A Hierarchical Graph-Based Framework for Cross-Subject Emotion Recognition Using Multimodal Physiological Signals

Yufan Du, Yiting Wei*, *Student Member, IEEE*, Chenpei Xie,
Mostafa Haghi, *Member, IEEE*, Nima TaheriNejad, *Member, IEEE*

***Abstract*—Cross-subject multimodal emotion recognition is challenged by inter-subject physiological variability and the continuous, ambiguous nature of the Arousal–Valence space. To address this, we propose a Hierarchical Graph-based Feature Processing (HGFP) framework that learns interpretable feature-importance weights by propagating supervision on a fixed intra-/inter-modality graph, enabling graph-guided re-weighting and selection, and we further adopt soft-label training to model uncertainty in continuous Arousal–Valence. On the ASCERTAIN dataset, under a strict subject-wise evaluation protocol, our method achieves 77.61%/0.77 (Arousal Acc./F1) and 74.35%/0.73 (Valence), and reaches 80.14%/0.86 and 75.63%/0.83 under 5-fold cross-validation, markedly outperforms other methods in the same category. Ablation studies show that re-weighting and selection each yield significant gains and are complementary, with their combination delivering the best performance. Interpretability analyses further indicate that the learned modality-importance differentiation is physiologically consistent, supporting that the model learns meaningful representations rather than relying on purely statistical fusion. Overall, the proposed method maintains strong performance and interpretability under the more stringent cross-subject setting, providing a transferable and reusable technical foundation for emotion-related health-state assessment and continuous monitoring. Its cross-individual generalization helps reduce the need for subject-specific labeling and frequent retraining, thereby lowering deployment and maintenance costs while improving accessibility and consistency across different populations.**

***Index Terms*—Multimodal signal, Emotion recognition, Hierarchical graph learning, Arousal–Valence, Feature Processing**

I. Introduction

Emotion is a fundamental component of human cognition and behavioral decision-making, exerting profound influence on mental health assessment, human–computer interaction, intelligent healthcare, and affective computing systems [1]. Accurate and objective recognition of individual emotional states not only facilitates a deeper understanding of internal psychological processes, but also provides critical support for the early detection and intervention of emotion-related disorders such as depression and anxiety [2]. Moreover, in intelligent interactive systems and wearable health monitoring devices, emotion recognition serves as a key enabler for personalized services and adaptive feedback. Consequently, developing robust emotion recognition models with strong generalization capability has become a central research problem in the field of affective computing.

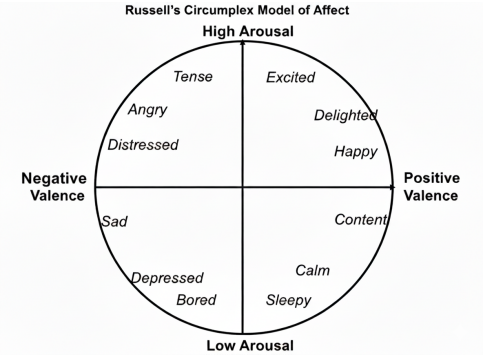

Fig. 1: Russell's circumplex model of affect.

Compared with discrete emotion classification models (e.g., happiness, sadness, and anger), continuous emotion representation frameworks demonstrate superior expressive power in both theory and practice. Among them, the Arousal–Valence two-dimensional emotion model has been widely adopted to characterize emotional intensity and emotional polarity [3], as illustrated in Fig. 1. Specifically, arousal reflects the level of physiological activation, while valence describes the degree of pleasantness or unpleasantness of an emotional experience [4]. This model enables a more fine-grained representation of complex and continuously evolving emotional states and exhibits strong consistency with theories in neuroscience and psychology, making it one of the most prevalent modeling paradigms in contemporary emotion recognition research.

In recent years, advances in wearable sensing technologies have driven a growing interest in emotion recognition based on multimodal physiological signals [5]. This trend is grounded in well-established physiological principles: emotional changes induce measurable responses in both the autonomic and central nervous systems [6]. Specifically, Electrocardiography (ECG) signals reflect the dynamic balance between sympathetic and parasympathetic nervous activity [7], where variations in emotional arousal are typically associated with significant changes in heart rate and heart rate variability; high-arousal emotional states are often accompanied by increased heart rate and reduced Heart rate variability (HRV). Galvanic Skin Response (GSR) directly captures sympathetic nervous system

This work has been partially funded by Hector Stiftung.

