Symmetry-Aware Structured Representation Learning for Unified Multi-Modal Physiological Modeling in Affective State and Preference Inference

Wenli Qu and Mu-Jiang-Shan Wang *

Posted Date: 27 February 2026

doi: 10.20944/preprints202602.1827.v1

Keywords: symmetry-aware learning; structured representation; multi-modal symmetry modeling; physiological signal analysis; affective computing; EEG and peripheral signals; token-based representation learning; music therapy

 doi:10.20944/preprints202602.1827.v1

Article

Symmetry-Aware Structured Representation Learning for Unified Multi-Modal Physiological Modeling in Affective State and Preference Inference

Wenli Qu [1] and Mu-Jiang-Shan Wang [2,3,*]

[1] College of Arts Management, Shandong University of Arts, Jinan 250300, China
[2] Shenzhen Kaihong Digital Industry Development Co., Ltd., Shenzhen, China
[3] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
* Correspondence: mjs.wang@siat.ac.cn

Abstract

Decoding affective states and personal preferences from physiological responses remains a fundamental challenge in affective computing due to strong heterogeneity across neural, autonomic, and attentional signals, as well as the coupling between transient emotions and long-term preferences. Most existing methods address these factors independently and lack explicit mechanisms to preserve the intrinsic structural regularities and invariances of physiological affective responses, limiting their applicability in real-world scenarios such as music therapy. In this paper, we propose a symmetry-aware and structured multi-modal physiological modeling framework for joint affective state and preference inference. The framework integrates electroencephalography (EEG), peripheral physiological signals (GSR, BVP, EMG, respiration, and temperature), and eye-movement data (EOG) within a unified temporal modeling paradigm. At its core, a Dynamic Token Feature Extractor (DTFE) converts raw physiological time series into compact token representations without handcrafted features, and explicitly decomposes representation learning into cross-series symmetry and intra-series symmetry. These two complementary symmetry dimensions are realized through Cross-Series Intersection (CSI) and Intra-Series Intersection (ISI) mechanisms, enabling structured and interpretable physiological representations. A hierarchical cross-modal fusion strategy further integrates modality-level tokens in a symmetry-consistent manner, capturing dependencies among neural, autonomic, and attentional modalities. Extensive experiments on the DEAP dataset demonstrate consistent improvements over state-of-the-art methods under both single-task and multi-task settings. The proposed model achieves 98.32% and 98.45% accuracy for valence and arousal prediction, respectively, and 97.96% accuracy for quadrant-based emotion classification in single-task evaluation, while attaining 92.8%, 91.8%, and 93.6% accuracy for valence, arousal, and liking prediction in joint multi-task settings. Additional robustness analyses under reduced training data confirm that symmetry-aware structured decomposition improves data efficiency and generalization. Overall, this work establishes a principled symmetry-preserving representation learning framework for robust affective decoding and intelligent, feedback-driven music therapy systems.

Keywords: symmetry-aware learning; structured representation; multi-modal symmetry modeling; physiological signal analysis; affective computing; EEG and peripheral signals; token-based representation learning; music therapy

1. Introduction

Music is a powerful medium capable of eliciting rich emotional experiences and modulating human affective states. In clinical and therapeutic contexts, music has been widely adopted as a non-invasive intervention for emotional regulation, cognitive rehabilitation, and mental health treatment, particularly for conditions such as depression, anxiety, and stress-related disorders. Beyond

 doi:10.20944/preprints202602.1827.v1

therapy, understanding emotional responses to music also plays a crucial role in personalized recommendation systems, human–computer interaction, and affect-aware intelligent systems. Accurately decoding affective states and personal preferences from physiological responses has therefore become a fundamental problem in affective computing and computational neurophysiology [1,2].

Emotions induced by music are inherently complex and multidimensional. Rather than being expressed through a single observable channel, affective responses emerge from the coordinated activity of multiple physiological and neurological systems. Electroencephalography (EEG) reflects cortical dynamics and neural oscillatory patterns associated with emotional perception and cognitive appraisal, while peripheral physiological signals such as galvanic skin response (GSR), blood volume pulse (BVP), electromyography (EMG), respiration, and skin temperature capture autonomic nervous system responses related to arousal, stress, and valence. Eye-movement signals (EOG) further provide valuable cues about attention allocation and cognitive engagement. These heterogeneous modalities jointly encode complementary aspects of emotional experience, making multi-modal physiological analysis a promising yet challenging research direction.

Despite substantial progress, existing emotion recognition approaches still exhibit several limitations. Early studies and classical methods primarily rely on handcrafted features and shallow classifiers, often focusing on a single modality or a limited subset of signals [3,4]. More recent deep learning-based approaches improve representation capacity but frequently adopt coarse-grained emotion formulations, such as binary classification or low-cardinality categorical schemes (e.g., valence–arousal quadrants) [5,6]. While effective in simplified settings, such formulations fail to capture the fine-grained variability of emotional intensity and subjective preference, which is particularly critical in music therapy scenarios.

Another important limitation lies in task formulation and representation structure. Most prior works treat affective dimensions—such as valence, arousal, or preference—as independent prediction problems and optimize them in isolation [7,8]. However, extensive psychological and neuroscientific evidence suggests that these dimensions are strongly interrelated. Ignoring such interdependencies not only limits expressive power, but also undermines generalization under subject variability and data perturbations. Similar challenges have been extensively studied in structured systems and network theory, where preserving connectivity, diagnosability, and functional consistency under disturbances is known to rely on symmetry-aware and structure-preserving designs [9–12]. Related studies on symmetric network structures further demonstrate that higher-order connectivity, conditional fault tolerance, and tightly constrained diagnosability are critical for maintaining system functionality under node or link failures, as shown in locally twisted cubes, Cayley graphs, and high-dimensional hypercube variants [13–15].

From a representation learning perspective, physiological affective modeling can therefore be viewed as a structured system inference problem, where robustness and generalization depend on preserving intrinsic structural regularities and invariant interaction patterns across channels and modalities. Nevertheless, most existing multi-modal models rely on straightforward feature concatenation or shallow fusion strategies, which are insufficient to capture complex cross-modal dependencies among neural, autonomic, and attentional signals [16,17].

To address these challenges, we propose a unified multi-task framework for fine-grained emotion and preference recognition from multi-modal physiological signals, with a particular focus on music therapy applications. Our framework jointly processes EEG, peripheral physiological signals (GSR, BVP, EMG, respiration, and temperature), and eye-movement data (EOG), enabling comprehensive modeling of affective responses across cortical, autonomic, and behavioral domains. Unlike previous approaches that predominantly employ binary or coarse-grained classification schemes [18,19], our method supports fine-grained 9-class prediction for each affective dimension, including valence, arousal, and liking.

At the core of the proposed framework lies a Dynamic Token Feature Extractor (DTFE), which transforms raw physiological time series into compact and discriminative token representations. DTFE

 doi:10.20944/preprints202602.1827.v1

explicitly decomposes representation learning into two complementary structural symmetry dimensions: *cross-series symmetry*, which models invariant interaction patterns among multiple physiological channels, and *intra-series symmetry*, which captures recurrent temporal–spectral structures within individual signals. On top of modality-specific processing, we further introduce a hierarchical cross-modal fusion mechanism that integrates modality-level representations in a symmetry-consistent manner, enabling synergistic affective reasoning across heterogeneous modalities. The entire framework is trained end-to-end under a unified multi-task learning paradigm, allowing shared representations to be jointly optimized across related affective objectives. Building upon a symmetry-aware perspective of multi-modal physiological affective modeling, the main contributions of this work are summarized as follows:

- **First structured multi-task affective modeling paradigm**: We propose the first unified and structured modeling paradigm that jointly infers emotional valence, arousal, and music liking from physiological signals. By explicitly encoding inter-task structural dependencies, this principled formulation enforces shared invariances across affective dimensions and yields more robust and accurate affective decoding than conventional single-task or loosely coupled approaches.
- **Novel symmetry-preserving token representation**: We introduce a novel and principled token-based representation learning strategy for physiological time series, in which affective representations are explicitly decomposed into cross-series and intra-series structural symmetry components. This structured decomposition is instantiated through Cross-Series Intersection (CSI) and Intra-Series Intersection (ISI), enabling invariant modeling of both inter-channel interactions and intra-channel temporal–spectral dynamics.
- **Principled hierarchical cross-modal integration**: We propose a principled and structured hierarchical fusion mechanism that integrates neural, autonomic, and attentional modalities in a symmetry-consistent manner. This design preserves modality-level structural regularities while enabling deep and interpretable cross-modal interaction, establishing a coherent multi-modal affective representation space.
- **State-of-the-art performance with structured generalization**: Extensive experiments on the DEAP dataset demonstrate that the proposed structured modeling paradigm achieves state-of-the-art performance under both single-task and multi-task learning settings. The model attains 98.32%, 98.45%, and 97.96% accuracy for valence, arousal, and quadrant classification in single-task evaluation, and maintains superior joint multi-task performance with 92.8%, 91.8%, and 93.6% accuracy for valence, arousal, and liking prediction, respectively.

By advancing symmetry-aware, structured representation learning for physiological signals, this work addresses fundamental limitations of prior affective computing methods and establishes a principled foundation for robust affective inference, intelligent music therapy, and affect-aware human–computer interaction systems.

