A Modality-Aware Graph-Structured and Symmetry-Aware Multi-Task Learning Framework for Joint Emotion Recognition and Immersion Estimation in Virtual Reality

Haibing Wang and Mu-Jiang-Shan Wang *

Posted Date: 16 January 2026

doi: 10.20944/preprints202601.1212.v1

Keywords: affective computing; graph-structured modeling; symmetry-aware learning; multimodal fusion; emotion–immersion modeling; virtual reality; human–computer interaction; adaptive systems

 doi:10.20944/preprints202601.1212.v1

Article

A Modality-Aware Graph-Structured and Symmetry-Aware Multi-Task Learning Framework for Joint Emotion Recognition and Immersion Estimation in Virtual Reality

Haibing Wang [1] and Mu-Jiang-Shan Wang [2,*]

[1] Design College, Shandong University of Arts, Jinan 250014, China
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
* Correspondence: mjs.wang@siat.ac.cn

Abstract

Virtual Reality (VR) has emerged as a powerful medium for immersive human–computer interaction, where users' emotional and experiential states play a pivotal role in shaping engagement and perception. However, existing affective computing approaches often model emotion recognition and immersion estimation as independent problems, overlooking their intrinsic coupling and the structured relationships underlying multimodal physiological signals. In this work, we propose a modality-aware multi-task learning framework that jointly models emotion recognition and immersion estimation from a graph-structured and symmetry-aware interaction perspective. Specifically, heterogeneous physiological and behavioral modalities—including eye-tracking, electrocardiogram (ECG), and galvanic skin response (GSR)—are treated as relational components with balanced structural roles, while their cross-modality dependencies are adaptively aggregated to preserve interaction symmetry and allow controlled asymmetry across tasks, without introducing explicit graph neural network architectures. To support reproducible evaluation, the VREED dataset is further extended with quantitative immersion annotations derived from presence-related self-reports via weighted aggregation and factor analysis. Extensive experiments demonstrate that the proposed framework consistently outperforms recurrent, convolutional, and Transformer-based baselines, achieving higher accuracy (75.42%), F1-score (74.19%), and Cohen's Kappa (0.66) for emotion recognition, as well as superior regression performance for immersion estimation (RMSE = 0.96, MAE = 0.74, R-squared = 0.63). Beyond empirical improvements, this study provides a structured interpretation of multimodal affective modeling that highlights symmetry, coupling, and symmetry breaking in multi-task learning, offering a principled foundation for adaptive VR systems, emotion-driven personalization, and dynamic user experience optimization.

Keywords: affective computing; graph-structured modeling; symmetry-aware learning; multimodal fusion; emotion–immersion modeling; virtual reality; human–computer interaction; adaptive systems

1. Introduction

With the rapid advancement of Virtual Reality (VR) technology, its applications in immersive art presentation, personalized recommendation, and educational training have expanded significantly. In virtual environments, users' emotional states and immersion experiences not only directly affect content receptiveness but also provide vital insights for human–computer interaction system design [1,2]. Recent studies have shown that immersive VR can substantially modulate affective responses and aesthetic judgments, highlighting the importance of emotion-aware modeling in virtual environments [3,4]. Traditional emotion recognition methods rely heavily on facial expressions or speech; however, such cues are often compromised in immersive VR scenarios due to headset occlusion or latency issues, resulting in reduced robustness and limited real-time applicability [5,6].