Yufan Du, Yiting Wei, Chenpei Xie, Mostafa Haghi and Nima TaheriNejad are with the EclectX Team, Institute of Computer Engineering, Heidelberg University, Heidelberg, Germany (*Corresponding author: Yiting Wei, email: yiting.wei@uni-heidelberg.de)

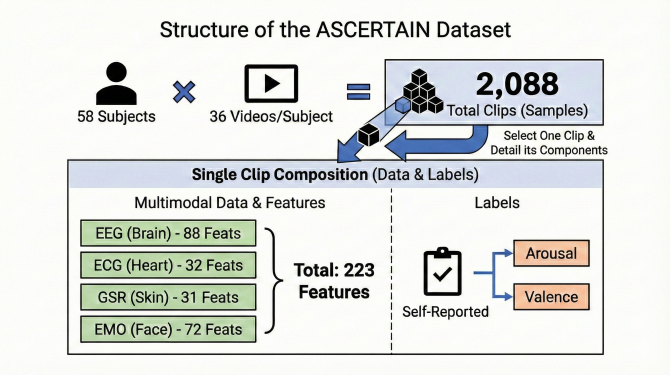

Fig. 2: Decription of ASCERTAIN dataset.

activation and is highly sensitive to emotional arousal, making it a widely used indicator for stress and emotional intensity assessment [8]. Electroencephalogram (EEG) signals provide insights into emotional processing at the neural level, where power variations across different frequency bands and brain regions exhibit stable associations with both arousal and valence [9]. Facial motion reflects emotion-related motor responses regulated by the central nervous system, with its intensity and temporal dynamics closely associated with arousal and valence [10]. As these physiological signals characterize emotional states from complementary systems and temporal scales, multimodal integration offers a solid foundation for building more comprehensive and robust emotion recognition models.

Despite substantial progress in physiological signal–based emotion recognition, existing approaches still face three key limitations. First, many studies rely on single-modality signals or simple multimodal feature concatenation, which fails to capture structured relationships across different physiological sources [11]. Second, most models are evaluated in within-subject or subject-dependent settings, limiting cross-subject generalization and hindering real-world deployment. Third, the majority of work adopts hard labels for emotion annotation, whereas emotional experiences are inherently continuous and ambiguous; as a result, hard labels often cannot adequately represent the underlying distribution and uncertainty of affective states. In contrast, soft labels provide a more natural way to model emotional uncertainty and individual variability, yet they remain insufficiently explored in emotion recognition.

To address these challenges, we propose a cross-subject emotion recognition framework based on Hierarchical Graph-based Feature Processing (HGFP). Our contributions are threefold: (1) We develop a hierarchical graph-based feature modeling scheme that captures high-order intra- and inter-modality relationships in multimodal physiological signals, yielding more discriminative representations. (2) We introduce a soft-label mechanism in the Arousal–Valence space to better model affective continuity and uncertainty, reducing the information loss of hard-label supervision. (3) We demonstrate strong cross-subject generalization, supporting scalable and reliable emotion-aware healthcare monitoring applications.

II. Related Work

Xu et al. proposed a joint model using fuzzy C-means and conditional random fields for emotion intensity and type recognition, but the method relies on manual annotation and may propagate intensity errors [12]. Skaramagkas et al. proposed a machine learning approach based on eye-tracking features to predict emotional arousal and valence, achieving high accuracy in binary classification, but showing reduced performance in multi-class settings when the neutral class is included [13]. He et al. proposed an EEG-based emotion recognition method using multivariate empirical mode decomposition (MEMD) to classify high/low arousal and high/low valence from multichannel IMF features, achieving competitive accuracy on the DEAP dataset, but with overall performance remaining at a moderate level [14]. Gannouni et al. proposed an adaptive EEG-based emotion recognition method under the valence–arousal–dominance model that improves classification accuracy by selecting subject-specific brain lobes and electrodes, but requires a complex multi-stage processing framework [15]. Greco et al. proposed a cvxEDA-based electrodermal activity method for recognizing arousal and valence induced by affective sounds, but the evaluation was limited to standardized stimuli in controlled settings [16]. Dresvyanskiy et al. proposed a multimodal audio–visual deep learning approach for arousal and valence estimation under noisy, in-the-wild conditions, achieving competitive performance, but showing limited generalization of multimodal fusion on the test set [17].

Existing emotion recognition methods span diverse modalities and models, yet they often rely on hard or manually annotated labels, fail to capture high-order multimodal physiological relationships, and degrade markedly in cross-subject settings where emotion is continuous and uncertain. This motivates methods that jointly model high-order feature relations, represent emotion continuity, and generalize robustly across subjects.

III. Materials and Methods

The overall pipeline of our proposed hierarchical graph-based framework for cross-subject emotion recognition is illustrated in Fig. 3. The framework first preprocesses multimodal physiological features, then constructs a hierarchical graph for HGFP-based feature weighting and selection. The selected features are finally used for soft-label SVR training and cross-subject emotion recognition.