2. Related Work

2.1. Emotion Recognition Based on EEG Signals

Electroencephalography (EEG) is one of the most widely used modalities for emotion recognition due to its high temporal resolution and direct correlation with neural activity. Existing EEG-based emotion recognition methods can be broadly categorized into handcrafted feature approaches and deep representation learning models.

Early studies predominantly relied on handcrafted features extracted from EEG signals, such as frequency-domain power spectra, time-domain statistics, and entropy-based descriptors, followed by conventional classifiers for binary or multi-class emotion recognition [1,2,20]. These approaches are computationally efficient and interpretable, but their performance is highly dependent on domain-specific feature engineering and tends to degrade under cross-subject variability.

To overcome these limitations, more sophisticated learning-based models have been proposed. Hierarchical neural architectures [1], graph-regularized sparse models [21], and deep forest frame-

 doi:10.20944/preprints202602.1827.v1

works [22] were introduced to capture spatial and temporal dependencies among EEG channels. More recent advances further incorporate attention mechanisms to enhance representation capacity. For example, Song et al. [23] proposed a dynamical graph convolutional neural network to model inter-channel relationships, while Zhang et al. [24] introduced hierarchical self-attention to localize emotion-relevant temporal segments within EEG signals. Models such as GLFANet [25] and DAST [26] additionally integrate spatio-temporal attention and domain adaptation strategies to improve generalization across subjects.

Beyond standard emotion recognition settings, EEG-based models have also been explored in special populations, such as patients with disorders of consciousness [27], as well as in real-time brain–computer interface (BCI) applications [28]. While these studies demonstrate the expressive power of EEG for affective analysis, most existing approaches implicitly entangle inter-channel interactions and intra-channel temporal dynamics within monolithic representations, making them sensitive to individual differences and EEG non-stationarity.

2.2. Emotion Recognition by Merging Multiple Physiological Signals

To address the inherent limitations of single-modality systems, an increasing body of work has investigated multimodal emotion recognition by integrating EEG with peripheral physiological signals, such as galvanic skin response (GSR), blood volume pulse (BVP), electromyography (EMG), respiration, and temperature. These approaches aim to exploit the complementary characteristics of cortical, autonomic, and behavioral responses to improve robustness and recognition accuracy.

Early multimodal methods primarily adopted ensemble learning or feature-level fusion. Representative examples include IRS [3] and MESAE [6], which combine ECG, GSR, EMG, and other signals using ensemble or attention-based architectures. More recent studies leverage deep learning to perform joint representation learning across modalities. Liu et al. [16] employed deep canonical correlation analysis (DCCA) to learn shared embeddings from EEG, GSR, and eye-movement signals, while Tang et al. [18] proposed a hierarchical fusion framework with contrastive alignment to enhance cross-modal consistency. Other approaches, such as BDAE [5], MM-ResLSTM [7], and FGSVM [8], utilize deep residual networks, recurrent encoders, or kernel-based fusion strategies to jointly model neural and peripheral physiological modalities.

Despite these advances, most multimodal emotion recognition models are formulated as single-task classification pipelines with limited output granularity, typically focusing on binary or low-cardinality emotion categories. Only a few methods, such as DEMA [19], support more fine-grained or multi-label emotion prediction. However, even these approaches primarily optimize individual affective dimensions in isolation and do not explicitly model the hierarchical structure or interdependencies among multiple affective attributes.

2.3. Summary and Motivation

In summary, existing studies on physiological emotion recognition—whether based on EEG alone or on multimodal signal fusion—have demonstrated the feasibility of decoding affective states from neural and peripheral responses. However, from a representation learning perspective, most prior approaches remain largely task-driven and data-driven, with limited emphasis on the intrinsic structural regularities and symmetries underlying physiological affective responses. As summarized in Table 1, the majority of existing methods are constrained to single-task settings and coarse-grained label spaces, such as binary or 4/5-class emotion schemes, which restrict both the expressiveness and interpretability of affective representations in complex scenarios such as music therapy.

More critically, few prior models are designed to simultaneously handle multiple interdependent affective dimensions, such as valence, arousal, and liking, within a unified learning framework that preserves their underlying structural relationships. From a symmetry-aware modeling viewpoint, existing deep learning and attention-based approaches typically entangle cross-channel interactions and intra-channel temporal–spectral dynamics in an implicit manner, without an explicit decomposition into complementary structural components. As a result, important invariances—such as recurrent

 doi:10.20944/preprints202602.1827.v1

temporal patterns within a channel and symmetric interaction structures across channels—are not explicitly preserved. This lack of structured and symmetry-aware modeling limits interpretability and degrades generalization under subject variability and multimodal heterogeneity.

In contrast, the proposed method is explicitly formulated from a structured and symmetry-aware perspective. It supports fine-grained 9-class prediction for each affective dimension within a unified multi-task learning framework, while preserving intrinsic regularities across modalities and tasks. By jointly modeling EEG, GSR, BVP, EMG, respiration, temperature, and EOG signals, and by explicitly decoupling cross-series interaction symmetry from intra-series temporal–spectral symmetry, the proposed framework overcomes the representational bottlenecks of prior works. This principled decomposition enables high-resolution, multi-dimensional emotion and preference estimation with improved interpretability and generalization.

Table 1. Comparison of recent multi-physiological emotion recognition methods in terms of task structure, emotion categories, and input modalities.

Method	Task Type	Emotion Categories	Modalities
SVM [4]	Single	5-class	EEG, EDA, GSR, SCR, skin temp
MT-MKL [29]	Single	2-class	EEG, GSR, RB, skin temp
IRS [3]	Single	4-class	ECG, GSR, PPG
BDAE [5]	Single	2-class	EEG, eye movement
Bimodal-LSTM [30]	Single	2-class	EDA, PPG, EMG
MESAE [6]	Single	5-class	EEG, EOG, EMG, GSR, temp, BP
DCCA [16]	Single	2/4-class	EEG, eye, GSR, EMG, PPG
MM-ResLSTM [7]	Single	2-class	EEG, peripheral signals
FGSVM [8]	Single	5-class	EDA, PPG, EMG
DPAN [31]	Single	2-class	EDA, PPG
Random Forest [32]	Single	2-class	EMG, EOG
RDFKM [33]	Single	2-class	EEG, EMG, GSR, RES
i-Isomap+DCNN [17]	Single	4-class	EEG, peripheral, eye
RHRPNet [18]	Single	2/4-class	EEG, peripheral signals
DEMA [19]	Single	2/5-class	EEG, GSR, BP, RB
Our Method	**Multi-task**	**3×9-class**	**EEG, GSR, BVP, EMG-Zyg, EMG-Trap, Resp., Temp., EOG**

3. Methodology

We propose a novel and structured multi-modal physiological modeling paradigm for recognizing emotional valence and musical preference during music therapy by analyzing physiological responses. As illustrated in Figure 1, the overall system is organized into four principal stages: (1) music-based stimulus presentation, (2) multi-channel physiological signal acquisition, (3) symmetry-preserving feature extraction via Dynamic Token Feature Extractors (DTFEs), and (4) hierarchical cross-modal fusion for joint affective inference. During the stimulus phase, participants are exposed to a diverse set of music videos designed to evoke varying emotional states and stylistic preferences, while their physiological signals are synchronously recorded. The collected signals include 32-channel electroencephalography (EEG), 6-channel peripheral physiological measurements (e.g., GSR, BVP, EMG, respiration, and skin temperature), and 2-channel electrooculography (EOG), providing complementary observations of cortical, autonomic, and attentional responses to music stimuli. Following preprocessing, each modality is processed by a modality-specific DTFE module. EEG, peripheral, and eye-movement signals are transformed into domain-specific token representations, denoted as $\mathbf{S}$, $\mathbf{L}$, and $\mathbf{I}$, respectively. Within each DTFE, both cross-series interactions among signal channels and intra-series temporal–spectral dynamics are explicitly modeled, yielding compact yet discriminative feature representations. These modality-level representations are subsequently fed into a hierarchical cross-modal fusion module, which integrates spatial, temporal, and frequency-domain patterns across modalities in a structured and symmetry-consistent manner. This design enables synergistic reasoning over heterogeneous physiological signals with distinct temporal characteristics. Finally, the fused representations are jointly optimized to predict emotional valence and music preference under a unified multi-task learning formulation. The entire model is trained end-to-end using a multi-objective loss that balances prediction accuracy and representation efficiency. Through this principled pipeline, the

 doi:10.20944/preprints202602.1827.v1

proposed approach provides a comprehensive and interpretable understanding of how physiological responses encode emotional states and personal preferences during music therapy sessions.

