Unimodal vs Multimodal Coupling in Emotion Recognition: An Explainable Framework using Physiological States and Transitions

Anubhav[1,*], Kantaro Fujiwara[1]

[1]The University of Tokyo, Tokyo, Japan

Abstract

Emotion recognition from physiological signals has traditionally prioritized classification accuracy over interpretability, often relying on complex black-box models such as deep neural networks. In contrast, this work proposes a lightweight and explainable framework for emotion recognition using multimodal physiological data from the DEAP dataset. We focus on all four emotional dimensions, valence, arousal, dominance, and liking, by analyzing binarized state and transition dynamics across seven physiological modalities: EEG, EOG, EMG, GSR, respiration rate, PPG, and temperature. Our framework explores three layers of explainability: modality-wise signal relevance, pairwise multimodal coupling, and the dynamics of state and state transitions. Using Spearman rank correlation with subject-wise and inter-subject aggregation, we identify interpretable physiological patterns and transition signatures correlating with emotional states. Results show that transition-based features from modality couplings consistently outperform unimodal analyses across all emotion dimensions. In the subject-wise setting, the strongest correlations were observed for valence also via GSR–Resp ($\rho = 0.578, p < 0.0001$), arousal via EOG–EMG ($\rho = 0.580$, $p < 0.0001$), dominance via EEG–PPG ($\rho = 0.594$, $p < 0.0001$), and liking via GSR–Resp ($\rho = 0.589, p < 0.0001$). However, during inter-subject aggregation, the strength of these correlations diminished, suggesting that emotional signatures in physiological signals exhibit significant inter-individual variability and benefit from subject-specific modeling. These findings underscore the utility of explainable multimodal coupling for real-time, interpretable prediction systems and lay the groundwork for future integration into adaptive, personalized frameworks with the potential to scale toward more generalizable, context-aware emotion recognition systems across diverse user populations.

Keywords

Multimodal AI, XAI, Physiological Signals, Emotion Recognition,

1. Introduction

Affective computing using physiological signals has gained significant momentum in recent years, driven by advances in wearable sensor technology, increased affordability, and improved signal fidelity. These developments have enabled continuous, non-invasive monitoring of internal states such as stress, arousal, and emotional reactivity. The importance of such tools has become particularly salient in the post-COVID-19 era, where rates of anxiety and depression have risen sharply, prompting renewed interest in unobtrusive methods for emotional well-being monitoring and regulation. In this context, physiological emotion recognition systems offer a promising foundation for real-time mental health support, biofeedback applications, and emotionally intelligent interfaces.

Among various sensing modalities, physiological signals such as electroencephalography (EEG), electrooculography (EOG), galvanic skin response (GSR), photoplethysmography (PPG), and respiration rate are increasingly recognized for their ability to provide objective and continuous indicators of emotional state. Unlike vision or audio-based emotion recognition, these biosignals can be measured in silent, private, or wearable settings without reliance on external context. In particular, the DEAP dataset has become a benchmark resource for physiological emotion research, offering a multimodal

MAI-XAI 2025: 2nd Workshop on Multimodal, Affective and Interactive eXplainable Artificial Intelligence, October 25-26, 2025, Bologna, Italy

*Corresponding author.

anubhav2901@g.ecc.u-tokyo.ac.jp (Anubhav); kantaro@g.ecc.u-tokyo.ac.jp (K. Fujiwara)

0000-0002-2480-6119 (Anubhav); 0000-0001-8114-7837 (K. Fujiwara)

collection of EEG and peripheral signals recorded across 32 subjects and annotated using four core emotion dimensions: valence, arousal, dominance, and liking [1].

Physiological signal-based models are especially valuable for wearable embedded BCI applications due to their low latency, low noise, and potential for unobtrusive acquisition. With the rise of real-time applications, there is a growing demand for interpretable emotion recognition models that balance performance with explainability and resource efficiency. While earlier studies have shown high classification accuracy using deep learning models such as Long Short-Term Memory networks (LSTM) [2, 3], or non-linear dynamics methods based on reservoir computing [4, 5], these approaches often lack transparency in revealing why specific physiological features contribute to emotion recognition outcomes.

Recent work in multimodal learning has highlighted the potential of fusing multiple physiological signals to improve classification robustness [6, 7]. However, most models focus only on valence and arousal, neglecting DEAP's complete label set, which includes dominance and liking. Furthermore, fusion techniques treat multimodal input holistically without systematically examining pairwise modality interactions or temporal dynamics. This omission represents a missed opportunity to explore how unimodal signals contribute individually and how their pairwise combinations, multimodal couplings, can reveal interpretable emotion-specific interactions and transition patterns.

In this context, there remains a clear gap in the literature. First, most existing approaches offer limited support for explainability, particularly in identifying interpretable biomarkers within or across modalities. Second, the dominance and liking dimensions remain underexplored emotion dimensions. Third, while state dynamics and transitions are implicit in sequential models, few studies extract these features explicitly in a computationally lightweight and interpretable fashion. In particular, prior research has not systematically compared unimodal versus multimodal coupling strategies in the context of emotion recognition.

To address these gaps, we propose a novel, explainable framework for multimodal physiological emotion recognition using the DEAP dataset. Our method is centered around three dimensions of analysis: (1) signal-level representation (unimodal), (2) pairwise modality coupling (multimodal), and (3) temporal dynamics in terms of binary state transitions. Each signal is binarized based on standardized trial-level statistics and transformed into interpretable features such as state frequency, transition count, and directional state-transition sequences. Using Spearman correlation, we relate these features to all four emotional dimensions: valence, arousal, dominance, and liking, at both intra-subject and inter-subject levels.

Our framework yields interpretable correlation scores and visual summaries identifying dominant modalities and modality pairs associated with specific emotional experiences. We demonstrate that our transition-based multimodal coupling features correlate more with emotional labels than unimodal models while preserving computational efficiency. The remainder of this paper is organized as follows: Section 2 reviews prior work on interpretable emotion recognition. Section 3 describes our proposed framework. Section 4 presents our empirical findings across the three analytical layers. Section 5 discusses the implications of these findings, and Section 6 concludes the paper with future research directions.

2. Related Work

2.1. Unimodal EEG-Based Emotion Recognition

Emotion recognition using EEG has been extensively studied, particularly within subject-dependent settings using supervised learning. Traditional approaches extract frequency-domain or time-frequency features from EEG signals and use classifiers such as SVMs or deep networks. For instance, Nath et al. [2] and Anubhav and Fujiwara [3] achieved classification accuracies exceeding 93% on valence and arousal using LSTM-based models. However, these methods are generally black-box in nature and rely on high-dimensional input, limiting their interpretability.