A. Dataset

ASCERTAIN is a multimodal dataset for emotion and personality recognition, comprising multi-source physiological and behavioral data collected from 58 healthy subjects while they watched emotion-eliciting video clips in a controlled experimental environment [18]. As shown in Fig. 2, each subject viewed 36 movie clips designed to induce different emotional states, resulting in a total of 2,088 samples. During each clip, EEG, ECG, GSR, and Facial motion (EMO) were synchronously recorded, and statistical and spectral features were extracted from the raw signals. Emotional labels were

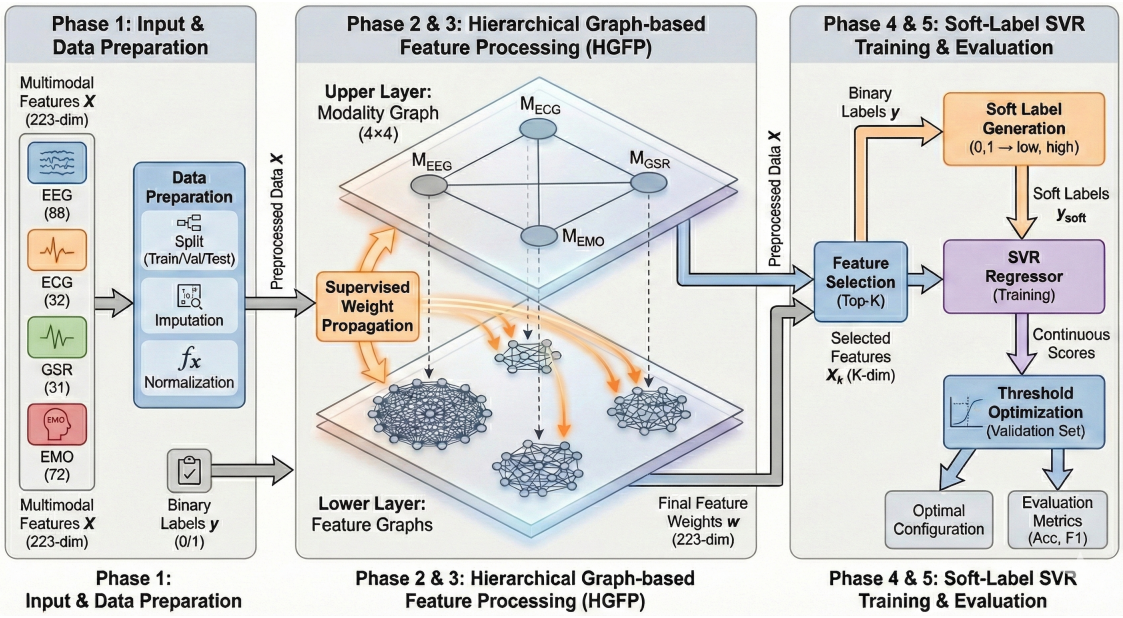

Fig. 3: Hierarchical graph-based framework for cross-subject emotion recognition.

obtained through self-report and annotated as continuous values in the two-dimensional arousal–valence emotion space, enabling the dataset to support multimodal emotion modeling and cross-subject emotion recognition studies.

B. Data Pre-processing

To ensure the quality of multimodal data, systematic data pre-processing was conducted on the ASCERTAIN dataset prior to model training. First, multimodal features including EEG, ECG, GSR, and facial expression data were loaded, and all samples were subjected to integrity checks. All feature matrices were scanned, and invalid values (e.g., positive or negative infinity) were uniformly treated as missing values.

Subsequently, feature-level cleaning was performed to remove non-informative features. Specifically, features with missing values across all samples were removed to ensure that the retained features had valid observations in the dataset. In parallel, sample-level inspection was conducted, and samples with all feature values missing were discarded.

To reduce the adverse impact of extreme outliers on model training stability, Winsorization was applied independently to each feature, where extreme values were truncated based on feature-wise distribution quantiles. Finally, global Z-score normalization was applied to the features to eliminate scale differences across modalities and improve training stability.

C. Extracted Features

At the feature level, ASCERTAIN provides four modalities with a total of 233 handcrafted features. Specifically, EEG (88 dimensions) includes per-channel statistics and descriptors of temporal dynamics derived from the multi-channel outputs of a single-electrode EEG device; ECG (32 dimensions) comprises statistical HRV measures and spectral power features computed from R-peaks/inter-beat intervals (IBI) and heart-rate (HR) sequences; GSR (31 dimensions) includes amplitude-, derivative-, and peak-based features, rise-time measures, as well as spectral power and zero-crossing rate; and EMO (72 dimensions) consists of statistical descriptors of displacement trajectories (mean, variance, skewness, kurtosis, and the proportion exceeding a threshold) computed from facial action units/keypoint movements.

D. Labeling Strategy

1) Cross-subject Data Partition: To avoid subject-dependent bias, a strict cross-subject data partitioning strategy is adopted. All subjects are split at the subject level into training, validation, and test sets with a ratio of 8:1:1, such that data from each subject appear in only one subset. This ensures that the model does not observe any samples from test subjects during training. Based on this partition, five-fold cross-validation is performed on the training and validation sets. In each fold, the model is trained exclusively on the training data, the validation set is used for selecting soft-label mapping parameters and decision thresholds, and the test set is reserved solely for final evaluation, thereby strictly preventing information leakage.