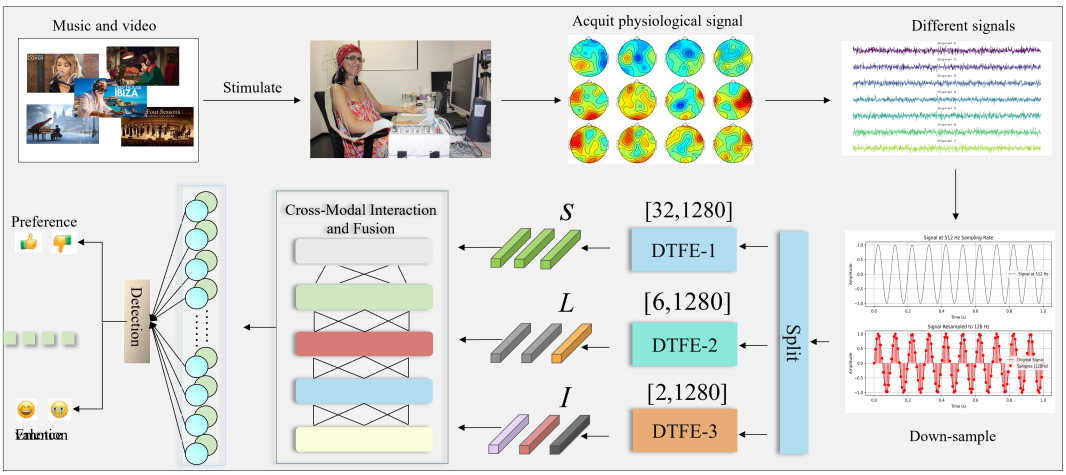

Figure 1. Structured multi-modal physiological modeling framework.

3.1. Dynamic Token Feature Extractor (DTFE)

The proposed *Dynamic Token Feature Extractor (DTFE)* is the core feature encoder in our multi-modal framework, aiming to transform heterogeneous physiological time-series into compact, task-discriminative token representations. As illustrated in Figure 2, DTFE follows a *normalize–project–tokenize–intersect–decode* pipeline, where learnable token matrices $\{Q_i, Q_w, Q_f, Q_o\}$ act as adaptive operators to selectively aggregate informative temporal dynamics and channel correlations. Compared with conventional sequence encoders (e.g., CNN/RNN) that scale linearly with the sequence length, DTFE leverages token-based mixing to achieve both expressive representation learning and computational efficiency, which is particularly desirable for multi-signal affective modeling.

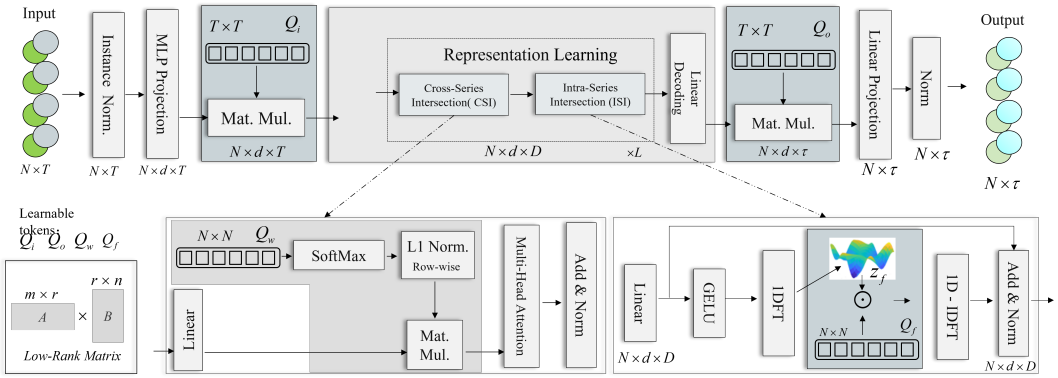

Figure 2. Architecture of the Dynamic Token Feature Extractor (DTFE). DTFE processes input physiological signals via instance normalization, multi-dimensional projection, token-based representation learning with Cross-Series Intersection (CSI) and Intra-Series Intersection (ISI), and output decoding. Learnable tokens Q_i, Q_o, Q_w, and Q_f enable adaptive feature extraction across different signal types.

3.1.1. Module Architecture

Input Formulation

For each modality, we denote the input physiological signal segment as

$$X \in \mathbb{R}^{B \times C \times T}, \tag{1}$$

where B is the batch size, C is the number of channels (e.g., C=32 for EEG, C=6 for peripheral signals, and C=2 for EOG), and T is the temporal length of the window (e.g., T=1280 for 10 s windows at

 doi:10.20944/preprints202602.1827.v1

128 Hz). This formulation explicitly distinguishes *inter-channel* structure from *intra-channel* temporal dynamics, which is essential for the subsequent CSI/ISI design.

Instance Normalization

To reduce inter-subject and inter-session amplitude shifts while preserving discriminative temporal patterns, DTFE first applies instance normalization on each channel:

$$\hat{X}_{b,c,t} = \frac{X_{b,c,t} - \mu_{b,c}}{\sigma_{b,c} + \epsilon}, \qquad \mu_{b,c} = \frac{1}{T}\sum_{t=1}^{T} X_{b,c,t}, \quad \sigma_{b,c} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(X_{b,c,t} - \mu_{b,c})^2}, \tag{2}$$

where ϵ is a small constant for numerical stability. This normalization improves robustness to subject-dependent baselines and facilitates stable optimization across heterogeneous modalities.

3.1.2. Multi-Dimensional Projection

Channel-Wise Projection to Latent Tokens

After normalization, we project each channel sequence into a d-dimensional latent space using a lightweight MLP (implemented as point-wise linear layers along the channel axis):

$$X_{\text{proj}} = \text{MLP}(\hat{X}) \in \mathbb{R}^{B\times d\times T}. \tag{3}$$

Concretely, we first collapse the channel dimension via a learnable mixing operator and then lift the representation to dimension d:

$$X_{\text{proj}}(b,:,t) = W_2\,\phi\big(W_1\,\hat{X}(b,:,t) + b_1\big) + b_2, \tag{4}$$

where $W_1 \in \mathbb{R}^{d\times C}$, $W_2 \in \mathbb{R}^{d\times d}$, $b_1 \in \mathbb{R}^{d}$, $b_2 \in \mathbb{R}^{d}$ are learnable parameters, and $\phi(\cdot)$ is a non-linear activation (e.g., ReLU/GELU). This step serves two purposes: (i) it aligns heterogeneous modalities into a shared latent space, and (ii) it provides sufficient representational capacity to capture subtle physiological cues linked to affective states.

3.1.3. Token-Based Processing

Learnable Temporal Tokenization

A key idea of DTFE is to replace fixed temporal pooling with *learnable token mixing*. Given the projected sequence $X_{\text{proj}} \in \mathbb{R}^{B\times d\times T}$, we introduce a learnable tokenization matrix

$$Q_i \in \mathbb{R}^{T\times T}, \tag{5}$$

and perform token-based temporal aggregation by

$$F_i = X_{\text{proj}}\,Q_i \in \mathbb{R}^{B\times d\times T}. \tag{6}$$

Here, Q_i acts as a data-adaptive temporal mixer: each column of Q_i can be interpreted as a learnable temporal filter that re-weights and combines time steps, enabling the model to emphasize informative sub-structures (e.g., transient peaks in GSR, rhythmic oscillations in EEG, or saccade-related bursts in EOG). Unlike handcrafted filters, Q_i is optimized end-to-end, thereby tailoring temporal aggregation to the downstream affective objectives.

Design Rationale

Physiological responses to music often exhibit (i) non-stationary temporal patterns and (ii) modality-dependent dynamics. The token-based formulation in Eq. (6) provides a unified mechanism to adaptively capture such patterns while keeping the computational cost controlled: the

 doi:10.20944/preprints202602.1827.v1

dominant operations are matrix multiplications on compact tokenized representations, avoiding heavy recurrent computation and enabling efficient training/inference across multiple modalities.

3.1.4. Representation Learning

The representation learning stage constitutes the core of the DTFE module, where tokenized features are progressively refined to capture both *cross-channel dependencies* and *intra-channel temporal–spectral dynamics*. Starting from the temporally tokenized representation $F_i \in \mathbb{R}^{B\times d\times T}$, DTFE employs two complementary components in a cascaded manner:

$$F_{\text{CSI}} = \text{CrossSeriesIntersection}(F_i), \qquad F_{\text{ISI}} = \text{IntraSeriesIntersection}(F_{\text{CSI}}), \tag{7}$$

where B denotes the batch size, T the temporal length, and d the latent feature dimension. The CSI module focuses on modeling inter-channel (cross-series) interactions, while ISI further refines each channel by exploiting temporal and frequency-domain structures. This design explicitly disentangles *where* information is shared across channels from *how* temporal patterns evolve within each channel.

Cross-Series Intersection (CSI)

The Cross-Series Intersection module is designed to capture structured dependencies among different physiological channels. In the case of EEG, CSI models spatial correlations across brain regions; for peripheral physiological signals, it reflects coordinated autonomic responses (e.g., between GSR and BVP); for EOG, it captures bilateral eye-movement consistency. Such cross-series relationships are known to be critical for affective state inference, yet are often overlooked by purely temporal encoders. To this end, CSI first summarizes the temporally tokenized features into compact channel descriptors by temporal pooling:

$$G = \frac{1}{T}\sum_{t=1}^{T} F_i(:,:,t) \in \mathbb{R}^{B\times d}. \tag{8}$$

These descriptors are then projected to a channel-interaction space using a low-rank bilinear formulation:

$$U = W_u G, \quad V = W_v G, \tag{9}$$

where $W_u, W_v \in \mathbb{R}^{C\times d}$ are learnable parameters and C denotes the number of channels for the corresponding modality. The cross-series affinity matrix is obtained as

$$Q_w = U^\top V \in \mathbb{R}^{B\times C\times C}, \tag{10}$$

which explicitly models pairwise channel correlations while maintaining parameter efficiency through low-rank factorization. A row-wise softmax is applied to normalize the affinities:

$$A_w(b,i,j) = \frac{\exp(Q_w(b,i,j))}{\sum_{j'=1}^{C}\exp(Q_w(b,i,j'))}, \tag{11}$$

ensuring that each channel aggregates information from other channels in a convex combination manner. The cross-series aggregated representation is then computed by

$$F_{\text{CSI}} = A_w \cdot F_i + F_i, \tag{12}$$

where the residual connection preserves the original tokenized features and stabilizes optimization. This operation enables each channel to adaptively incorporate complementary information from correlated channels, guided by the learned affinity structure. Compared with conventional self-attention over temporal tokens, CSI explicitly operates in the *channel domain*, thereby decoupling inter-channel modeling from temporal modeling. This separation not only improves interpretability—by

 doi:10.20944/preprints202602.1827.v1

yielding channel-wise interaction matrices—but also reduces computational overhead, since the channel dimension is typically much smaller than the temporal length in physiological signals.