Alternative modeling frameworks such as Reservoir Computing (RC) offer biologically inspired processing of EEG dynamics with reduced training complexity. Anubhav and Fujiwara proposed a reservoir splitting strategy [4] to isolate brain-lobe-specific activity. Their subsequent multi-reservoir framework [5] evaluated EEG generalization across trials, subjects, and contexts. While these approaches improve efficiency, they are still constrained to spatial EEG features and do not explicitly capture modality interactions or transitions.

2.2. Multimodal Physiological Fusion and Signal Coupling

Multimodal learning strategies integrate EEG with peripheral physiological signals such as EOG, EMG, GSR, and PPG to capture richer emotional information. Bălan et al. [7] used Random Forests to classify V-A-D emotional states from EEG and GSR, analyzing feature contributions from each modality. Gohumpu et al. [8] further demonstrated that incorporating multiple peripheral modalities enhances both classification performance and interpretability. Yet, most fusion strategies treat modalities holistically, without analyzing their pairwise interactions.

Hierarchical and attention-based fusion models have been introduced to address modality relationships. Zhang et al. [9] proposed a hierarchical fusion network, while more recent transformer models such as ST-SHAP [10] and ERTNet [11] provide partial explainability through SHAP and attention weights. These architectures improve accuracy but rely on computationally intensive training and lack transparency in physiological terms. Saxena et al. [12] summarized such trends, highlighting that the performance-interpretability trade-off remains unresolved in multimodal emotion recognition.

2.3. Interpretability and Temporal Dynamics in Emotion Modeling

Despite increasing focus on explainability, most models fall short of delivering physiologically meaningful interpretations. Attention mechanisms and SHAP values indicate signal importance but do not reveal inter-modality dynamics or temporal structure. Shu et al. [13] surveyed entropy and mutual information metrics as proxies for EEG signal complexity, yet their application in multimodal temporal analysis remains rare.

Temporal aspects of emotional experience, such as transitions between physiological states, are underexplored. While deep models like LSTMs implicitly capture sequence dynamics, explicit modeling of temporal transitions and their relation to emotion has not been central in existing frameworks. Furthermore, existing approaches primarily address valence and arousal, neglecting dominance and liking due to modeling complexity.

2.4. Our Contribution: Interpretable Multimodal Coupling with Temporal Transitions

We address these limitations by proposing a fully interpretable and computationally efficient framework for emotion recognition using all four DEAP emotional labels: valence, arousal, dominance, and liking. Our approach departs from high-dimensional, opaque models and instead focuses on three core contributions:

- **Binarized signal representation across modalities:** We discretize seven physiological channels to enable direct analysis of state activation and transitions.
- **Pairwise modality coupling:** Rather than treating the multimodal signal as a fused whole, we examine synergistic interactions between specific modality pairs to uncover functional relationships.
- **Temporal transition modeling:** We statistically quantify cross-time transitions using a correlation-based framework, making temporal structure and dynamics explicit and interpretable.

Our method does not require deep networks or supervised training, making it suitable for real-time applications. By analyzing modality transitions and their correlations with emotional states, we offer a

novel pathway toward understanding the physiological basis of emotion in a multimodal and temporally resolved manner. Importantly, our work builds on conceptual foundations from Energy Landscape Analysis (ELA), which has been used to model brain state transitions [14, 15]. Unlike traditional ELA, which relies on maximum entropy modeling and is limited to unimodal EEG, our framework extends this perspective to multimodal binarized transitions, providing a scalable and interpretable alternative for emotion research.

To contextualize our contribution, Table 1 summarizes selected studies emphasizing interpretable frameworks using unimodal and multimodal physiological signals. As shown, most approaches focus on valence and arousal only and rarely include modality-level transition features or pairwise signal coupling.

Table 1
Overview of interpretable physiological emotion recognition models.

Paper	Modality	Model Type	Explainability	Emotion Labels
Koelstra et al. (2012) [1]	EEG + peripherals	Naïve Bayes + fusion	Frequency–label correlation	Valence, Arousal, Liking
Atkinson and Campos (2016) [6]	EEG	SVM + mRMR	Feature importance (MI)	Valence, Arousal
Bălan et al. (2019) [7]	EEG + GSR + PPG	Random Forest	Feature ranking	Valence, Arousal, Dominance
Anubhav and Fujiwara (2024) [5]	EEG	Reservoir Computing	Topographic channel analysis	Valence, Arousal
Miao et al. (2024) [10]	EEG	Transformer + SHAP	Channel-wise SHAP weights	Valence, Arousal
Liu et al. (2024) [11]	EEG	Transformer (ERTNet)	Attention + SHAP	Valence, Arousal
Fusion-based works	Peripheral physiological signals (Gohumpu et al., 2023 [8])	Multimodal fusion	Modality contribution analysis	Valence, Arousal
	EEG + others (Zhang et al., 2021 [9])	Hierarchical Fusion Network	Fusion hierarchy modeling	Valence, Arousal
Shu et al. (2018) [13]	EEG + physiological	Entropy/MI measures	Signal dynamics interpretation	Valence, Arousal
This work	EEG + EOG + others	Correlation + transitions	State-transition interpretability	**Valence, Arousal, Dominance, Liking**

3. Methodology

3.1. Dataset and Preprocessing

The present study uses the DEAP dataset [1], a widely adopted multimodal emotion corpus that includes physiological recordings from 32 participants across 40 video trials. Each trial includes self-reported ratings on four emotional dimensions: valence, arousal, dominance, and liking, scored on a 9-point Likert scale.

We utilize all seven physiological modalities recorded in the dataset: 32 EEG channels, 2 EOG channels, 2 EMG channels, GSR, respiration (RESP), photoplethysmography (PPG), and skin temperature (TEMP). To ensure uniformity across modalities and reduce dimensionality, we average signals across the

corresponding channels: EEG, EOG, and EMG signals are each reduced to a single average vector per trial, resulting in a final 7-dimensional time series per trial.

All signals are standardized on a per-trial basis using z-score normalization. We do not downsample the signals and maintain the 128 Hz sampling rate provided in the preprocessed DEAP dataset. The emotional labels are also standardized within subjects but are preserved in their continuous form for correlation analysis.