 doi:10.20944/preprints202601.1212.v1

In recent years, physiological signals such as eye-tracking, electrocardiogram (ECG), and galvanic skin response (GSR) have attracted increasing attention as reliable and stable indicators of affective and experiential states. These signals capture fine-grained emotional fluctuations and subjective immersion more accurately under VR conditions and have been widely adopted in multimodal emotion recognition systems [3,7–11]. For example, Koelstra et al. constructed the DEAP dataset to demonstrate correlations between EEG signals and emotional responses [12], while Tabbaa et al. introduced the VREED dataset integrating eye-tracking and GSR signals for emotion evaluation in VR environments [13]. Beyond affective recognition, recent VR studies have also explored how advanced interaction paradigms, including large language model–powered guidance and cinematic VR simulations, influence user behavior, learning outcomes, and physiological responses [14,15]. Moreover, an emerging line of research has leveraged immersive VR not only as a sensing platform but also as an active intervention medium for emotional regulation and mental health support. Recent works have demonstrated the effectiveness of VR-based systems in treating emotional disorders, promoting psychological well-being, and alleviating chronic pain through embodied and interactive experiences [16–18]. These application-oriented findings further emphasize the necessity of accurate emotion and immersion modeling, as reliable affective perception constitutes a fundamental prerequisite for adaptive, personalized, and therapeutic VR systems. From a structural perspective, these multimodal signals form a set of heterogeneous yet interrelated components whose interactions exhibit inherent relational symmetry across modalities.

Meanwhile, psychological constructs such as sense of presence and flow state have been shown to play a mediating role in affective experience. Lønne et al. reported that higher immersion levels are associated with more positive emotional responses [1], while Ochs and Sonderegger revealed complex interdependencies between presence, engagement, and learning motivation [19,20]. These findings suggest that emotion and immersion should not be treated as isolated phenomena, but rather as coupled dimensions within a unified experiential structure. Such coupling naturally reflects a form of task-level symmetry, where both dimensions share latent representations while contributing asymmetrically to overall user experience.

Multi-task learning (MTL) has therefore emerged as a promising paradigm for jointly modeling emotion recognition and immersion estimation by enabling representation sharing across related tasks. Nevertheless, conventional MTL approaches often suffer from task interference and gradient conflicts, particularly when handling heterogeneous objectives such as emotion classification and immersion regression. Recent studies have proposed conflict-aware optimization strategies to alleviate these issues and improve training stability [21,22]. From a learning-theoretic viewpoint, these challenges can be interpreted as manifestations of symmetry breaking during joint optimization, where balanced task coupling must be carefully controlled to avoid dominance by a single objective.

Despite these advances, current studies on multimodal emotion and immersion assessment still face three key limitations:

1. **Lack of a unified modeling framework** that jointly captures emotion recognition and immersion estimation as structurally coupled tasks. Most existing methods address these problems independently, thereby failing to exploit their intrinsic relational symmetry [1,10,20].
2. **Insufficient design of modality-aware feature extraction and fusion mechanisms** for heterogeneous physiological signals. Many approaches overlook the balanced yet distinct roles of eye-tracking, ECG, and GSR modalities, resulting in suboptimal representation learning and weak cross-modality coordination [8,11,23,24].
3. **Severe task interference in multi-task optimization**, arising from differences in label structures and learning objectives, which can lead to unstable convergence and asymmetric task dominance during training [21,22,25,26].

From a structural reliability perspective, diagnosability theory in interconnection networks has provided a rigorous framework for characterizing fault tolerance under comparison-based models. Early studies on nature diagnosability investigated intrinsic fault identification limits in symmetric

 doi:10.20944/preprints202601.1212.v1

graph structures, such as bubble-sort star graphs under the PMC and MM* models [27]. Subsequent extensions systematically generalized these ideas through connectivity-aware and g-good-neighbor diagnosability analyses, establishing global reliability criteria for complex networks with structured symmetry and constrained fault propagation [28–33].