Intra-Series Intersection (ISI)

The Intra-Series Intersection (ISI) module focuses on refining temporal dynamics *within each individual channel* by jointly modeling time-domain and frequency-domain characteristics. While CSI captures cross-channel correlations, ISI aims to enhance channel-wise representations by exploiting the spectral signatures that are known to be highly informative for affective physiology (e.g., EEG rhythms, low-frequency GSR trends, and EOG saccadic patterns).

Time-domain projection and nonlinearity.

Given the cross-series refined features $F_{\text{CSI}} \in \mathbb{R}^{B\times d\times T}$, ISI first applies a channel-wise linear transformation followed by a non-linear activation:

$$F_{\text{lin}} = W_l F_{\text{CSI}} + b_l \in \mathbb{R}^{B\times d\times T}, \tag{13}$$

$$F_{\text{act}} = \text{GELU}(F_{\text{lin}}), \tag{14}$$

where $W_l \in \mathbb{R}^{d\times d}$ and $b_l \in \mathbb{R}^d$ are learnable parameters shared across time steps. The GELU activation introduces smooth nonlinearity while preserving small-amplitude temporal variations, which is desirable for noisy physiological signals.

Frequency-Domain Transformation

To explicitly capture oscillatory and rhythmic patterns, ISI transforms the activated features into the frequency domain via a 1D Fast Fourier Transform (FFT) applied along the temporal dimension:

$$\mathcal{F}(F_{\text{act}})(b,:,k) = \sum_{t=0}^{T-1} F_{\text{act}}(b,:,t)\, e^{-i2\pi kt/T}, \quad k = 0,1,\ldots,T-1, \tag{15}$$

where $\mathcal{F}(\cdot)$ denotes the discrete Fourier transform. This operation decomposes each channel's temporal signal into frequency components, enabling explicit modeling of physiologically meaningful bands.

Learnable Frequency Gating

Instead of using fixed frequency filters, ISI introduces a learnable frequency gating token

$$Q_f \in \mathbb{R}^T, \tag{16}$$

which modulates the spectral amplitudes in a data-driven manner:

$$F_{\text{freq}} = \mathcal{F}(F_{\text{act}}) \odot Q_f, \tag{17}$$

where $\odot$ denotes element-wise multiplication with broadcasting along the batch and feature dimensions. The token Q_f acts as a soft frequency selector, allowing the network to emphasize or suppress specific spectral components according to their relevance to emotion and preference recognition.

Inverse Transformation and Residual Fusion

After frequency-domain modulation, the refined representation is mapped back to the time domain using the inverse FFT:

$$\mathcal{F}^{-1}(F_{\text{freq}})(b,:,t) = \frac{1}{T}\sum_{k=0}^{T-1} F_{\text{freq}}(b,:,k)\, e^{i2\pi kt/T}. \tag{18}$$

 doi:10.20944/preprints202602.1827.v1

The final output of the ISI module is obtained via a residual fusion with the time-domain features:

$$F_{\text{ISI}} = \mathcal{F}^{-1}(F_{\text{freq}}) + \text{LayerNorm}(F_{\text{act}}), \tag{19}$$

where LayerNorm stabilizes the feature scale across channels and time steps. This residual design ensures that frequency-aware refinement complements, rather than overrides, the original temporal representation.

Design Rationale

By integrating learnable frequency gating with residual time-domain refinement, ISI provides a flexible mechanism to capture both transient temporal events and stable oscillatory patterns. This is particularly important in affective computing, where emotional responses are encoded across multiple time scales and frequency bands. Together with CSI, ISI completes a structured decomposition of physiological representation learning into cross-series and intra-series components.

3.1.5. Output Generation

After intra-series refinement, the DTFE module produces a compact, modality-specific representation that summarizes the affect-relevant information of the input physiological signal. Given the refined features $F_{\text{ISI}} \in \mathbb{R}^{B\times d\times T}$, we first apply a linear projection to unify the feature dimension:

$$F_{\text{proj}} = W_d F_{\text{ISI}} + b_d \in \mathbb{R}^{B\times d\times T}, \tag{20}$$

where $W_d \in \mathbb{R}^{d\times d}$ and $b_d \in \mathbb{R}^d$ are learnable parameters shared across time steps.

Temporal aggregation via learnable attention pooling.

Instead of fixed temporal pooling (e.g., average or max pooling), we employ a learnable attention-based aggregation to adaptively summarize the temporal dynamics. Specifically, a temporal importance vector is computed as:

$$\alpha = \text{Softmax}\left(\frac{w_a^\top F_{\text{proj}}}{\sqrt{d}}\right) \in \mathbb{R}^{B\times T}, \tag{21}$$

where $w_a \in \mathbb{R}^d$ is a learnable attention vector. The aggregated representation is obtained as:

$$z = \sum_{t=1}^{T} \alpha_t \, F_{\text{proj}}(:,:,t) \in \mathbb{R}^{B\times d}. \tag{22}$$

This mechanism allows DTFE to focus on temporally salient segments, such as transient physiological responses or sustained emotional patterns.

Output Embedding Projection

The aggregated feature is further mapped to the output embedding space:

$$Y = W_p z + b_p \in \mathbb{R}^{B\times \tau}, \tag{23}$$

where τ denotes the output embedding dimension, and $W_p \in \mathbb{R}^{\tau\times d}$, $b_p \in \mathbb{R}^\tau$ are learnable parameters. The resulting vector Y serves as the final modality-specific representation produced by DTFE, which is subsequently fed into the cross-modal interaction and fusion module.

3.2. Cross-Modal Interaction and Fusion

Overview. As illustrated in Figure 3, the proposed cross-modal interaction and fusion module aims to integrate modality-specific representations extracted by DTFE into a unified affective embedding. Unlike early fusion strategies that directly concatenate raw features, our design explicitly models

 doi:10.20944/preprints202602.1827.v1

inter-modality dependencies at the representation level, enabling complementary physiological cues to be jointly exploited.

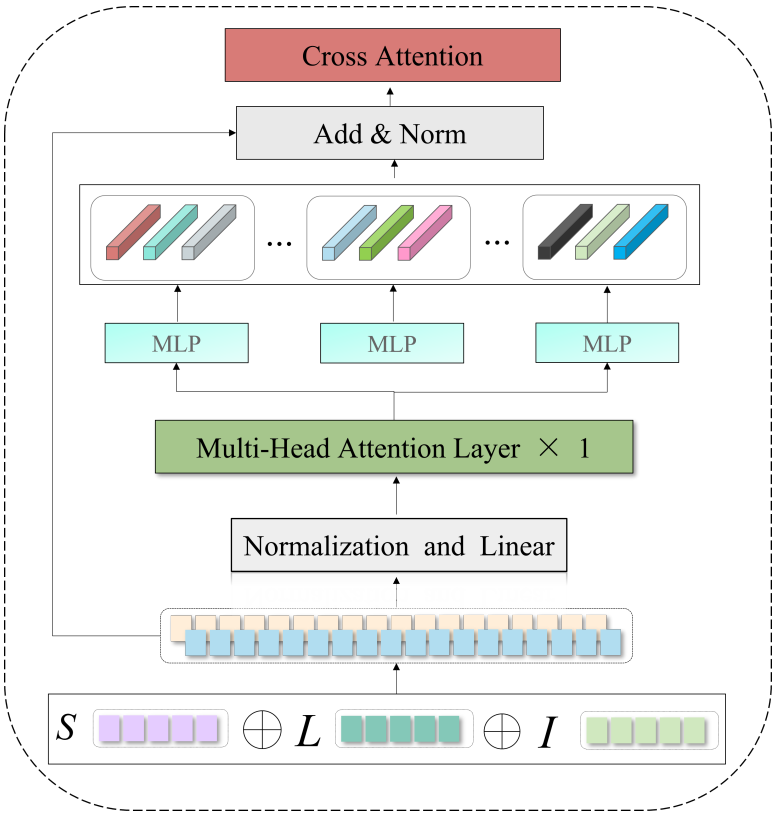

Figure 3. Cross-modal interaction and fusion architecture integrating EEG, peripheral physiological, and eye movement features.

Modality-Level Feature Construction

Let $S \in \mathbb{R}^{B\times d}$, $L \in \mathbb{R}^{B\times d}$, and $I \in \mathbb{R}^{B\times d}$ denote the modality-specific embeddings produced by the EEG, peripheral physiological, and EOG DTFEs, respectively. These embeddings are stacked to form a modality token sequence:

$$F = [S; L; I] \in \mathbb{R}^{B\times M\times d}, \tag{24}$$

where $M = 3$ denotes the number of modalities. Each token encodes high-level affective characteristics specific to one physiological modality.