3.2. Binarization and Feature Representation

Each of the seven physiological signals is binarized within a trial such that a value of 1 is assigned when the standardized signal exceeds 0 (i.e., above the mean), and 0 otherwise. This binarization allows us to extract interpretable features based on binary state logic.

We extract three categories of features from each modality and modality-pair:

- **State-based:** Frequency of signal being in state 1 over the trial duration.
- **Transition-based:** Counts of $0 \to 1$ and $1 \to 0$ transitions within each trial.
- **State-transition-based:** Joint transition features over time, including:
  - For unimodal: transitions such as $0 \to 0$, $0 \to 1$, $1 \to 0$, $1 \to 1$.
  - For multimodal pairs: coupled transitions such as $00 \to 01$, $10 \to 11$, $01 \to 00$, etc., over time-aligned signals.

This feature structure enables both intra- and inter-modality interpretability.

3.3. Analysis Framework

Our methodology is structured along three analytical layers:

- **Modality-Level Analysis:** We examine both unimodal signals and all pairwise combinations of modalities. For 7 modalities, this results in 21 pairwise modality combinations. For each combination, we evaluate the informativeness of state and transition features independently.
- **Subject-Level Analysis:** We conduct both intra-subject and inter-subject analyses. In the intra-subject setup, features and emotion labels are correlated per subject and then aggregated. In the inter-subject analysis, features across all trials and subjects are pooled before correlation analysis.
- **Feature-Type Analysis:** For each modality or modality-pair, we compute features using the three schemes above (state-based, transition-based, state-transition-based). This layered setup allows us to assess the importance of dynamic signal behavior as opposed to static values.

We use Spearman rank correlation to assess the relationship between extracted features and the four continuous emotion labels. Spearman's ρ is chosen due to its non-parametric nature and robustness to outliers. For each feature and label pair, we compute both the correlation coefficient ρ and associated p-value. The feature with the highest ρ (with $p < 0.05$) for each modality and modality-pair is retained for further visualization.

4. Results

4.1. Unimodal Analysis

4.1.1. State-Based Analysis

We begin by analyzing the binary state frequency for each of the seven physiological modalities across all trials. Figure 1 shows the best-performing modality per subject and emotion label. The heatmaps indicate that **EOG** and **PPG** consistently emerge as top modalities for *liking* and *dominance*, while **EEG** and **GSR** dominate in the *valence* and *arousal* dimensions.

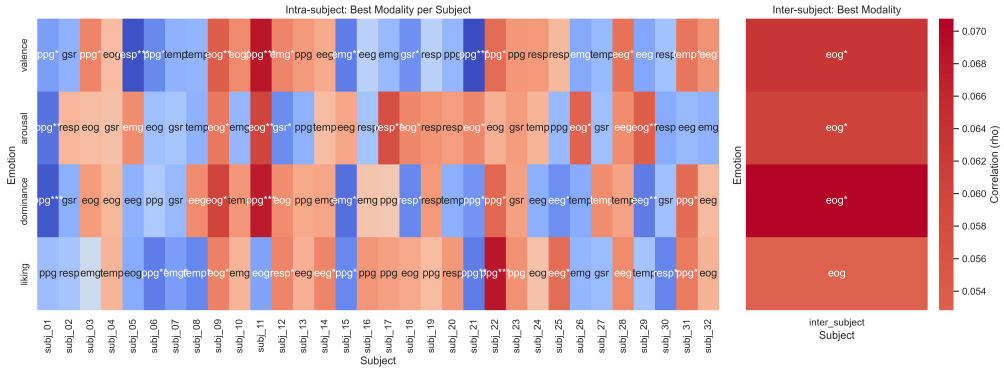

Figure 1: Intra-subject distribution of best-performing unimodal signals across the four DEAP labels. Each color indicates the modality with highest ρ per subject.

To assess statistical significance across modalities, Figure 2 shows boxplots of Spearman correlation coefficients (ρ) between state frequency and each emotion label across all subjects. Overall, the distribution of correlation values is centered near zero, with the strongest modality-label combinations reaching $\rho \approx 0.10$. Notably, the corrected p-values from pairwise comparisons indicate no statistically significant differences between modalities (all $p > 0.12$), as summarized in Appendix A.

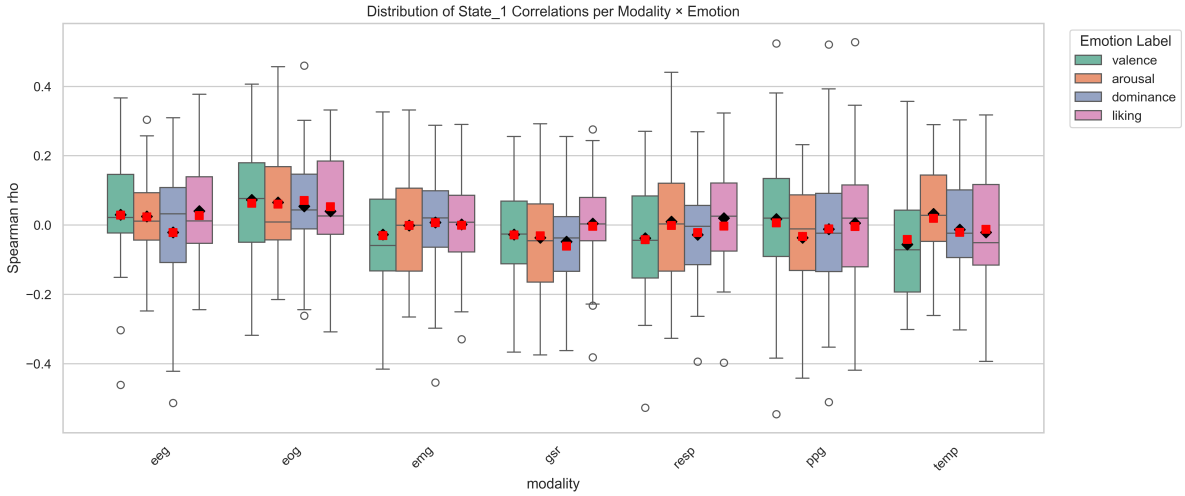

Figure 2: Distribution of Spearman ρ for unimodal state frequency across subjects. No statistically significant differences were observed between modalities.

4.1.2. Transition-Based Analysis

We next evaluated unimodal transitions between states, particularly the counts of $1 \to 0$ and $0 \to 1$ events. Figure 3 summarizes the transition feature that yields the highest correlation with each emotion label.