To address these challenges, we propose **MMEA-Net**, a modality-aware framework for **joint emotion recognition and immersion estimation** in VR environments. MMEA-Net adopts a structured interaction perspective in which multimodal physiological signals are treated as relational components and emotion–immersion objectives are modeled as coupled tasks with controlled symmetry and asymmetry. The main contributions of this work are summarized as follows:

- We propose a unified multi-task framework, **MMEA-Net**, for jointly modeling emotion classification and immersion regression in VR environments. Compared with the best-performing Transformer-based baseline, MMEA-Net achieves a **+2.55%** improvement in test accuracy (75.42% vs. 72.87%), a **+2.54%** increase in F1-score (74.19% vs. 71.65%), and a **+0.04** gain in Cohen's Kappa (0.66 vs. 0.62) for emotion classification. For immersion estimation, the proposed model reduces RMSE by **0.09** (0.96 vs. 1.05), lowers MAE by **0.07** (0.74 vs. 0.81), and improves the coefficient of determination R^2 by **+0.05** (0.63 vs. 0.58), demonstrating effective cross-task synergy.
- We design a **Hybrid-M modality-aware module** and a **Cross-Domain Fusion mechanism**. Hybrid-M employs dedicated sub-networks to encode temporal characteristics of eye-tracking, ECG, and GSR signals, while the fusion mechanism facilitates structured and symmetry-aware interaction across modalities and tasks.
- We extend the VREED dataset by annotating quantitative immersion scores, enabling **dual-task benchmarking** and supporting reproducible research in multimodal affective computing.

2. Related Work

2.1. Emotion Recognition Using Physiological Signals

Emotion recognition using physiological signals has become an important research direction in affective computing. The DEAP dataset constructed by Koelstra et al. [12] is one of the earliest large-scale multimodal resources, comprising EEG, GSR, and facial signals annotated with affective states. This dataset has inspired extensive subsequent studies on robust representation learning from physiological inputs, particularly in settings where conventional behavioral cues are unreliable.

Recent research has increasingly adopted deep neural architectures to capture temporal and inter-signal dependencies. For instance, Hu et al. [34] and Xiao et al. [35] proposed convolutional recurrent models augmented with self-attention mechanisms to model spatiotemporal patterns in EEG sequences, achieving strong performance on arousal and valence prediction. Lin and Li [8] provided a comprehensive survey of physiological modalities, including ECG and GSR, emphasizing their relative stability compared with facial expressions and speech, especially under constrained or immersive conditions.

Beyond EEG-centric approaches, a growing body of work has explored multimodal fusion strategies. Fu et al. [10] combined eye-tracking, ECG, and GSR signals with transfer learning to enhance classification robustness. Katada and Okada [24] focused on user-independent modeling by incorporating modality importance weighting, while Moin et al. [6] developed a multimodal fusion pipeline for real-time human–machine interaction. Collectively, these studies suggest that physiological modalities can be regarded as heterogeneous yet structurally related components whose coordinated modeling is essential for reliable emotion recognition.

2.2. Multimodal Emotion Modeling in Virtual Reality

Virtual Reality (VR) provides a natural environment for eliciting complex emotional responses through immersive and interactive stimuli. The VREED dataset introduced by Tabbaa et al. [13] collected eye-tracking and physiological signals during emotion-inducing VR experiences, offering a valuable benchmark for affective analysis in immersive settings. Building on this dataset, Alharbi [36]

 doi:10.20944/preprints202601.1212.v1

proposed a transparent deep learning framework for VR-based emotion recognition, emphasizing explainable feature selection across multiple physiological channels.

A number of studies have further examined the relationships among presence, immersion, and emotional experience. Lønne et al. [1] showed that higher immersion levels are associated with enhanced affective responses in educational VR environments. Souza et al. [37] surveyed metrics for measuring presence in virtual environments, while Ochs and Sonderegger [20] identified links between immersion, cognitive engagement, and emotional outcomes. Related investigations by Yang et al. [2] and Liu et al. [38] explored how immersive art and educational visualization influence user flow, neural activity, and emotional engagement. These findings collectively indicate that emotion and immersion are not independent constructs, but rather interrelated dimensions that evolve jointly during immersive experiences.

2.3. Multi-Task Learning for Affective and Immersion-Aware Modeling

Multi-task learning (MTL) has been widely adopted to improve generalization by sharing representations across related tasks. Transformer-based architectures, including the original Transformer [39], T5 [40], and Vision Transformer (ViT) [41], have demonstrated strong capability in capturing shared structural patterns across modalities and objectives.