Cross-Modal Attention-Based Interaction

To explicitly capture inter-modality correlations, we apply a multi-head self-attention mechanism over the modality tokens:

$$F_{\text{att}} = \text{MultiHeadAttention}(F, F, F), \tag{25}$$

where queries, keys, and values are all derived from the modality token set. This operation enables each modality to selectively attend to others, allowing the model to emphasize synergistic patterns such as EEG–autonomic coupling or attention–arousal alignment.

Feature Refinement and Residual Learning

The attended representations are further refined via a position-wise feed-forward network (FFN) with residual connection:

$$F_{\text{ref}} = \text{LayerNorm}(\text{FFN}(F_{\text{att}}) + F_{\text{att}}). \tag{26}$$

This step enhances non-linear interactions among modalities while preserving stable optimization behavior.

 doi:10.20944/preprints202602.1827.v1

Global Fusion and Unified Representation

Finally, a global pooling operation aggregates the refined modality tokens into a single fused embedding:

$$z_{\text{fusion}} = \frac{1}{M} \sum_{m=1}^{M} F_{\text{ref}}^{(m)} \in \mathbb{R}^{B \times d}. \tag{27}$$

The resulting representation z_{fusion} serves as the unified affective descriptor and is subsequently fed into the task-specific prediction heads for emotion and preference recognition.

Discussion. This fusion strategy offers three key advantages: (1) it explicitly models interactions among EEG, autonomic, and attentional modalities at a semantic level; (2) it avoids redundant temporal modeling by operating on compact modality embeddings; and (3) it provides adaptive, data-driven weighting of different physiological cues depending on emotional context. These properties make the proposed fusion module particularly suitable for music therapy scenarios characterized by heterogeneous and complementary physiological responses.

3.2.1. Multi-Task Loss Function

We formulate emotional valence, arousal, and music preference recognition as parallel multi-class classification tasks with nine ordinal categories each. These affective dimensions are inherently correlated yet exhibit different levels of label imbalance, subjectivity, and learning difficulty. To jointly optimize these tasks while maintaining robustness to noisy annotations, we design a unified multi-task objective that combines focal cross-entropy with label smoothing and adaptive task weighting.

Overall Objective

The total training loss is defined as a weighted sum of task-specific losses:

$$\mathcal{L}_{\text{total}} = \lambda_e \mathcal{L}_e + \lambda_a \mathcal{L}_a + \lambda_p \mathcal{L}_p, \tag{28}$$

where $\mathcal{L}_e$, $\mathcal{L}_a$, and $\mathcal{L}_p$ correspond to the losses for valence, arousal, and preference, respectively. The weights λ_e, λ_a, and λ_p are learnable scalar parameters, initialized to $\frac{1}{3}$ and jointly optimized with the network. This design enables the model to automatically balance task contributions during training, alleviating dominance by easier or over-confident tasks.

Task-Specific Loss Formulation

For each affective dimension $t \in \{e, a, p\}$, we adopt a focal cross-entropy loss with label smoothing:

$$\mathcal{L}_t = -\sum_{i=1}^{9} (1 - \hat{y}_t^i)^\gamma \cdot \left[(1 - \alpha)\, y_t^i + \frac{\alpha}{9} \right] \log(\hat{y}_t^i), \tag{29}$$

where y_t^i denotes the one-hot ground-truth label for class i, and $\hat{y}_t^i$ is the predicted softmax probability. The focal factor $(1 - \hat{y}_t^i)^\gamma$ down-weights well-classified samples and emphasizes harder or ambiguous instances, which are common in affective annotations. The label smoothing term, controlled by α, mitigates over-confidence and accounts for subjective uncertainty in emotion perception, particularly at category boundaries.

4. Experiments

4.1. Dataset and Task Setup

The evaluation of the proposed framework is conducted on the well-known DEAP dataset [34], a multimodal benchmark extensively used in affective computing research. The dataset comprises recordings from 32 participants, each viewing 40 one-minute-long music video clips while multiple physiological modalities were simultaneously recorded. Specifically, DEAP provides 32-channel electroencephalography (EEG), six peripheral physiological signals (galvanic skin response (GSR), blood volume pulse (BVP), zygomaticus and trapezius electromyograms (EMG-Zyg, EMG-Trap),

 doi:10.20944/preprints202602.1827.v1

respiration, and temperature), along with two channels of electrooculography (EOG). All signals were originally sampled at 512 Hz and subsequently downsampled to 128 Hz for computational efficiency. The EEG data were bandpass-filtered between 4 and 45 Hz to retain relevant neural oscillatory components. For temporal modeling, the recordings were segmented into overlapping 10-second windows (1280 time steps) with a stride of 1 second. Each window yields three modality-specific tensors: EEG ($[32, 1280]$), peripheral physiological signals ($[6, 1280]$), and EOG ($[2, 1280]$). All signal streams are z-score normalized on a per-subject basis to eliminate inter-individual variability. Unless otherwise specified, all experiments follow a subject-wise data splitting protocol with an 8:2 ratio for training and testing. Specifically, subjects are partitioned into mutually exclusive splits, and all sliding windows extracted from the same subject are strictly assigned to either the training or the testing set, but never both. This design prevents subject identity leakage and enables a fair evaluation of cross-subject generalization. Within the training set, a subset of subjects is further held out for validation to support hyperparameter tuning and early stopping. To generate compact and discriminative representations, a modality-specific tokenization layer is applied within the proposed DTFE module. The EEG stream is projected into four learnable tokens, the peripheral stream into two tokens, and the EOG stream into one token. This asymmetric token design offers a balanced trade-off between representational richness and parameter efficiency, reflecting the relative complexity and dimensionality of each signal modality. A multi-task learning paradigm is adopted to jointly predict three key affective dimensions:

- *Valence*, *Arousal*, and *Liking* scores are discretized into nine ordinal categories, formulated as parallel 9-class classification tasks. Each subtask utilizes an independent prediction head with softmax activation and is optimized using focal loss with label smoothing to address potential class imbalance.

  In addition, an auxiliary branch is introduced for discrete affective quadrant classification:

- *Quadrants (Q1–Q4)* and *Neutral* states are modeled as a five-dimensional multi-label classification task, trained independently from the valence–arousal–liking pipeline. This separation enables an explicit evaluation of categorical emotion distribution without mutual interference from ordinal regression targets.

Model performance is quantitatively evaluated using three standard metrics: **Accuracy (Acc)** measures the proportion of correctly classified samples among the total number of samples, reflecting overall recognition correctness. **F1-score (F1)** represents the harmonic mean of precision and recall, providing a balanced assessment under potential class imbalance. **Precision (Prec)** quantifies the ratio of true positive predictions to all positive predictions, indicating the reliability of positive classifications.

4.2. Training Configuration

For the multi-task classification of valence, arousal, and liking, we use the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 32, and weight decay of 1×10^{-5}. The model is trained for 120 epochs with early stopping based on the validation loss of the valence prediction task. The total loss is a weighted sum of the three task-specific losses, with learnable weights λ_e, λ_a, and λ_p initialized to $1/3$. For the Q1–Q4 + neutral classifier, we train a separate model using sigmoid activation and binary cross-entropy loss. The same optimizer and training schedule are applied. All experiments are implemented in PyTorch and executed on a single NVIDIA RTX 3090 GPU.

4.3. Comparison with Prior Works on DEAP Dataset

To benchmark our proposed framework, we compare its single-task performance with a series of state-of-the-art emotion recognition models evaluated on the DEAP dataset. Since most existing works are designed as single-task pipelines and report classification accuracy only for valence, arousal, or quadrant-based labels, we adapt our framework to match these single-task settings under a strict cross-subject evaluation protocol for fair comparison. In particular, subject-wise data splitting is enforced, ensuring that physiological recordings from the same participant never appear in both training and

 doi:10.20944/preprints202602.1827.v1

testing sets. Furthermore, non-overlapping temporal windows are used during preprocessing to eliminate any potential information leakage across samples. All comparisons in this subsection strictly follow an identical subject-independent evaluation protocol for fair assessment. Specifically, training and testing sets are split at the subject level, such that no subjects appear in both sets, and no overlapping temporal windows are shared between training and evaluation data.

Table 2. Comparison with prior works evaluated on DEAP (accuracy %). Results are directly reported from the respective papers.

Method	Valence (%)	Arousal (%)	Q1–Q4 + Neutral (%)
SVM [4]	81.45	–	–
MT-MKL [29]	60.00	58.00	–
BDAE [5]	85.20	80.50	–
Bimodal-LSTM [30]	83.82	83.23	–
MESAE [6]	83.04	84.18	84.18
DCCA [16]	85.62	84.33	–
MM-ResLSTM [7]	92.30	92.87	–
FGSVM [8]	–	–	89.53
DPAN [31]	78.72	79.03	–
Random Forest [32]	62.58	–	–
RDFKM [33]	64.50	63.10	–
i-Isomap+DCNN [17]	–	–	90.05
RHRPNet [18]	74.17	74.34	–
DEMA [19]	97.55	97.61	97.01
Ours (Single-task)	**98.32**	**98.45**	**97.96**

Discussion. The proposed method consistently outperforms prior models across all three affective dimensions under the same subject-independent evaluation protocol. In particular, compared with the recent DEMA [19], our model achieves gains of $+0.77\%$ in valence, $+0.84\%$ in arousal, and $+0.95\%$ in Q1–Q4 plus neutral classification accuracy. Importantly, these results are obtained under strict cross-subject evaluation settings, where subjects in the test set are completely unseen during training and no overlapping temporal windows are shared between training and testing samples. This ensures that the reported performance reflects genuine generalization across subjects, rather than memorization of subject-specific physiological patterns. The observed performance improvements can be attributed to two key factors. First, the proposed symmetry-aware DTFE explicitly captures complementary temporal–spectral structures across EEG and peripheral signals, leading to more invariant and discriminative representations under subject variability. Second, the hierarchical cross-modal fusion strategy enables structured interaction among neural, autonomic, and attentional modalities, promoting effective alignment between modality-specific cues and task-specific objectives. In contrast to prior methods that rely on limited modalities or shallow fusion schemes, our model leverages a richer set of eight physiological signal types, which contributes to more robust affective decoding even under challenging cross-subject conditions.