From the inter-subject analysis, we observe:

- For **arousal**, the $1 \to 0$ transition in **GSR** exhibits the highest correlation ($\rho = 0.10$, $p < 0.01$).
- For **liking** and **valence**, the $1 \to 0$ transition in **EOG** shows the highest correlations ($\rho = 0.09$ and $\rho = 0.07$ respectively, both $p < 0.05$).
- For **dominance**, the strongest correlation was again from **GSR**'s $1 \to 0$ transition ($\rho = 0.05$, $p < 0.05$).

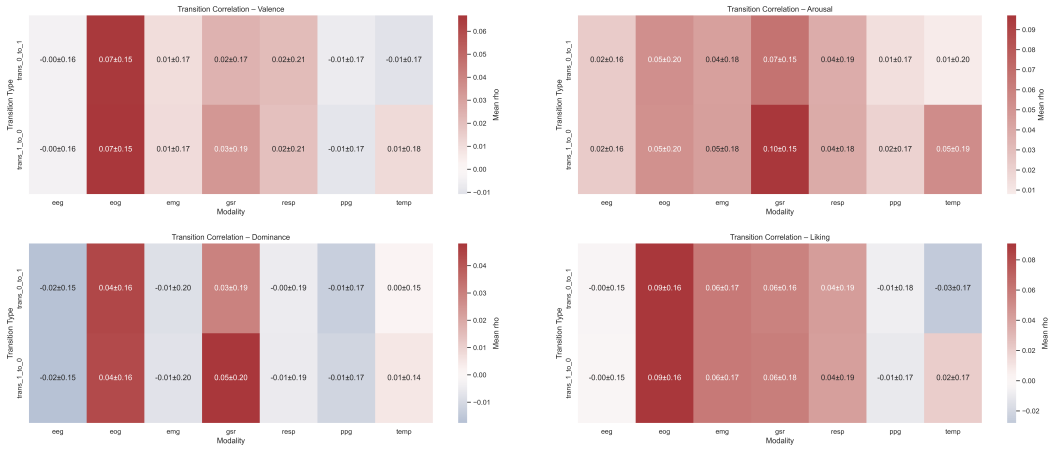

Figure 3: Unimodal transition-based analysis: correlation heatmaps across subjects. High-performing transitions are seen in GSR and EOG for different emotion labels.

These findings highlight that while the absolute strength of unimodal correlations remains modest, incorporating transition dynamics allows for capturing signal patterns not revealed in state-based analysis alone.

4.2. Multimodal Analysis

4.2.1. State-Based Analysis

In the multimodal setting, we evaluate all $\binom{7}{2} = 21$ pairwise combinations of physiological modalities. For each modality pair, we extract joint binary state frequencies (e.g., fraction of time both signals are in state 0).

Figure 4 visualizes the best-performing state for each pair across emotion labels. In the inter-subject analysis, we observe:

- **Valence:** highest correlation from `state_00` of **EMG–RESP** ($\rho = 0.09$, $p < 0.01$)
- **Arousal:** best feature is `state_00` of **EEG–EOG** ($\rho = 0.07$, $p < 0.05$)
- **Dominance:** `state_10` of **EEG–EOG** ($\rho = 0.07$, $p < 0.05$)
- **Liking:** `state_01` of **EOG–TEMP** ($\rho = 0.07$, $p < 0.05$)

The corresponding network diagrams in Figure 5 show the top modality pairs for each label across intra- and inter-subject aggregations. Particularly, EEG–EOG and EOG–TEMP consistently appear as high-weight nodes, suggesting strong modality coupling for valence and liking.

4.2.2. Transition-Based Analysis

Finally, we analyze state-transition dynamics in modality pairs. Transition pairs are defined as time-aligned 2-tuple changes across two binarized signals. The best transitions across the combined dataset are visualized in Figure 6.

- **Valence:** `trans_0_2` from **GSR–RESP** ($\rho = 0.578$, $p < 0.001$)
- **Arousal:** `trans_0_2` from **EOG–EMG** ($\rho = 0.580$, $p < 0.001$)
- **Dominance:** `trans_1_1` from **EEG–PPG** ($\rho = 0.594$, $p < 0.001$)
- **Liking:** `trans_0_2` from **GSR–RESP** ($\rho = 0.589$, $p < 0.001$)

Inter-subject network structures visualized in Figure 7 further emphasize the strength of transition-based features, especially in PPG and GSR. These features not only surpass all unimodal results but also introduce interpretable coupling effects between central (EEG/EOG) and peripheral (GSR/PPG) signals.

In summary, state-transition-based multimodal features provide the highest interpretability and effect sizes, demonstrating strong, statistically significant associations with all four emotional labels.

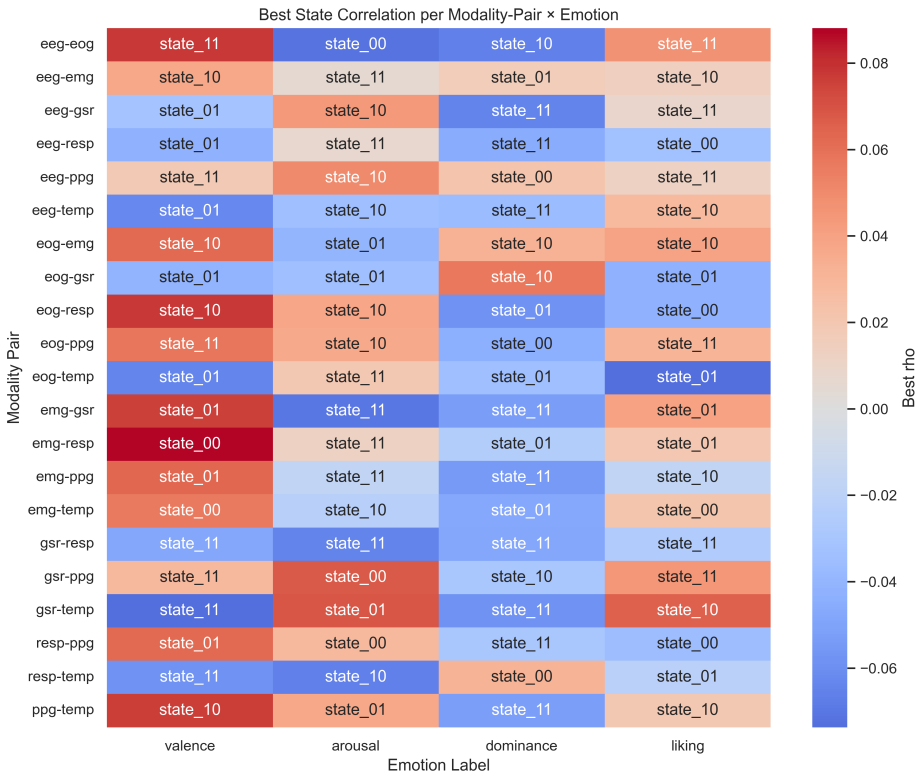

Figure 4: Multimodal state-based heatmap showing strongest state-pair correlation for each modality combination across the four emotion labels.