2) Hard Label Definition Based on Training Data: In this study, emotion recognition is formulated as a binary classification task along the Arousal and Valence dimensions. To avoid subjective threshold selection, an adaptive labeling strategy based on training data statistics is employed. For each cross-validation fold, the median of the continuous emotion scores in the training set is computed and used as the decision threshold. Samples with scores above the median are labeled as 1 (High), while the others are labeled as 0 (Low).

3) Soft Label Modeling and Parameter Selection: Although hard labels provide clear class assignments, emotional states are inherently continuous and ambiguous. As shown by the distance distribution analysis in Fig. 4, substantial overlap

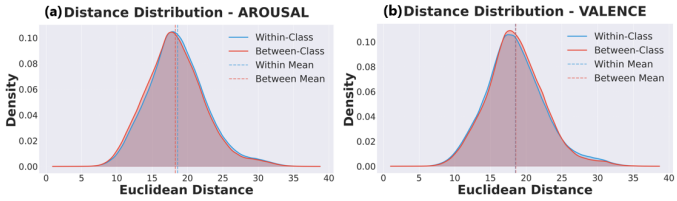

Fig. 4: Kernel density estimation (KDE) of Euclidean distance distributions for within-class and between-class samples in the (a) Arousal and (b) Valence dimensions.

exists between within-class and between-class samples in the feature space for both Arousal and Valence dimensions, indicating that the samples are not strictly separable. Consequently, directly applying hard labels may introduce boundary noise, leading to unstable model behavior or overfitting.

To address this issue, a soft-label strategy is introduced by mapping binary labels to continuous target values. Specifically, labels 0 and 1 are mapped to confidence-based value pairs within predefined intervals, such as $[0.1, 0.9]$ and $[0.2, 0.8]$. Different mapping configurations represent different levels of class confidence, thereby explicitly modeling uncertainty in the supervision signal. In implementation, multiple candidate soft-label mappings are evaluated. For each mapping configuration, a regression model is trained on the training set, and its continuous outputs are mapped back to binary predictions on the validation set using a tunable threshold. These predictions are compared with the corresponding hard labels to evaluate classification performance. The optimal soft-label mapping and decision threshold are jointly selected based on validation performance.

4) Final Prediction with Soft Labels: After determining the optimal soft-label mapping and decision threshold, the trained model is applied to the test set. The model first produces continuous regression outputs, which are subsequently converted into final binary predictions using the selected threshold. These predictions are reported as the final emotion recognition results on the test set.

E. Hierarchical Graph-based Feature Processing

1) Problem Definition and Notation: Assume a dataset consisting of N samples, where each sample is represented as $\mathbf{x}_n \in R^F$ with $F = 233$. The features are extracted from $M = 4$ physiological modalities, namely EEG, ECG, GSR, and EMO.

Let $\mathcal{F}_m$ denote the feature index set corresponding to the m-th modality, with dimensionality $F_m = |\mathcal{F}_m|$, satisfying $\sum_{m=1}^{M} F_m = F$. The feature matrix of modality m is denoted as $\mathbf{X}_m \in R^{N \times F_m}$.

2) Hierarchical Graph Construction: The hierarchical graph structure is constructed solely based on statistical relationships among features, without using any label information, and remains fixed across all cross-validation folds. The structure consists of a modality-level graph and modality-specific intra-modality feature graphs, as illustrated in Fig. 5.

a) Modality-level Graph Construction: To achieve unified modeling across different modalities, features within each

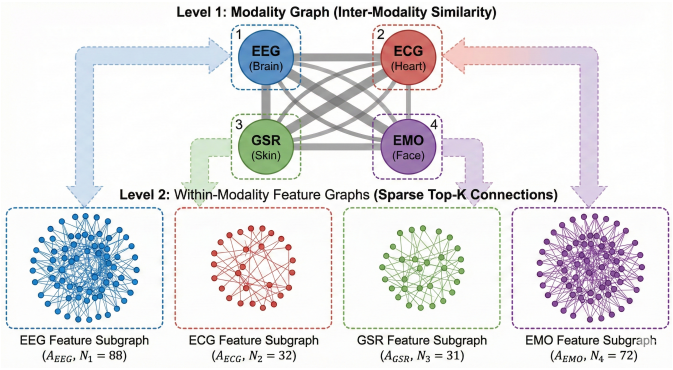

Fig. 5: Hierarchical graph construction.

modality are aggregated along the feature dimension. For modality m, the modality-level representation vector is defined as

$$r_m[n] = \frac{1}{F_m} \sum_{j=1}^{F_m} X_m[n, j], \quad n = 1, \ldots, N, \tag{1}$$

yielding $\mathbf{r}_m \in R^N$. This representation preserves the overall variation trend of each modality across samples while eliminating inconsistencies caused by different feature dimensionalities.

The modality-level graph characterizes statistical similarity between physiological modalities. Thus, this graph describes how strongly different physiological modalities vary together across samples. For any two modalities m and k, the edge weight is defined using Pearson correlation, $A_{\text{mod}}(m, k) = \text{corr}(\mathbf{r}_m, \mathbf{r}_k)$, and its absolute value is adopted to focus on correlation strength. The resulting adjacency matrix is $\mathbf{A}_{\text{mod}} \in R^{M \times M}$.