4.4. Multi-Task Classification Results and Analysis

Table 3 reports the performance of various baseline models equipped with our proposed multi-task classification heads under a consistent training pipeline. All methods are evaluated on the DEAP dataset with a 9-class setup for valence, arousal, and liking dimensions. Unless otherwise specified, all experiments in this section follow a subject-wise 8:2 train–test split, ensuring that samples from the same subject do not appear in both training and testing sets, thereby avoiding data leakage. Since no prior work explicitly focuses on multi-task modeling of valence, arousal, and liking jointly, we include

 doi:10.20944/preprints202602.1827.v1

a comprehensive set of representative single-task models as baselines. These include common recurrent structures (LSTM, BiLSTM, GRU), convolutional models (1D-CNN, CNN-LSTM), and attention-based or non-deep learning models (Transformer, XGBoost). Each baseline is reconfigured to support our multi-head multi-task setting and trained under the same subject-wise split for fair comparison. As shown in Table 3 and visualized in Figure 4, our full model consistently achieves superior results across all affective dimensions and evaluation metrics. Specifically, our method reaches 92.8% accuracy on valence, 91.8% on arousal, and 93.6% on liking—each surpassing the closest baseline by 2–5 percentage points. F1-scores and Precision follow the same trend, highlighting not only the accuracy but also the robustness of our predictions. Compared to strong single-task models such as Transformer-based [35] or BiLSTM-based [36], our method exhibits more balanced performance across all tasks. For instance, while Transformer performs well on arousal (90.1% accuracy), it struggles on liking (84.2%), indicating limited generalization in multi-target scenarios. Similarly, BiLSTM underperforms on both arousal and liking dimensions. Tree-based models like XGBoost [37] and hybrid models such as CNN-LSTM [38], though competitive on individual metrics, show inconsistent behavior across tasks, further underscoring the necessity of joint modeling. Overall, the superior and stable performance of our method demonstrates the effectiveness of shared feature representation, enhanced by the proposed DTFE module and cross-modal fusion strategy. These results validate that our approach not only outperforms all baselines but also provides a scalable and unified solution for multi-dimensional affective modeling with strong generalization ability.

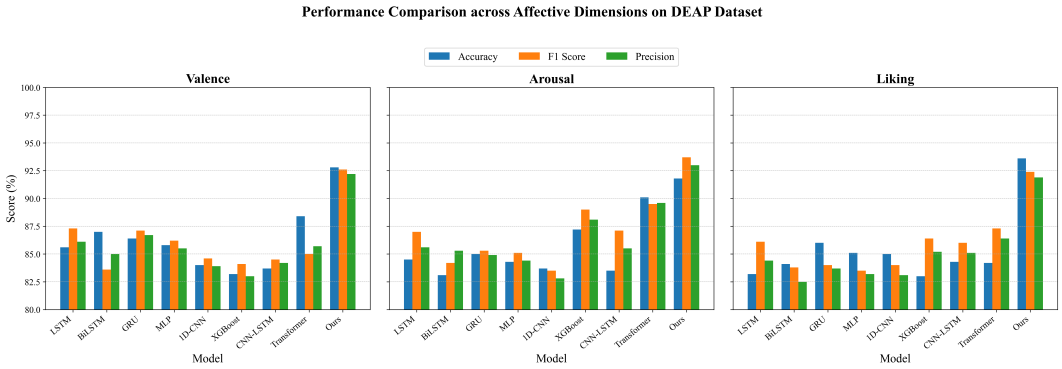

Figure 4. Performance comparison of various models across Valence, Arousal, and Liking dimensions on the DEAP dataset. Each subfigure reports Accuracy, F1-score, and Precision under the 9-class setup. Our proposed method consistently outperforms all baselines across all metrics.

Table 3. Multi-task classification results on DEAP dataset (9-class setup). All models are evaluated under a unified multi-task framework with shared classification heads. Baselines are originally single-task.

Model	Valence			Arousal			Liking		
	Acc (%)	F1 (%)	Prec (%)	Acc (%)	F1 (%)	Prec (%)	Acc (%)	F1 (%)	Prec (%)
LSTM [39]	85.6	87.3	86.1	84.5	87.0	85.6	83.2	86.1	84.4
BiLSTM [36]	87.0	83.6	85.0	83.1	84.2	85.3	84.1	83.8	82.5
GRU [40]	86.4	87.1	86.7	85.0	85.3	84.9	86.0	84.0	83.7
MLP [41]	85.8	86.2	85.5	84.3	85.1	84.4	85.1	83.5	83.2
1D-CNN [42]	84.0	84.6	83.9	83.7	83.5	82.8	85.0	84.0	83.1
XGBoost [37]	83.2	84.1	83.0	87.2	89.0	88.1	83.0	86.4	85.2
CNN-LSTM [38]	83.7	84.5	84.2	83.5	87.1	85.5	84.3	86.0	85.1
Transformer [35]	88.4	85.0	85.7	90.1	89.5	89.6	84.2	87.3	86.4
Ours (Multi-Task)	**92.8**	**92.6**	**92.2**	**91.8**	**93.7**	**93.0**	**93.6**	**92.4**	**91.9**

Note: Baseline models were originally designed for single-task learning and are adapted here with identical multi-task heads and training pipeline for fair evaluation.

 doi:10.20944/preprints202602.1827.v1

4.5. Robustness Analysis Under Different Train–Test Splits

To further evaluate the robustness and generalization ability of the proposed framework, we conduct additional experiments under more challenging train–test splits. Besides the default 8:2 setting, we consider subject-wise 7:3 and 6:4 splits, where fewer training samples are available and the risk of overfitting is more pronounced. All splits are performed in a subject-independent manner to strictly avoid data leakage. Table 4 and Table 5 report the multi-task classification results under the 7:3 and 6:4 splits, respectively. As expected, reducing the proportion of training data leads to a consistent performance degradation across all baseline methods and affective dimensions. This trend is particularly evident for recurrent and convolutional baselines, which exhibit noticeable drops in both accuracy and F1-score as the training data becomes more limited. Despite the increased difficulty, our proposed model maintains a clear performance advantage under both settings. Under the 7:3 split, our method achieves 91.7%, 90.6%, and 92.4% accuracy on valence, arousal, and liking, respectively, consistently outperforming the strongest baselines by a clear margin. When the split is further reduced to 6:4, the performance of all methods decreases more substantially; nevertheless, our framework still attains 90.2% accuracy on valence, 88.9% on arousal, and 90.9% on liking, remaining the best-performing approach across all three affective dimensions. Notably, compared with Transformer-based and recurrent models, our method exhibits a slower performance degradation as the amount of training data decreases. This observation indicates that the proposed Dynamic Token Feature Extractor (DTFE) and the hierarchical cross-modal fusion strategy enable more data-efficient representation learning by effectively capturing complementary temporal and frequency-domain cues across heterogeneous physiological signals. Overall, these results demonstrate that the proposed framework is not only effective under standard evaluation protocols, but also robust to reduced training data, highlighting its strong generalization capability in realistic, data-constrained affective computing scenarios.

Table 4. Multi-task classification results on DEAP dataset (9-class setup) under a subject-wise 7:3 train–test split.

Model	Valence			Arousal			Liking		
	Acc (%)	F1 (%)	Prec (%)	Acc (%)	F1 (%)	Prec (%)	Acc (%)	F1 (%)	Prec (%)
LSTM [39]	84.3	85.8	84.6	83.2	85.4	84.1	82.1	84.7	83.0
BiLSTM [36]	85.6	82.4	83.7	82.0	83.1	84.0	83.0	82.6	81.5
GRU [40]	85.1	85.6	85.2	83.8	84.1	83.6	84.6	82.9	82.6
MLP [41]	84.7	84.9	84.2	83.1	83.8	83.2	84.0	82.4	82.0
1D-CNN [42]	82.9	83.5	82.7	82.6	82.4	81.9	83.8	82.7	81.9
XGBoost [37]	81.9	82.7	81.6	86.1	87.6	86.7	81.8	85.1	84.0
CNN-LSTM [38]	82.6	83.4	83.0	82.4	85.8	84.1	83.1	84.7	83.8
Transformer [35]	87.2	83.6	84.4	88.9	88.1	88.2	82.9	85.8	84.9
Ours (Multi-Task)	**91.7**	**91.4**	**91.0**	**90.6**	**92.5**	**91.8**	**92.4**	**91.2**	**90.7**

Table 5. Multi-task classification results on DEAP dataset (9-class setup) under a subject-wise 6:4 train–test split.