5. Discussion

Our findings show that transition-based multimodal features consistently outperform unimodal and static state-based features. Notable pairings, such as GSR–RESP ($\rho = 0.578$, $p < 0.001$ for valence) and EEG–PPG ($\rho = 0.594$, $p < 0.001$ for dominance, demonstrate clear physiological relevance. Our framework also highlights the importance of transition patterns, such as $0 \to 2$ and $1 \to 1$, as discriminative features for emotion recognition.

Importantly, our analysis includes dominance and liking, which are often omitted in affective computing literature, providing interpretable correlations for previously underexplored modality pairs like EEG–PPG (dominance) and EOG–TEMP (liking) [6, 7]. This result extends prior studies that focused on EEG-only data [2, 3] or employed multimodal fusion without interpretability [9].

5.1. Comparison with Other Methods

In terms of methodology, our framework is computationally efficient and avoids model training, parameter tuning, or high-dimensional feature extraction. This stands in contrast to deep learning models such as LSTM [3] and reservoir computing approaches [5], making our method suitable for real-time, low-power applications. Unlike SHAP-based transformer models [10] or attention-based explanations in ERTNet [11], our correlation-based analysis offers direct interpretability by linking each ρ score to simple physiological patterns. Furthermore, our framework conceptually aligns with ELA, which models metastable brain dynamics using binarized EEG [14, 15]. While ELA traditionally requires complex parameter fitting and is limited to unimodal data, our method offers a tractable and extensible alternative for future multimodal applications in emotional state modeling.

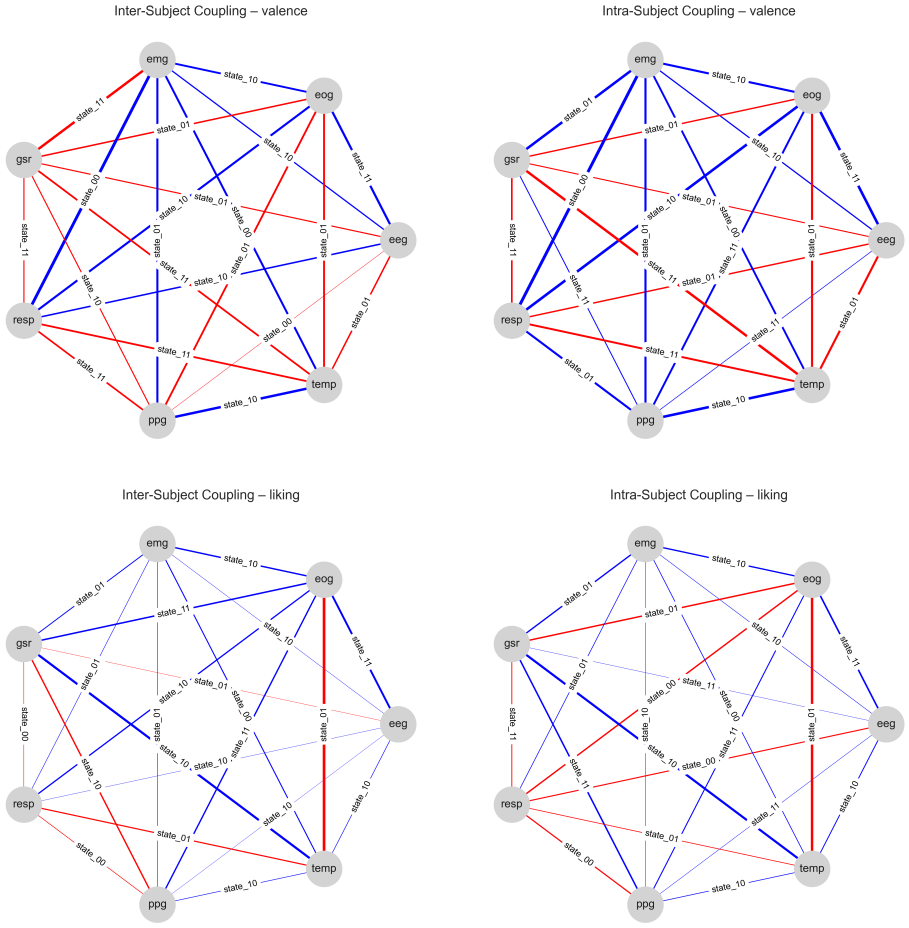

Figure 5: Network graphs showing high-correlation multimodal pairings for valence and liking. Edge thickness reflects inter-subject correlation strength.

5.2. Predictive Performance

To complement our correlation-based analysis, we conducted a lightweight predictive comparison using logistic regression over the four interpretable feature families (unimodal state, unimodal transitions, multimodal state, multimodal transitions) with both within-subject cross-validation and leave-one-subject-out (LOSO) evaluation. Within-subject, *valence* achieved $\mathrm{AUROC} = 0.557\,[0.523,\ 0.592]$ with multimodal transition features, and *arousal* reached $0.550\,[0.515,\ 0.584]$ with unimodal transitions, both significantly above chance. *Dominance* and *liking* were weaker ($\mathrm{AUROC} \approx 0.52$–$0.54$, confidence intervals overlapping 0.5). Under LOSO, performance was modest (most AUROCs $\approx$ 0.51–0.54), reflecting strong inter-subject variability, in line with our correlation analyses. These results indicate that the proposed interpretable transition features are not only statistically associated with emotion labels but are also predictive while remaining computationally lightweight.