To ensure numerical stability, symmetric normalization is applied:

$$\tilde{\mathbf{A}}_{\text{mod}} = \mathbf{D}^{-1/2} \mathbf{A}_{\text{mod}} \mathbf{D}^{-1/2}, \tag{2}$$

where $\mathbf{D}$ denotes the degree matrix.

b) Intra-modality Feature Graph Construction: For each modality, an intra-modality feature graph is constructed to model statistical relationships among features within the same modality. In this lower layer, each feature is treated as a graph node, and the Pearson correlation between features is computed along the sample dimension as $C_m(i, j) = \text{corr}(\mathbf{X}_m[:, i], \mathbf{X}_m[:, j])$, where the absolute value is taken and diagonal elements are set to zero.

Since C_m forms a fully connected graph, a Top-K sparsification strategy is applied to reduce redundancy and noise. For each feature node, only the K strongest connections are retained, yielding a sparse adjacency matrix $\mathbf{A}_m$.

As Top-K sparsification may introduce asymmetry, $\mathbf{A}_m$ is symmetrized and normalized as

$$\tilde{\mathbf{A}}_m = \mathbf{D}_m^{-1/2} \mathbf{A}_m \mathbf{D}_m^{-1/2}, \tag{3}$$

where $\mathbf{D}_m$ is the corresponding degree matrix.

3) Fold-wise Supervised Feature Importance Propagation: Given the fixed hierarchical graph structure, feature importance is learned independently for each cross-validation fold.

All statistics involving supervision are computed exclusively based on the training data of the current fold, while test data are strictly excluded to avoid information leakage. Specifically, we employ a residual one-step propagation scheme on the fixed graphs, in which the propagated importance is the sum of the original seed scores and a diffusion term.

a) Initial Feature Importance: For each feature i, its initial importance score q_i is computed solely from the training data $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$ to measure its discriminative ability. All seed scores form a vector $\mathbf{q} \in R^F$, which is further partitioned into modality-specific subvectors $\mathbf{q}_m \in R^{F_m}$.

b) Intra-modality Feature Importance Propagation: Within each modality, seed scores are propagated once over the intra-modality feature graph using a seed-preserving formulation:

$$\mathbf{s}_m = \mathbf{q}_m + \beta_{\text{feat}} \tilde{\mathbf{A}}_m \mathbf{q}_m, \tag{4}$$

where β_{feat} controls the intra-modality diffusion strength and the first term preserves the original seed evidence.

c) Modality-level Importance Propagation: Feature importance within each modality is aggregated as $g_m = \text{Agg}(\mathbf{s}_m)$, and all modality scores are stacked into $\mathbf{g} \in R^M$. Propagation on the modality-level graph is then performed as

$$\mathbf{w} = \mathbf{g} + \beta_{\text{mod}} \tilde{\mathbf{A}}_{\text{mod}} \mathbf{g}, \tag{5}$$

where β_{mod} controls the diffusion strength on the modality graph, yielding the modality weight vector $\mathbf{w} \in R^M$.

d) Feature Importance Fusion: For a feature i belonging to modality m, the final importance is defined as

$$W_i = w_m \cdot s_{m,i}. \tag{6}$$

All feature importance scores are normalized to obtain the final weight vector $\mathbf{W} \in R^F$.

4) Feature Re-weighting and Selection Based on Feature Importance: After obtaining the final feature weight vector $\mathbf{W}$, features are re-weighted and selected to construct the final representation for downstream models.

a) Feature Re-weighting: For any sample feature vector $\mathbf{x} \in R^F$, element-wise re-weighting is applied as

$$\mathbf{x}' = \mathbf{x} \odot \sqrt{\mathbf{W}}, \tag{7}$$

where $\odot$ denotes the Hadamard product.

b) Weight-based Feature Selection (Threshold-then-Top-K): A candidate feature set is constructed using a threshold τ as $\mathcal{S}_\tau = \{i \mid W_i \geq \tau\}$. If $|\mathcal{S}_\tau| \geq K$, the final feature subset is obtained by selecting the K features with the largest weights; otherwise, the global Top-K strategy is applied.

c) Output Feature Representation: After re-weighting and selection, the reduced training and test feature matrices are obtained as $\mathbf{X}_{\text{train,sel}}$ and $\mathbf{X}_{\text{test,sel}}$.

F. Evaluation Method

The proposed method is evaluated under a strict cross-subject setting to assess its generalization performance on unseen subjects. Five-fold cross-validation is adopted during the training and validation stages to ensure robust model selection and parameter tuning. The performance of the model is quantitatively evaluated using Accuracy and F1-score, which measure overall classification correctness and robustness under potential class imbalance, respectively.