Model	Valence			Arousal			Liking		
	Acc (%)	F1 (%)	Prec (%)	Acc (%)	F1 (%)	Prec (%)	Acc (%)	F1 (%)	Prec (%)
LSTM [39]	82.6	84.1	82.9	81.4	83.2	82.0	80.3	82.8	81.2
BiLSTM [36]	83.8	80.9	82.1	80.6	81.7	82.5	81.5	81.1	80.1
GRU [40]	83.5	84.0	83.6	82.1	82.4	81.9	82.7	81.2	80.9
MLP [41]	83.0	83.2	82.5	81.6	82.3	81.7	82.0	80.6	80.3
1D-CNN [42]	81.2	81.8	81.1	80.9	80.7	80.2	82.1	81.0	80.2
XGBoost [37]	80.1	80.9	79.8	84.7	86.1	85.3	80.4	83.7	82.5
CNN-LSTM [38]	81.5	82.3	81.9	81.1	84.2	82.8	81.7	83.3	82.3
Transformer [35]	85.6	82.1	82.8	87.4	86.8	86.9	81.4	84.0	83.1
Ours (Multi-Task)	**90.2**	**89.9**	**89.4**	**88.9**	**90.8**	**90.1**	**90.9**	**89.7**	**89.2**

 doi:10.20944/preprints202602.1827.v1

4.6. Ablation Study on DTFE and Fusion Modules

To further assess the contribution of each key component, we conduct ablation experiments by selectively removing the Dynamic Token Feature Extractor (DTFE) and the Cross-Modal Interaction and Fusion module. As shown in Table 6, removing the DTFE modules (*w/o DTFE*) significantly degrades performance across all metrics, with valence accuracy dropping from 92.8% to 88.5%. This confirms the importance of DTFE's token-based temporal-frequency modeling in capturing emotional patterns. Similarly, disabling the Cross-Modal Fusion module (*w/o Cross-Modal Fusion*) leads to a consistent drop in performance, particularly on the liking dimension (from 93.6% to 89.1%), suggesting that cross-modal alignment is critical for modeling subjective preference responses. In both cases, our full model consistently outperforms the ablated variants, validating the necessity of both DTFE and hierarchical fusion for effective multi-modal emotion and preference recognition.

Table 6. Ablation study of DTFE and Cross-Modal Fusion on DEAP dataset.

Model Variant	Valence			Arousal			Liking		
	Acc (%)	F1 (%)	Prec (%)	Acc (%)	F1 (%)	Prec (%)	Acc (%)	F1 (%)	Prec (%)
w/o DTFE	88.5	87.3	86.8	87.1	85.9	85.5	88.0	86.2	85.4
w/o Cross-Modal Fusion	89.4	88.2	87.9	88.3	87.0	86.4	89.1	87.1	86.8
Ours (Full)	**92.8**	**92.6**	**92.2**	**91.8**	**93.7**	**93.0**	**93.6**	**92.4**	**91.9**

4.7. Ablation Study: Learnable vs. Fixed Loss Weights

To evaluate the effectiveness of our learnable loss weight mechanism, we compare the performance of models trained with fixed weights $\lambda_e = \lambda_a = \lambda_p = \frac{1}{3}$ and models trained with learnable weights initialized to the same values and optimized jointly with the main network. Both settings share the same architecture and training schedule, isolating the impact of dynamic loss weighting.

Analysis. As shown in Table 7, enabling the loss weights to be learnable significantly improves classification performance across all three affective dimensions. Compared to fixed equal weights, the dynamic weighting mechanism allows the model to adaptively prioritize more difficult or underperforming tasks during training. This leads to more balanced optimization and consistent gains in accuracy, F1-score, and precision. We conclude that learnable task weights are essential for achieving optimal performance in multi-task affective modeling.

Table 7. Comparison between fixed and learnable loss weights on DEAP dataset (9-class classification).

Loss Weight Type	Valence (%)			Arousal (%)			Liking (%)		
	Acc	F1	Prec	Acc	F1	Prec	Acc	F1	Prec
Fixed Weights	90.3	89.7	88.9	89.1	89.4	88.7	89.5	88.8	88.0
Learnable Weights (Ours)	**92.8**	**92.6**	**92.2**	**91.8**	**93.7**	**93.0**	**93.6**	**92.4**	**91.9**

5. Potential Applications

Beyond academic contributions, our proposed framework holds promising potential for real-world applications in the fields of personalized healthcare, music-based therapy, and affective content recommendation. By leveraging fine-grained physiological responses to musical stimuli, the system can be deployed to support a variety of emotion-centered tasks:

- **Therapeutic Efficacy Monitoring:** The framework can serve as a non-invasive tool for evaluating emotional changes during music therapy sessions. By analyzing trends in predicted valence, arousal, and liking scores over time, clinicians can quantitatively assess the therapeutic impact of specific musical interventions for individuals with depression, anxiety, or cognitive disorders.
- **Adaptive Music Recommendation:** The multi-modal emotional profiling capability of the system can be integrated into intelligent music recommendation engines. Unlike traditional behavior-based systems, our model enables real-time adaptation of musical content based on a listener's physiological feedback, facilitating mood enhancement and emotional regulation.

 doi:10.20944/preprints202602.1827.v1

- **Biofeedback-Driven Interactive Systems:** The proposed approach can be embedded into immersive environments, such as virtual reality (VR) or meditation platforms, where users' physiological states are continuously monitored to personalize audio-visual content, adjust difficulty levels, or enhance engagement.
- **Clinical Decision Support:** In therapeutic contexts, longitudinal data collected through our system can contribute to clinical dashboards, enabling psychologists or music therapists to identify emotional baselines, detect anomalies, and tailor interventions based on real-time physiological patterns.

6. Conclusion

In this work, we presented a unified and symmetry-aware multi-modal framework for emotion and preference recognition in the context of music therapy, leveraging a comprehensive set of physiological signals including EEG, GSR, BVP, EMG, respiration, temperature, and EOG. Rather than treating multimodal affective modeling as a purely data-driven fusion problem, the proposed framework is explicitly grounded in structured representation learning, aiming to preserve intrinsic regularities and invariances underlying physiological affective responses.

At the core of the framework lies the Dynamic Token Feature Extractor (DTFE), which provides a principled mechanism to transform raw physiological time series into compact and discriminative token embeddings. By explicitly decomposing representation learning into cross-series interaction symmetry and intra-series temporal–spectral symmetry, DTFE enables structured and interpretable feature extraction across heterogeneous signal modalities. This symmetry-preserving design distinguishes the proposed approach from conventional sequence encoders that implicitly entangle different dependency structures.

By integrating DTFE outputs through a hierarchical cross-modal fusion mechanism, the proposed system further achieves symmetry-consistent information integration across neural, autonomic, and attentional modalities. The resulting unified representation supports simultaneous multi-task prediction of emotional valence, arousal, liking, and quadrant-based affective indicators, effectively addressing long-standing challenges in affective computing, including modality heterogeneity, inter-subject variability, and task interdependence.

Extensive experiments on the DEAP dataset demonstrate that the proposed method consistently outperforms state-of-the-art approaches under both single-task and multi-task settings, achieving superior accuracy, F1-score, and precision across all affective dimensions. Comprehensive ablation studies further confirm the necessity of each structural component, validating the complementary roles of symmetry-aware feature decomposition, hierarchical fusion, and adaptive loss balancing in robust multi-task affective learning. Beyond quantitative performance improvements, this work highlights the broader value of token-based and symmetry-preserving physiological modeling as a principled paradigm for structured affective representation learning. By aligning model design with the inherent structure of physiological signals, the proposed framework improves generalization and interpretability, offering a scalable and fully end-to-end solution for real-world affective applications. These findings not only advance the methodological foundation of multi-modal affective computing, but also pave the way for intelligent, feedback-driven music therapy systems, personalized affect-aware applications, and broader human–computer interaction scenarios. Future work will explore longitudinal affect modeling, real-time adaptive intervention strategies, and validation on larger and more diverse clinical datasets to further investigate the role of symmetry-aware structured learning in affective computing.