5.3. Implications for Personalization and Generalization

The predictive trends above dovetail with our correlation findings and reinforce a central takeaway of this work: subject-specific modeling matters. The clear within-subject gains alongside modest LOSO performance indicate that interpretable transition features capture stable, person-dependent signatures that dilute under cross-subject pooling. Practically, this favors lightweight, on-device personalization (e.g., brief calibration or adaptive thresholds) for real-time use, while encouraging future work on transfer and normalization strategies that preserve interpretability. Taken together with our correlation results, these baselines strengthen the case that multimodal transition structure is a meaningful, human

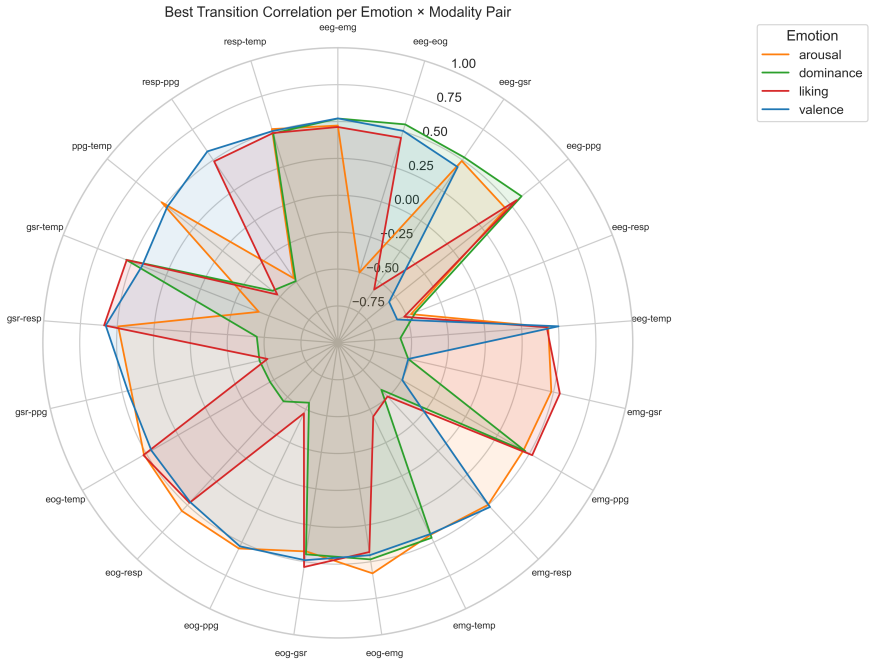

Figure 6: Radial plot showing best multimodal transitions correlated with each emotion label across all subjects. Each arc shows a modality pair with highest ρ.

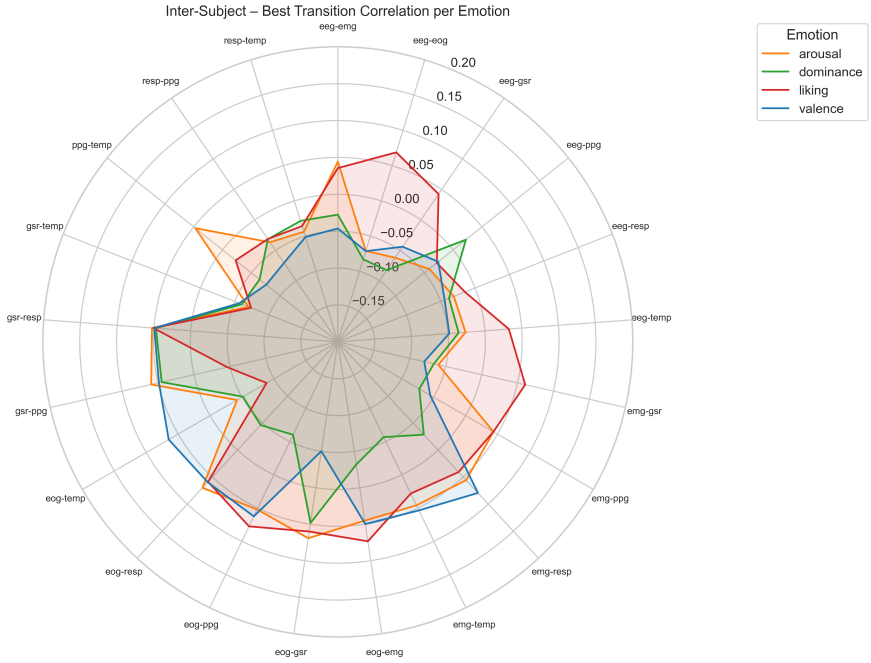

Figure 7: Inter-subject radial summary of best-performing transitions. GSR–RESP and EEG–PPG show strong coupling patterns for emotional relevance.

parsable substrate for emotion modeling even when absolute cross-subject accuracy remains challenging.

5.4. Limitations and Future Work

However, several limitations must be acknowledged. First, the binarization process, while useful for interpretability, simplifies continuous signals and may overlook finer-grained patterns. Second, while Spearman correlation captures monotonic relationships, it does not model conditional dependencies or

interactions involving three or more modalities. Third, our analysis does not yet incorporate subject-level variability modeling or personalized baselines, which have shown utility in previous EEG studies [5]. Furthermore, we currently analyze each trial as an independent unit, which may limit generalizability in time-locked or context-rich scenarios.

Future work will address these limitations in multiple directions. One promising extension is to explore multi-threshold or ternary state representations that preserve more information without compromising interpretability. Additionally, integrating this framework with shallow learning classifiers (e.g., logistic regression with binary state features) could bridge the gap between interpretable analysis and predictive modeling. Incorporating personalized baselines or adapting transition features to dynamic resting-state estimates can further improve robustness across subjects. Lastly, a key future direction involves explicitly linking our statistical transition framework to multimodal energy landscape models. This would allow us to define basins of attraction, compute energy gradients across modalities, and map emotion trajectories in a formalized state-space. Applying this methodology to other benchmark datasets such as DREAMER [16] or AMIGOS [17] would validate and extend the current interpretations across acquisition conditions and emotional contexts, offering better generalizability.

6. Conclusion

This paper presents an interpretable and computationally efficient framework for emotion recognition using multimodal physiological signals. Using DEAP dataset, we conduct a comprehensive correlation-based analysis across four emotion dimensions: valence, arousal, dominance, and liking, by examining both unimodal and multimodal signal pairings through state- and transition-based features. The framework operates entirely on binarized representations and interpretable signal dynamics, requiring no complex learning algorithms, and is therefore well-suited for low-latency, real-time emotion tracking applications.

Our results consistently demonstrate that transition-based multimodal features outperform unimodal and state-based counterparts in both intra- and inter-subject analyses. These findings underscore the importance of examining temporal dynamics and cross-modal interactions to understand the physiological basis of emotional experience. The approach not only confirms known relationships, such as EEG's relevance to valence and arousal, but also identifies underutilized modalities such as GSR, TEMP, and PPG as critical in characterizing dominance and liking, which often overlooked in prior studies.