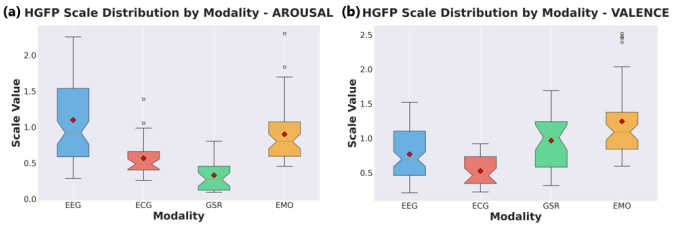

Fig. 6: HGFP scale distributions across modalities for (a) arousal and (b) valence.

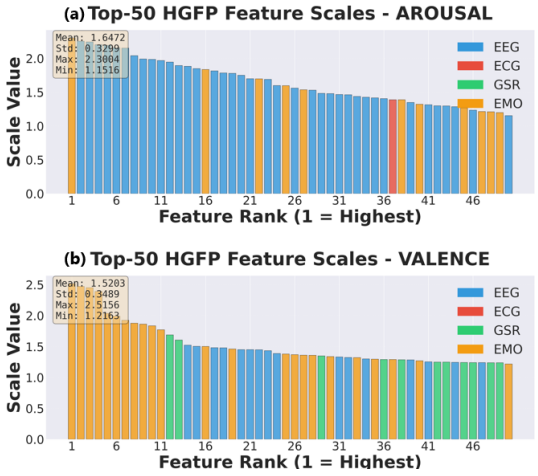

Fig. 7: Top-50 HGFP feature scales for (a) arousal and (b) valence, ranked by scale magnitude.

IV. Results and Discussion

A. Comprehensive Comparison with Published Methods

As shown in Table I, we conduct a comprehensive comparison between our method and representative published approaches. Importantly, all results reported in the table are taken from publicly reported evaluations on the same dataset, ensuring dataset-level comparability. Nevertheless, the literature adopts markedly different evaluation protocols, including subject-dependent random splits, standard 5-fold cross-validation (CV), LOVO, and less stringent transductive settings. In general, random-split or subject-dependent evaluations tend to yield higher scores, whereas subject-wise (subject-independent) cross-subject evaluation is more challenging due to stronger inter-subject variability and distribution shift, and is therefore more aligned with real-world deployment. Consequently, high performance under a relatively relaxed protocol alone does not necessarily imply strong cross-subject generalization.

Within this context, our advantage is twofold. First, under the relatively relaxed random-split 5-fold CV setting, our method already achieves leading performance (Arousal: 80.14% / 0.86; Valence: 75.63% / 0.83 in Acc./F1), demonstrating strong discriminative capability under conventional evaluation. More importantly, under the substantially stricter

TABLE I: COMPARISON WITH EXISTING METHODS ON ASCERTAIN FOR BINARY AROUSAL AND VALENCE RECOGNITION. ACC. IS REPORTED IN %, AND F1 IS THE F1-SCORE.

Src.	Method	Arousal		Valence		SI	Validation
		Acc.(%)	F1	Acc.(%)	F1		
[18]	NB / SVM + Decision Fusion	–	0.69	–	0.71	–	LOVO
[19]	GCN	76.67	–	72.4	–	–	5-fold CV
[20]	VM2HL	72.54	–	68.53	–	–	Subject-dependent 50/50 split
[21]	VGLCN	70.1	–	73.26	–	–	5-fold CV
[22]	SVM / KNN / RF	71.79	–	69.11	–	–	5-fold CV
[19]	LocalGSL	59.81	–	72.72	–	–	Transductive eval.
Ours	**HGFP**	**80.14**	**0.86**	**75.63**	**0.83**	–	5-fold CV
Ours	**HGFP**	**77.61**	**0.77**	**74.35**	**0.73**	**Yes**	5-fold CV

cross-subject subject-wise (subject-independent) evaluation, our method remains the best or achieves a clear lead (Arousal: 77.61% / 0.77; Valence: 74.35% / 0.73). This indicates that our gains are not driven by potential advantages of random splitting, but persist in a more challenging generalization scenario, reflecting robust modeling of inter-subject differences.

Overall, when comparing on the same dataset while accounting for the strictness of evaluation protocols, our method consistently remains competitive and shows its most pronounced advantage under the stricter cross-subject setting. These results further validate the effectiveness and practical value of the proposed hierarchical graph-driven feature processing and soft-label training mechanism for cross-subject multimodal emotion recognition.