In affective computing, several studies have applied MTL to mitigate overfitting and enhance task synergy. Liu and Yu [22] proposed MT2ST, which dynamically alternates between multi-task and single-task training phases. Liu et al. [21] introduced the Conflict-Averse Gradient Descent (CAGrad) algorithm to reduce gradient interference, an issue that becomes particularly pronounced when jointly optimizing heterogeneous objectives such as emotion classification and continuous state estimation. From a structural perspective, these approaches aim to balance task coupling while preventing dominance by a single objective during training.

Despite these advances, relatively few existing frameworks explicitly incorporate immersion or presence estimation as learning targets alongside emotion recognition. This omission overlooks the influence of immersive experience on emotional processing, as consistently reported in VR-based studies [1,37]. Jointly modeling affective and immersive dimensions within a multi-task framework therefore represents a promising direction for capturing the structured dependencies underlying user experience in virtual environments.

Similar challenges arise in other scientific domains that rely on multi-source, noisy observations. For example, geophysical studies infer subsurface velocity anomalies by integrating triplicated seismic waveforms across multiple phases and scales, where reliable conclusions emerge from coordinated evidence rather than isolated signals [42]. This cross-domain analogy highlights the importance of structured fusion and multi-scale consistency, reinforcing our design choice to jointly model emotion and immersion through coordinated multimodal interactions within a unified learning framework.

3. Methodology

3.1. Overview of MMEA-Net

In this section, we present MMEA-Net (Multimodal Emotion and Engagement Assessment Network), a deep learning framework designed for the joint modeling of emotion state classification and immersion level estimation in virtual reality environments. Figure 1 provides an overview of the proposed architecture, highlighting its structured organization and task-coupled design.

MMEA-Net adopts a multimodal formulation that integrates three heterogeneous yet structurally related data streams: visual and behavioral information derived from eye tracking, physiological dynamics captured by electrocardiogram (ECG), and autonomic responses measured through galvanic skin response (GSR). Treating these modalities as complementary components within a unified representation space enables the model to capture coordinated patterns of user responses that are difficult to observe through single-modality analysis.

 doi:10.20944/preprints202601.1212.v1

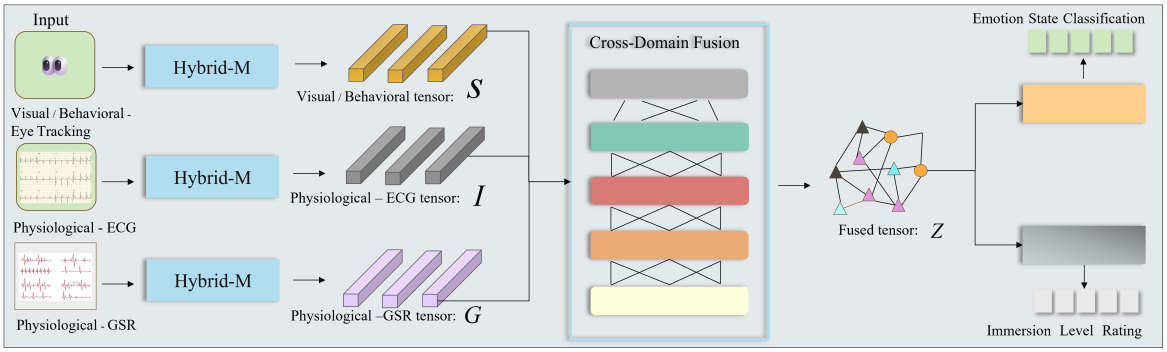

Figure 1. Overall architecture of the proposed MMEA-Net. The framework consists of three main components: (1) modality-specific feature extraction modules for eye-tracking, ECG, and GSR signals; (2) a cross-domain fusion module for structured multimodal integration; and (3) dual prediction heads for emotion classification and immersion estimation.