Beyond the empirical results, this study contributes to the field of explainable artificial intelligence (XAI) by offering a framework grounded in transparent, domain-relevant features. The correlation outputs are directly interpretable, and the modality-level analysis provides physiologically meaningful insights. In contrast to black-box deep models, our method reveals interpretable transition structures that align with human-understandable physiological behavior. The proposed methodology also opens up new possibilities for bridging physiological emotion recognition with statistical physics-based frameworks such as energy landscape analysis. Our observations of recurring stable transition motifs and modality-pair coupling patterns suggest a latent structure in emotional state space that mirrors metastable attractor dynamics in energy landscapes.

Future research should explore more expressive discrete encodings, dynamic baseline adjustments, and integration with shallow predictive models. Importantly, a promising extension involves modeling the transition networks as multimodal energy landscapes, identifying basins of emotional stability, and capturing the temporal evolution of emotional states through energy gradients. Such an approach could provide a powerful formalism for describing emotion dynamics in physiological data while maintaining interpretability. Extending this analysis to other datasets and contextual conditions will further enhance its generalizability and applicability to real-world emotion-aware systems.

Acknowledgments

JSPS KAKENHI Grant Numbers JP22K18419, JP24K15161, JP25H00451, JST Moonshot RD Grant No. JPMJMS2021

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o and Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis using physiological signals, IEEE Transactions on Affective Computing 3 (2012) 18–31. doi:10.1109/T-AFFC.2011.15.

[2] D. Nath, Anubhav, M. Singh, D. Sethia, D. Kalra, S. Indu, A comparative study of subject-dependent and subject-independent strategies for eeg-based emotion recognition using lstm network, in: Proceedings of the 2020 4th International Conference on Compute and Data Analysis, ACM, 2020, pp. 142–147. URL: https://doi.org/10.1145/3388142.3388167. doi:10.1145/3388142.3388167.

[3] Anubhav, D. Nath, M. Singh, D. Sethia, D. Kalra, S. Indu, An efficient approach to eeg-based emotion recognition using lstm network, in: 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), IEEE, 2020, pp. 88–92. URL: https://doi.org/10.1109/CSPA48992.2020.9068691. doi:10.1109/CSPA48992.2020.9068691.

[4] Anubhav, K. Fujiwara, Reservoir splitting method for eeg-based emotion recognition, in: 2023 11th International Winter Conference on Brain-Computer Interface (BCI), IEEE, 2023, pp. 1–5. URL: https://doi.org/10.1109/BCI57258.2023.10078629. doi:10.1109/BCI57258.2023.10078629.

[5] Anubhav, K. Fujiwara, Across trials vs subjects vs contexts: A multi-reservoir computing approach for eeg variations in emotion recognition, in: Proceedings of the 26th International Conference on Multimodal Interaction, ACM, New York, NY, USA, 2024, pp. 518–525. URL: https://doi.org/10.1145/3678957.3685730. doi:10.1145/3678957.3685730.

[6] J. Atkinson, D. Campos, Improving bci-based emotion recognition by combining eeg, peripheral physiological signals, and eye-related measures, in: Proceedings of the International Conference on Physiological Computing Systems, SCITEPRESS, 2016, pp. 61–72. doi:10.5220/0006022100610072.

[7] R. Balan, J. Rajagopal, T. S. Kumar, Emotion recognition using EEG and peripheral physiological signals: A review, in: 2019 11th International Conference on Advanced Computing (ICoAC), IEEE, 2019, pp. 143–150. doi:10.1109/ICoAC48765.2019.246841.

[8] J. Gohumpu, M. Xue, Y. Bao, Emotion recognition with multi-modal peripheral physiological signals, Frontiers in Computer Science 5 (2023). doi:10.3389/fcomp.2023.1264713.

[9] Y. Zhang, C. Cheng, Y. Zhang, Multimodal Emotion Recognition Using a Hierarchical Fusion Convolutional Neural Network, IEEE Access 9 (2021) 7943–7951. doi:10.1109/ACCESS.2021.3049516.

[10] M. Miao, J. Liang, Z. Sheng, W. Liu, B. Xu, W. Hu, ST-SHAP: A hierarchical and explainable attention network for emotional EEG representation learning and decoding, Journal of Neuroscience Methods 414 (2025) 110317. doi:10.1016/j.jneumeth.2024.110317.

[11] R. Liu, Y. Chao, X. Ma, X. Sha, L. Sun, S. Li, S. Chang, ERTNet: An interpretable transformer-based framework for EEG emotion recognition, Frontiers in Neuroscience 18 (2024). doi:10.3389/fnins.2024.1320645.

[12] A. Saxena, A. Khanna, D. Gupta, Emotion Recognition and Detection Methods: A Comprehensive

Survey, Journal of Artificial Intelligence and Systems 2 (2020) 53–79. doi:10.33969/ais.2020.21005.
[13] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, X. Yang, A Review of Emotion Recognition Using Physiological Signals, Sensors 18 (2018) 2074. doi:10.3390/s18072074.
[14] T. Ezaki, T. Watanabe, M. Ohzeki, N. Masuda, Energy landscape analysis of neuroimaging data, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 375 (2017) 20160287. doi:10.1098/rsta.2016.0287.
[15] T. Watanabe, N. Masuda, F. Megumi, R. Kanai, G. Rees, Energy landscape and dynamics of brain activity during human bistable perception, Nature communications 5 (2014) 4765. doi:10.1038/ncomms5765.
[16] S. Katsigiannis, N. Ramzan, DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices, IEEE Journal of Biomedical and Health Informatics 22 (2018) 98–107. doi:10.1109/JBHI.2017.2688239.
[17] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, I. Patras, AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups, IEEE Transactions on Affective Computing 12 (2021) 479–493. doi:10.1109/TAFFC.2018.2884461.

A. Appendix

This appendix provides comprehensive correlation results referenced in the main paper. We include intra-subject and inter-subject correlation tables for both state-based and transition-based features, covering unimodal and multimodal analyses. Additionally, we report normalized state distributions by modality and emotion label.

A.1. Unimodal State-Based Correlations

Table 2 presents the top and statistically significant unimodal state-based correlations between physiological modalities and emotional dimensions across inter-subject aggregation. Notably, the EOG signal shows weak but consistent correlations for all four emotion labels, especially for dominance ($\rho = 0.07$, $p = 0.011$). GSR signals also correlate with dominance, albeit with lower effect sizes.