B. Ablation Study: Contributions of Graph-driven Re-weighting and Selection

Table II reports an ablation study under the subject-wise evaluation protocol to quantify the individual contributions and the synergy of the proposed feature processing modules. The SVR-only baseline, which applies no graph-based feature processing, achieves only 56.11%/0.53 on Arousal and 54.31%/0.52 on Valence (Acc./F1), indicating that directly learning from the raw features is insufficient for robust cross-subject generalization. Incorporating feature re-weighting only (without selection) yields a substantial improvement to 66.38%/0.64 (Arousal) and 64.52%/0.65 (Valence), where the re-weighting coefficients are learned by our hierarchical graph mechanism and used to adaptively rescale features, suggesting that the graph-derived weights effectively emphasize emotion-relevant information while suppressing redundant noise. Applying feature selection only (without re-weighting) further improves performance to 71.04%/0.72 on Arousal and 66.83%/0.64 on Valence; this variant ranks features using the same graph-derived weights and removes those below a threshold of 0.8, validating the discriminative utility of the graph-guided ranking. Finally, combining both re-weighting and selection in our full method (HGFP graph with soft-label SVR) delivers the best results, reaching 77.61%/0.77 on Arousal and 74.35%/0.73 on Valence. Overall, the results demonstrate that graph-driven re-weighting and selection provide complementary benefits: re-weighting improves the transferability and robustness of feature representations, while selection reduces redundancy and sharpens the decision boundary, and their integration maximizes performance gains in cross-subject emotion prediction.

C. Comparison with Alternative Feature Processing Strategies

Table III compares several feature processing strategies for Arousal and Valence prediction. To ensure a fair comparison, we keep the modeling component strictly unchanged across all settings, including the same data split, the same predictor configuration, and the same training and evaluation protocol; thus, the observed performance differences are attributable solely to the feature processing strategy. Methods based on feature selection alone yield limited performance, with Arousal accuracy around 62% and Valence accuracy around 59%. Introducing re-weighting-based baselines provides only marginal gains. In contrast, our proposed HGFP-based feature re-weighting and selection with soft-label SVR achieves the best results on both dimensions, reaching 77.61% accuracy and 0.77 F1 on Arousal, and 74.35% accuracy and 0.73 F1 on Valence. These consistent improvements indicate that HGFP learns more effective and transferable feature importance patterns under cross-subject settings, thereby enhancing prediction performance without altering the downstream model.

D. Physiologically Consistent Modality-Importance Differentiation

As shown in Figs. 6 and 7, the feature-scale distributions learned by HGFP and the Top-50 feature-scale rankings reveal a clear dimension–modality differentiation. For Arousal, high-scale features are predominantly contributed by EEG, indicating that the model relies more on central neural activity to discriminate between low and high arousal. In contrast, for Valence, EMO (facial/expression-related modality) features appear more frequently among the top-ranked features and exhibit higher overall scales, suggesting that the model tends to leverage overt expressive cues to distinguish positive versus negative valence. This pattern is consistent with established psychophysiological mechanisms: arousal is more directly linked to changes in cortical excitability and vigilance, to which EEG is particularly sensitive, whereas valence is more readily conveyed through facial expressions and semantically interpretable behavioral cues, making EMO a more stable contributor under cross-subject settings [23]. Overall, these results suggest that HGFP goes beyond purely statistical fusion by learning modality selection and weighting that align with underlying physiological/behavioral representations, thereby providing interpretable evidence for the effectiveness of the proposed method.

V. CONCLUSION

This paper proposes a hierarchical graph-driven framework for cross-subject multimodal emotion recognition. By

TABLE II: ABLATION STUDY UNDER SUBJECT-WISE EVALUATION. ACC. IS REPORTED IN %, AND F1 IS THE F1-SCORE.

Variant	ReW	FS	Arousal		Valence	
			Acc.(%)	F1	Acc.(%)	F1
SVR-only	–	–	56.11	0.53	54.31	0.52
+ Re-weighting	Yes	–	66.38	0.64	64.52	0.65
+ Feature selection	–	Yes	71.04	0.72	66.83	0.64
Ours	**Yes**	**Yes**	**77.61**	**0.77**	**74.35**	**0.73**

TABLE III: COMPARISON OF FEATURE PROCESSING STRATEGIES FOR AROUSAL AND VALENCE PREDICTION. ACC. IS REPORTED IN %, AND F1 IS THE F1-SCORE.

Combination	Methods	Arousal		Valence	
		Acc.(%)	F1	Acc.(%)	F1
Feature selection	Top-k Fisher-score	62.31	0.66	59.28	0.63
	Top-k Random Forest	62.79	0.65	59.81	0.54
Feature re-weighting + selection	Logreg + β	63.15	0.65	59.43	0.62
	Linear-SVM + β	62.08	0.62	58.53	0.62
Ours (Feature re-weighting + selection)	**HGFP + Soft-Label SVR**	**77.61**	**0.77**	**74.35**	**0.73**

propagating supervision through a hierarchical graph, it captures high-order intra-/inter-modality relations and produces interpretable weights for graph-guided feature re-weighting and selection (HGFP). We further adopt soft-label training in the continuous Arousal–Valence space to model affective continuity and uncertainty, improving cross-subject robustness.

On ASCERTAIN, under a strict subject-wise (subject-independent) protocol, our method achieves leading and stable performance on both Arousal and Valence, outperforming prior reports on the same dataset. Ablations confirm that graph-based re-weighting and selection each provide significant gains and are complementary. Interpretability analyses also reveal physiologically consistent modality importance, with EEG contributing more to Arousal and EMO to Valence, indicating meaningful learning beyond purely statistical fusion.