Author Contributions: Conceptualization, W.Q. and M.-J.-S.W.; methodology, W.Q.; software, W.Q.; validation, W.Q. and M.-J.-S.W.; formal analysis, W.Q.; investigation, W.Q.; resources, M.-J.-S.W.; data curation, W.Q.; writing—original draft preparation, W.Q.; writing—review and editing, M.-J.-S.W.; visualization, W.Q.; supervision, M.-J.-S.W.; project administration, M.-J.-S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the authors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on reasonable request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Acknowledgments: The authors would like to thank the administrative and technical staff of the College of Arts Management, Shandong University of Arts, and the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, for their support during this study. During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.2) for language polishing and formatting assistance. The authors have reviewed and edited the generated content and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yang, Y.; Wu, Q.M.J.; Zheng, W.L.; Lu, B.L. EEG-Based Emotion Recognition Using Hierarchical Network with Subnetwork Nodes. *IEEE Transactions on Cognitive and Developmental Systems* **2018**, *10*, 408–419.
2. Wang, X.W.; Nie, D.; Lu, B.L. Emotional State Classification from EEG Data Using Machine Learning Approach. *Neurocomputing* **2014**, *129*, 94–106.
3. Li, C.; Xu, C.; Feng, Z. Analysis of Physiological Signals for Emotion Recognition with the IRS Model. *Neurocomputing* **2016**, *178*, 103–111.
4. Verma, G.K.; Tiwary, U.S. Multimodal Fusion Framework: A Multiresolution Approach for Emotion Classification and Recognition from Physiological Signals. *NeuroImage* **2014**, *102*, 162–172.
5. Liu, W.; Zheng, W.L.; Lu, B.L. Emotion Recognition Using Multimodal Deep Learning. In Proceedings of the Proceedings of the 23rd International Conference on Neural Information Processing (ICONIP), Kyoto, Japan, 2016; pp. 521–529.
6. Yin, Z.; Zhao, M.; Wang, Y.; Yang, J.; Zhang, J. Recognition of Emotions Using Multimodal Physiological Signals and an Ensemble Deep Learning Model. *Computer Methods and Programs in Biomedicine* **2017**, *140*, 93–110.
7. Ma, J.; Tang, H.; Zheng, W.L.; Lu, B.L. Emotion Recognition Using Multimodal Residual LSTM Network. In Proceedings of the Proceedings of the 27th ACM International Conference on Multimedia. ACM, 2019, pp. 176–183.
8. Hassan, M.M.; Alam, M.G.R.; Uddin, M.Z.; Huda, S.; Almogren, A.; Fortino, G. Human Emotion Recognition Using Deep Belief Network Architecture. *Information Fusion* **2019**, *51*, 10–18.
9. Wang, M.; Yang, W.; Guo, Y.; Wang, S. Conditional fault tolerance in a class of Cayley graphs. *International Journal of Computer Mathematics* **2016**, *93*, 67–82.
10. Wang, M.; Lin, Y.; Wang, S. The connectivity and nature diagnosability of expanded k-ary n-cubes. *RAIRO-Theoretical Informatics and Applications-Informatique Théorique et Applications* **2017**, *51*, 71–89.
11. Wang, S.; Wang, Y.; Wang, M. Connectivity and matching preclusion for leaf-sort graphs. *Journal of Interconnection Networks* **2019**, *19*, 1940007.
12. Wang, M.; Xu, S.; Jiang, J.; Xiang, D.; Hsieh, S.Y. Global reliable diagnosis of networks based on Self-Comparative Diagnosis Model and g-good-neighbor property. *Journal of Computer and System Sciences* **2025**, p. 103698.
13. Wang, M.; Yang, W.; Wang, S. Conditional matching preclusion number for the Cayley graph on the symmetric group. *Acta Math. Appl. Sin.(Chinese Series)* **2013**, *36*, 813–820.
14. Wang, M.; Ren, Y.; Lin, Y.; Wang, S. The tightly super 3-extra connectivity and diagnosability of locally twisted cubes. *American Journal of Computational Mathematics* **2017**, *7*, 127–144.
15. Wang, S.; Wang, M. A Note on the Connectivity of m-Ary n-Dimensional Hypercubes. *Parallel Processing Letters* **2019**, *29*, 1950017.
16. Liu, W.; Qiu, J.L.; Zheng, W.L.; Lu, B.L. Multimodal Emotion Recognition Using Deep Canonical Correlation Analysis. *arXiv preprint arXiv:1908.05349* **2019**.
17. Zhang, Y.; Cheng, C.; Zhang, Y. Multimodal Emotion Recognition Based on Manifold Learning and Convolution Neural Network. *Multimedia Tools and Applications* **2022**, *81*, 33253–33268.

 doi:10.20944/preprints202602.1827.v1

18. Tang, J.; Ma, Z.; Gan, K.; Zhang, J.; Yin, Z. Hierarchical Multimodal Fusion of Physiological Signals for Emotion Recognition with Scenario Adaption and Contrastive Alignment. *Information Fusion* **2024**, *103*, 102129.
19. Li, Q.; Jin, D.; Huang, J.; Zhang, L.; Liu, H.; Wu, J.; Wang, Y. DEMA: Deep EEG-first Multi-Physiological Affect Model for Emotion Recognition, 2025. Manuscript under review / publication details to be updated.
20. Lan, Z.; Sourina, O.; Wang, L.; Liu, Y. Real-Time EEG-Based Emotion Monitoring Using Stable Features. *The Visual Computer* **2016**, *32*, 347–358.
21. Li, Y.; Zheng, W.; Cui, Z.; Zong, Y.; Ge, S. EEG Emotion Recognition Based on Graph Regularized Sparse Linear Regression. *Neural Processing Letters* **2019**, *49*, 555–571.
22. Cheng, J.; Chen, M.; Li, C.; Liu, Y.; Song, R.; Liu, A.; Chen, X. Emotion Recognition from Multi-Channel EEG via Deep Forest. *IEEE Journal of Biomedical and Health Informatics* **2020**, *25*, 453–464.
23. Song, T.; Zheng, W.; Song, P.; Cui, Z. EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing* **2020**, *11*, 532–541.
24. Zhang, Y.; Liu, H.; Zhang, D.; Chen, X.; Qin, T.; Zheng, Q. EEG-Based Emotion Recognition with Emotion Localization via Hierarchical Self-Attention. *IEEE Transactions on Affective Computing* **2023**. Early Access.
25. Liu, S.; Zhao, Y.; An, Y.; Zhao, J.; Wang, S.H.; Yan, J. GLFANet: A Global to Local Feature Aggregation Network for EEG Emotion Recognition. *Biomedical Signal Processing and Control* **2023**, *85*, 104799.
26. Jin, H.; Gao, Y.; Wang, T.; Gao, P. DAST: A Domain-Adaptive Learning Combining Spatio-Temporal Dynamic Attention for Electroencephalography Emotion Recognition. *IEEE Journal of Biomedical and Health Informatics* **2023**. Early Access.
27. Huang, H.; Xie, Q.; Pan, J.; He, Y.; Wen, Z.; Yu, R.; Li, Y. An EEG-Based Brain–Computer Interface for Emotion Recognition and Its Application in Patients with Disorder of Consciousness. *IEEE Transactions on Affective Computing* **2019**, *12*, 832–842.
28. Gu, X.; Cao, Z.; Jolfaei, A.; Xu, P.; Wu, D.; Jung, T.P.; Lin, C.T. EEG-Based Brain–Computer Interfaces (BCIs): A Survey of Recent Studies on Signal Sensing Technologies and Computational Intelligence Approaches and Their Applications. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2021**, *18*, 1645–1666.
29. Kandemir, M.; Vetek, A.; Gönen, M.; Klami, A.; Kaski, S. Multi-Task and Multi-View Learning of User State. *Neurocomputing* **2014**, *139*, 97–106.
30. Tang, H.; Liu, W.; Zheng, W.L.; Lu, B.L. Multimodal Emotion Recognition Using Deep Neural Networks. In Proceedings of the Proceedings of the 24th International Conference on Neural Information Processing (ICONIP), Guangzhou, China, 2017; pp. 811–819.
31. Kim, B.H.; Jo, S. Deep Physiological Affect Network for the Recognition of Human Emotions. *IEEE Transactions on Affective Computing* **2020**, *11*, 230–243.
32. Kusumaningrum, T.D.; Faqih, A.; Kusumoputro, B. Emotion Recognition Based on DEAP Database Using EEG Time-Frequency Features and Machine Learning Methods. *Journal of Physics: Conference Series* **2020**, *1501*, 012020.
33. Zhang, X.; Liu, J.; Shen, J.; Li, S.; Hou, K.; Hu, B.; Gao, J.; Zhang, T. Emotion Recognition from Multimodal Physiological Signals Using a Regularized Deep Fusion of Kernel Machine. *IEEE Transactions on Cybernetics* **2021**, *51*, 4386–4399.
34. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Transactions on Affective Computing* **2012**, *3*, 18–31.
35. Valanarasu, J.M.J.; Patel, V.M. UNeXt: MLP-Based Rapid Medical Image Segmentation Network. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI 2022), Singapore, 2022; pp. 23–33.
36. Siami-Namini, S.; Tavakoli, N.; Namin, A.S. The Performance of LSTM and BiLSTM in Forecasting Time Series. In Proceedings of the Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019; pp. 3285–3292. https://doi.org/10.1109/BigData47090.2019.9005997.
37. Lu, W.; Li, J.; Li, Y.; Sun, A.; Wang, J. A CNN–LSTM–Based Model to Forecast Stock Prices. *Complexity* **2020**, *2020*, 6622927. https://doi.org/10.1155/2020/6622927.
38. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in Transformer. *Advances in Neural Information Processing Systems* **2021**, *34*, 15908–15919.
39. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to Forget: Continual Prediction with LSTM. *Neural Computation* **2000**, *12*, 2451–2471. https://doi.org/10.1162/089976600300015015.

 doi:10.20944/preprints202602.1827.v1

40. Rana, R. Gated Recurrent Unit (GRU) for Emotion Classification from Noisy Speech. *arXiv preprint arXiv:1612.07778* **2016**.
41. Azizjon, M.; Jumabek, A.; Kim, W. 1D CNN Based Network Intrusion Detection with Normalization on Imbalanced Data. In Proceedings of the Proceedings of the 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020; pp. 218–224. https://doi.org/10.1109/ICAIIC48513.2020.9064912.
42. Nielsen, D. Tree Boosting with XGBoost—Why Does XGBoost Win "Every" Machine Learning Competition? Master's thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2016.