The overall architecture is organized around three tightly coupled components:

- Hybrid-M modules, which perform modality-specific feature extraction to preserve the intrinsic temporal characteristics of each signal while maintaining a consistent structural interface across modalities.
- A Cross-Domain Fusion module, which integrates modality-specific representations through coordinated interactions, allowing information to be exchanged while maintaining balance among modalities.
- A Multi-scale Feature Extraction (MFE) mechanism, embedded within both the modality-specific and fusion stages, to capture temporal patterns at different resolutions in a structurally consistent manner.

This structured design allows MMEA-Net to address key challenges in multimodal emotion and immersion analysis, including heterogeneity across physiological signals and imbalance between learning objectives. By organizing modality processing and task prediction in a coordinated and symmetric fashion, the framework promotes stable representation sharing while avoiding dominance by any single modality or task.

Furthermore, the dual-task learning formulation enables emotion recognition and immersion estimation to be optimized jointly. Rather than treating these objectives independently, MMEA-Net exploits their intrinsic coupling, encouraging the learning of shared representations that reflect the mutual influence between affective and immersive states.

The model is trained and evaluated using the VREED dataset, which provides synchronized eye-tracking, ECG, and GSR recordings collected from participants exposed to diverse VR scenarios. The availability of aligned emotional labels and immersion-related self-reports makes this dataset particularly suitable for studying the structured relationship between emotion and immersion within a unified learning framework.

3.2. Hybrid-M: Modality-Specific Feature Extraction

The Hybrid-M module constitutes the foundational layer of the proposed framework and is responsible for processing each input modality independently prior to cross-domain fusion. This design follows a parallel and structurally consistent processing strategy, ensuring that heterogeneous physiological and behavioral signals are encoded through comparable transformation pipelines while preserving their modality-specific characteristics.

For each modality $m \in \{S, I, G\}$, where S denotes visual and behavioral eye-tracking signals, I corresponds to physiological ECG data, and G represents physiological GSR data, the Hybrid-M module performs feature extraction according to

$$T_m = \text{Hybrid-M}(X_m), \tag{1}$$

 doi:10.20944/preprints202601.1212.v1

where X_m is the raw input of modality m, and T_m is the resulting tensor representation. This formulation enforces a uniform mapping structure across modalities, allowing modality-specific information to be encoded within a shared representational framework.

As illustrated in Figure 2 (right), the Hybrid-M module is composed of several sequential processing stages:

1. Vectorization. Raw input signals are first transformed into vectorized representations suitable for neural processing. For eye-tracking data, this includes spatial gaze coordinates, pupil dilation, and fixation-related statistics. For ECG signals, features such as R–R intervals and heart rate variability are extracted, while GSR signals are decomposed into tonic and phasic components. This step establishes a common vector-level interface across modalities.
2. Multi-scale feature extraction. To capture temporal patterns occurring at different resolutions, a multi-scale feature extraction mechanism is applied:

$$F_{MFE} = \mathrm{MLP}\big(\mathrm{Conv}_{1\times1}(F) \oplus \mathrm{Conv}_{3\times3}(F) \oplus \mathrm{Conv}_{7\times7}(F)\big), \tag{2}$$

where F denotes the input feature sequence, $\mathrm{Conv}_{k\times k}$ represents convolution with kernel size $k \times k$, and $\oplus$ denotes concatenation. This design enables the model to capture both short-term fluctuations and longer-term trends in a balanced manner.
3. Normalization and linear transformation. The extracted features are normalized and linearly transformed to stabilize optimization and align feature distributions across modalities:

$$F_{norm} = \mathrm{LayerNorm}\big(\mathrm{Linear}(F_{MFE} \oplus F_{vec})\big), \tag{3}$$

where F_{vec} denotes the original vectorized features retained through a skip connection. This operation preserves modality-specific information while enforcing scale consistency.
4. Transformer-based temporal modeling. A transformer decoder is employed to model sequential dependencies:

$$Z = \mathrm{TransformerDecoder}(F_{norm}), \tag{4}$$

enabling the extraction of long-range temporal relationships that are critical for modeling evolving emotional and physiological responses.
5. Linear projection and residual refinement. The transformer output is further processed by linear and fully connected layers:

$$F_{FC} = \mathrm{FC}(\mathrm{Linear}(Z)), \tag{5}$$

followed by a residual connection and normalization:

$$F_{res} = \mathrm{LayerNorm}(F_{FC} + F_{norm}). \tag{6}$$

This step refines the representation while maintaining structural consistency across modalities.
6. Multi-head attention. Finally, a multi-head attention mechanism is applied to emphasize informative temporal segments:

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O, \tag{7}$$

where each attention head is computed as

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{8}$$

and

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{9}$$

 doi:10.20944/preprints202601.1212.v1

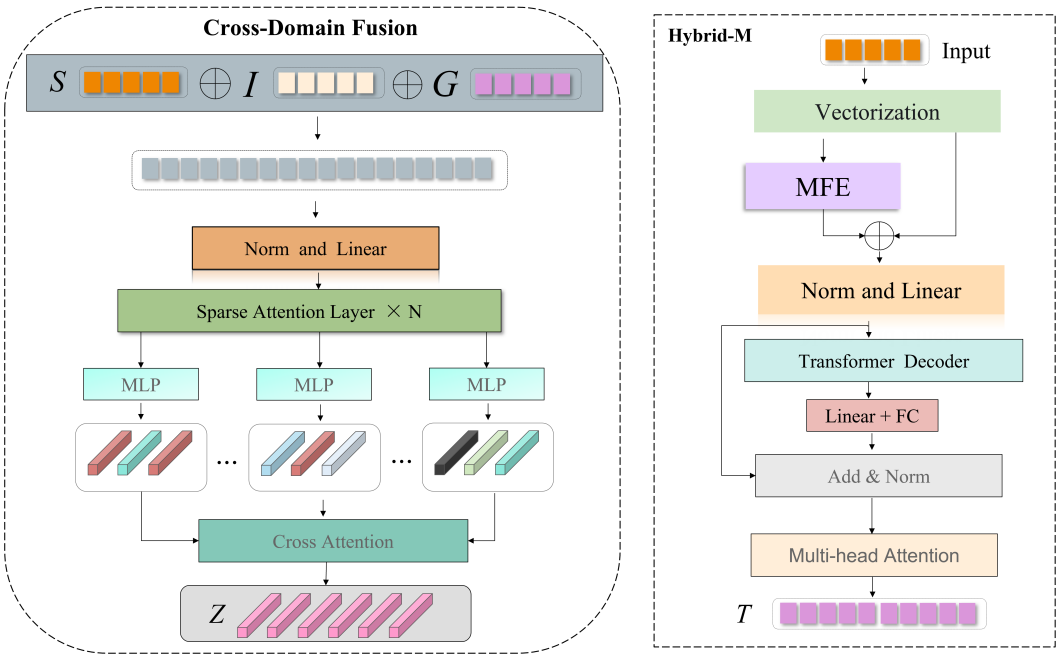

Figure 2. Detailed architecture of the Hybrid-M module (right) and Cross-Domain Fusion module (left). The Hybrid-M module transforms raw multimodal inputs into structured tensor representations through vectorization, multi-scale feature extraction, normalization, transformer-based temporal modeling, and attention mechanisms.

The output of the Hybrid-M module is a structured tensor representation T_m that encodes modality-specific temporal dynamics within a unified processing framework. By enforcing parallel and structurally aligned transformations across modalities, Hybrid-M preserves signal heterogeneity while enabling balanced integration in subsequent fusion stages. This design provides a stable and symmetric foundation for cross-domain interaction and joint emotion–immersion modeling.

