.

[14] Zhuen Guo, Mingqing Yang, Li Lin, Jisong Li, Shuyue Zhang, Qianbo He, Jiaqi Gao, Heling Meng, Xinran Chen, Yuehao Tao, et al. 2024. E-MFNN: an emotion-multimodal fusion neural network framework for emotion recognition. *PeerJ Computer Science* 10 (2024), e1977.

[15] Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. 2024. Fuse-moe: Mixture-of-experts transformers for fleximodal fusion. *Advances in Neural Information Processing Systems* 37 (2024), 67850–67900.

[16] Essam H Houssein, Asmaa Hammad, and Abdelmgeid A Ali. 2022. Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. *Neural Computing and Applications* 34, 15 (2022), 12527–12557.

[17] Hyo-Jeong Jang, Hye-Bin Shin, and Seong-Whan Lee. 2025. Uncertainty-Aware Cross-Modal Knowledge Distillation with Prototype Learning for Multimodal Brain-Computer Interfaces. *arXiv preprint arXiv:2507.13092* (2025).

[18] Xiaofang Jin, Jieyu Xiao, Libiao Jin, and Xinruo Zhang. 2024. Residual multimodal Transformer for expression-EEG fusion continuous emotion recognition. *CAAI Transactions on Intelligence Technology* 9, 5 (2024), 1290–1304.

[19] Kranti Kamble and Joydeep Sengupta. 2023. A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals. *Multimedia Tools and Applications* 82, 18 (2023), 27269–27304.

[20] Jun-Ho Maeng, Dong-Hyun Kang, and Deok-Hwan Kim. 2020. Deep learning method for selecting effective models and feature groups in emotion recognition using an Asian multimodal database. *Electronics* 9, 12 (2020), 1988.

[21] Haidy H Mustafa, Nagy R Darwish, and Hesham A Hefny. 2024. Automatic speech emotion recognition: a systematic literature review. *International Journal of Speech Technology* 27, 1 (2024), 267–285.

[22] Priyadarsini Samal and Mohammad Farukh Hashmi. 2024. Role of machine learning and deep learning techniques in EEG-based BCI emotion recognition system: a review. *Artificial Intelligence Review* 57, 3 (2024), 50.

[23] Premjeet Singh, Goutam Saha, and Md Sahidullah. 2021. Non-linear frequency warping using constant-Q transformation for speech emotion recognition. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 1–6.

[24] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing* 3, 1 (2011), 42–55.

[25] Yaxin Zhang, Xuegang Xu, Yuetao Du, and Ningchao Zhang. 2025. TMU-Net: A Transformer-Based Multimodal Framework with Uncertainty Quantification for Driver Fatigue Detection. *Sensors* 25, 17 (2025), 5364.

[26] Hongyu Zhu, Lin Chen, and Mingsheng Shang. 2026. BiMoE: Brain-Inspired Experts for EEG-Dominant Affective State Recognition. *arXiv preprint arXiv:2603.29205* (2026).