Table 2
Top state-based correlations (inter-subject) for each emotion label.

Label	Modality	State	Correlation (ρ)	p-value
Valence	EOG	state_1	0.06	0.024
Arousal	EOG	state_1	0.06	0.031
Dominance	EOG	state_1	0.07	0.011
Dominance	GSR	state_0	0.06	0.030
Liking	EOG	state_1	0.05	0.059

B. Unimodal Transition-Based Correlations

This section presents the correlation-based results for transition patterns within individual physiological modalities, independently assessed for their association with emotional dimensions: valence, arousal, dominance, and liking. For each modality, we compute transition frequencies of the form $i \to j$ (where $i, j \in \{0, 1\}$) over binarized state representations. The resulting features are evaluated using Spearman correlation with emotion scores, aggregated at the inter-subject level and presented in Table 3. Among unimodal analyses, only the EOG and GSR modalities yielded statistically significant results. The EOG transition from high to low ($1 \to 0$) was moderately correlated with liking ($\rho = 0.06$,

$p = 0.039$). However, across all emotion labels, unimodal transition-based features generally showed weak correlations, reinforcing the necessity of multimodal coupling for robust emotional inference.

Table 3
Top correlation results for transition-based unimodal features (inter-subject level).

Emotion Label	Modality + Transition	ρ	p-value
Valence	EOG – trans_1_to_0	0.05	0.083
Arousal	GSR – trans_1_to_0	0.05	0.083
Dominance	GSR – trans_1_to_0	0.03	0.233
Liking	EOG – trans_1_to_0	0.06	0.039

C. Multimodal State-Based Correlations

This appendix presents detailed inter-subject correlation results for the state-based multimodal emotion recognition framework using binarized state combinations across modality pairs. For each emotion label—valence, arousal, dominance, and liking—we report statistically significant correlations between binary state patterns and emotion ratings, using Spearman's rho and the corresponding p-values.

C.1. Top Correlations by Emotion Dimension

Table 4 summarizes the highest correlations observed for each emotion label in the inter-subject multimodal state-based analysis with Table 5 shows of statistically significant ($p < 0.05$) correlations observed in the inter-subject analysis for each emotion label. These results reflect the joint state patterns across modality pairs that are most informative of emotional states.

Table 4
Top multimodal state-based correlations (inter-subject) per emotion label.

Emotion Label	Modality Pair	State	Correlation (ρ, p)
Valence	EMG–RESP	state_00	$\rho = 0.09, p = 0.002$
Arousal	GSR–TEMP	state_01	$\rho = 0.07, p = 0.008$
Dominance	EEG–EOG	state_10	$\rho = -0.08, p = 0.005$
Liking	EOG–TEMP	state_01	$\rho = -0.08, p = 0.004$

C.2. Interpretation

While the overall correlation values are moderate ($|\rho| < 0.1$), the consistency of specific modality pairs across emotional dimensions, especially involving EEG–EOG, GSR–TEMP, and EMG–RESP, suggests robust inter-modality state dependencies. Interestingly, state_01 and state_11 emerge as frequently informative binary configurations, indicating patterns where both modalities are either concurrently low or concurrently high. These results complement the intra-subject and transition-based findings in the main text by reinforcing the value of analyzing simple, interpretable joint states across physiological signals.

C.3. Multimodal Transition-Based Correlations

C.3.1. Intra-Subject

Table 6 lists the top-performing multimodal transition features based on intra-subject correlation analysis. These pairs exhibit significantly stronger correlations ($\rho > 0.57, p < 0.001$), underscoring the utility of modeling cross-modality transitions for capturing affective states.

Table 5
Significant multimodal state-based correlations (inter-subject) across all emotion labels. Each cell shows modality pair (top) and binary state (bottom) with corresponding Spearman correlation and p-value.

Valence		Arousal		Dominance		Liking	
Modality	ρ, p	**Modality**	ρ, p	**Modality**	ρ, p	**Modality**	ρ, p
EMG–RESP	0.09,	GSR–TEMP	0.07,	EEG–EOG	-0.08,	EOG–TEMP	-0.08,
state_00	0.002	state_01	0.008	state_10	0.005	state_01	0.004
EMG–GSR	-0.07,	GSR–TEMP	-0.07,	EEG–EOG	0.07,	EOG–TEMP	0.07,
state_11	0.009	state_11	0.016	state_01	0.008	state_11	0.012
PPG–TEMP	0.07,	EEG–EOG	-0.06,	EEG–GSR	0.07,	GSR–TEMP	0.07,
state_10	0.009	state_00	0.027	state_00	0.013	state_10	0.017
PPG–TEMP	-0.07,	EEG–EOG	0.06,	EMG–PPG	0.07,	GSR–TEMP	-0.06,
state_11	0.009	state_11	0.033	state_01	0.014	state_11	0.021
EMG–GSR	0.07,	EEG–EOG	0.06,	EMG–PPG	-0.07,	EEG–EOG	0.06,
state_01	0.014	state_01	0.042	state_11	0.016	state_11	0.040
EEG–EOG	0.07,	GSR–TEMP	-0.06,				
state_11	0.019	state_00	0.048				
EOG–RESP	0.06,						
state_10	0.023						
EEG–EOG	-0.06,						
state_00	0.024						
EEG–EOG	-0.06,						
state_10	0.025						
EMG–PPG	0.06,						
state_01	0.028						

Table 6
Top multimodal transition features (intra-subject).

Emotion	Signal Pair	Transition	ρ (p-value)
Arousal	EOG–EMG	trans_0_2	0.580 (8.84×10^{-5})
Dominance	EEG–PPG	trans_1_1	0.594 (5.41×10^{-5})
Liking	GSR–RESP	trans_0_2	0.589 (6.35×10^{-5})
Valence	GSR–RESP	trans_0_2	0.578 (9.30×10^{-5})

Table 7
Top multimodal transition features (inter-subject).

Emotion	Signal Pair	Transition	ρ (p-value)
Arousal	EOG–EMG	trans_0_2	0.240 (0.17)
Dominance	EEG–PPG	trans_1_1	0.274 (0.11)
Liking	GSR–RESP	trans_0_2	0.265 (0.12)
Valence	GSR–RESP	trans_0_2	0.228 (0.19)

C.3.2. Inter-Subject

Table 7 presents the same analysis applied across subjects. As noted in Section 5, the inter-subject correlations are weaker, indicating substantial variability in how individuals physiologically express emotion. This supports our emphasis on subject-specific modeling in real-time systems.