Overall, the framework delivers strong, interpretable cross-subject transfer, reducing reliance on subject-specific labeling and frequent retraining for emotion-related health-state assessment and continuous monitoring. Future work will extend validation to larger, more diverse cohorts and more realistic settings, and explore lightweight implementations for deployment.

REFERENCES

[1] E. Hudlicka, "Computational modeling of cognition–emotion interactions: Theoretical and practical relevance for behavioral healthcare," in *Emotions and affect in human factors and human-computer interaction*. Elsevier, 2017, pp. 383–436.

[2] C. A. Mazefsky, J. Herrington, M. Siegel, A. Scarpa, B. B. Maddox, L. Scahill, and S. W. White, "The role of emotion regulation in autism spectrum disorder," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 52, no. 7, pp. 679–688, 2013.

[3] J. Pinto, A. Fred, and H. P. Da Silva, "Biosignal-based multimodal emotion recognition in a valence-arousal affective framework applied to immersive video visualization," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 3577–3583.

[4] G. Colombetti and P. Kuppens, "How should we understand valence, arousal, and their relation?" in *Emotion Theory: The Routledge Comprehensive Guide*. Routledge, 2024, pp. 599–620.

[5] P. S. Kumar, P. K. Govarthan, A. A. S. Gadda, N. Ganapathy, and J. F. A. Ronickom, "Deep learning-based automated emotion recognition using multimodal physiological signals and time-frequency methods," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024.

[6] E. Molefi, I. McLoughlin, and R. Palaniappan, "Heart rate variability responses to visually induced motion sickness," in *2023 45th annual international conference of the IEEE engineering in medicine & biology society (EMBC)*. IEEE, 2023, pp. 1–4.

[7] J.-M. Grégoire, C. Gilon, S. Carlier, and H. Bersini, "Autonomic nervous system assessment using heart rate variability," *Acta cardiologica*, vol. 78, no. 6, pp. 648–662, 2023.

[8] M. Z. Baig and M. Kavakli, "A survey on psycho-physiological analysis & measurement methods in multimodal systems," *Multimodal Technologies and Interaction*, vol. 3, no. 2, p. 37, 2019.

[9] M. Zangeneh Soroush, K. Maghooli, S. Kamaledin Setarehdan, and A. M. Nasrabadi, "A review on eeg signals based emotion recognition," *International Clinical Neuroscience Journal*, vol. 4, no. 4, pp. 118–129, 2017.

[10] E. G. Krumhuber, L. I. Skora, H. C. Hill, and K. Lander, "The role of facial movements in emotion recognition," *Nature Reviews Psychology*, vol. 2, no. 5, pp. 283–296, 2023.

[11] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 6, p. e1563, 2024.

[12] M. Xu, C. Xu, X. He, J. S. Jin, S. Luo, and Y. Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Processing*, vol. 93, no. 8, pp. 2140–2150, 2013.

[13] V. Skaramagkas, E. Ktistakis, D. Manousos, N. S. Tachos, E. Kazantzaki, E. E. Tripoliti, D. I. Fotiadis, and M. Tsiknakis, "A machine learning approach to predict emotional arousal and valence from gaze extracted features," in *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2021, pp. 1–5.

[14] Y. He, Q. Ai, and K. Chen, "A memd method of human emotion recognition based on valence-arousal model," in *2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 2. IEEE, 2017, pp. 399–402.

[15] S. Gannouni, A. Aledaily, K. Belwafi, and H. Aboalsamh, "Adaptive emotion detection using the valence-arousal-dominance model and eeg brain rhythmic activity changes in relevant brain lobes," *IEEE Access*, vol. 8, pp. 67 444–67 455, 2020.

[16] A. Greco, G. Valenza, L. Citi, and E. P. Scilingo, "Arousal and valence recognition of affective sounds based on electrodermal activity," *IEEE Sensors Journal*, vol. 17, no. 3, pp. 716–725, 2016.

[17] D. Dresvyanskiy, M. Markitantov, J. Yu, H. Kaya, and A. Karpov, "Multi-modal arousal and valence estimation under noisy conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4773–4783.

[18] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.

[19] W. S. Chien, M. Y. Tsai, and C. C. Lee, "Graph structure learning with local connectivity refinement for improved physiological emotion recognition," in *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2025, pp. 1–6.

[20] S. Zhao, G. Ding, J. Han, and Y. Gao, "Personality-aware personalized emotion recognition from physiological signals." in *IJCAI*, 2018, pp. 1660–1667.

[21] W.-S. Chien, H.-C. Yang, and C.-C. Lee, "Cross corpus physiological-based emotion recognition using a learnable visual semantic graph convolutional network," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2999–3006.

[22] K. Yang, B. Tag, Y. Gu, C. Wang, T. Dingler, G. Wadley, and J. Goncalves, "Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 562–570.

[23] J. T. Coull, "Neural correlates of attention and arousal: insights from electrophysiology, functional neuroimaging and psychopharmacology," *Progress in neurobiology*, vol. 55, no. 4, pp. 343–361, 1998.