3.3. Cross-Domain Fusion

The Cross-Domain Fusion module, shown in Figure 2 (left), is designed to integrate modality-specific representations while maintaining their individual structural properties. This module aims to coordinate information from heterogeneous sources in a balanced manner, enabling structured interaction among physiological and behavioral signals for joint emotion and immersion modeling.

Given the tensor representations S, I, and G extracted from eye-tracking, ECG, and GSR modalities, respectively, the fusion process is formulated as

$$Z_{fused} = \text{CrossDomainFusion}(S, I, G). \tag{10}$$

Rather than collapsing modalities into a single representation at an early stage, this formulation preserves modality-specific structure while allowing controlled interaction across domains.

The fusion procedure consists of several sequential stages:

1. Initial tensor combination. The modality-specific tensors are first combined through learnable weighting:

$$Z_{initial} = \alpha S \oplus \beta I \oplus \gamma G, \tag{11}$$

where α, β, and γ are trainable parameters and $\oplus$ denotes concatenation. This operation establishes a balanced aggregation scheme in which each modality contributes proportionally, preventing dominance by any single source.

2. Normalization and linear transformation. The combined representation is then normalized and linearly transformed:

$$Z_{norm} = \text{Norm}(\text{Linear}(Z_{initial})), \tag{12}$$

 doi:10.20944/preprints202601.1212.v1

ensuring scale alignment and structural consistency across modalities before higher-order interaction.

3. Sparse attention layers. A stack of sparse attention layers refines the normalized representation:

$$Z_i = \text{SparseAttention}(Z_{i-1}), \quad i = 1, \ldots, N, \tag{13}$$

with $Z_0 = Z_{norm}$. The sparsity constraint restricts attention to a subset of informative relationships, promoting efficient and structured interaction among features while reducing redundancy.

4. Parallel feature projection. The output of the sparse attention layers is processed by multiple parallel multilayer perceptrons:

$$F_j = \text{MLP}_j(Z_N), \quad j = 1, 2, 3, \tag{14}$$

where each MLP captures a distinct transformation perspective. This parallel design preserves symmetry in feature processing, allowing multiple coordinated views of the fused representation.

5. Cross-attention integration. The parallel feature projections are integrated through a cross-attention mechanism:

$$Z_{fused} = \text{CrossAttention}(F_1, F_2, F_3), \tag{15}$$

defined as

$$\text{CrossAttention}(F_1, F_2, F_3) = \text{softmax}\left(\frac{F_1 F_2^T}{\sqrt{d_k}}\right) F_3. \tag{16}$$

This operation enables structured information exchange across feature perspectives, reinforcing complementary patterns while maintaining internal consistency.

The resulting fused representation Z_{fused} encodes coordinated multimodal information in a unified yet structured form. By organizing fusion through balanced weighting, sparse interaction, parallel transformation, and cross-perspective integration, the Cross-Domain Fusion module captures inter-modal relationships without sacrificing modality-specific integrity. This design supports stable downstream prediction for both emotion classification and immersion estimation by preserving symmetry and coordination across modalities.

3.4. Multi-Scale Feature Extraction (MFE)

The Multi-scale Feature Extraction (MFE) module is designed to capture temporal patterns at different resolutions, which is essential for modeling the complex dynamics of emotional responses and immersion levels in virtual reality environments. Physiological and behavioral signals inherently exhibit variations across multiple time scales, ranging from rapid transient reactions to slowly evolving trends. The MFE module addresses this characteristic by organizing feature extraction into parallel and structurally consistent branches operating at different temporal scales.

Given an input feature representation x, the MFE module computes the multi-scale output as

$$x' = \text{MFE}(x). \tag{17}$$

This formulation allows features extracted at different temporal resolutions to be integrated within a single structured representation.

As illustrated in Figure 3, the MFE module consists of the following stages:

1. Initial feature projection. The input representation is first transformed using a multilayer perceptron:

$$F_{initial} = \text{MLP}(x), \tag{18}$$

which maps the input features into a space suitable for parallel multi-scale processing.

 doi:10.20944/preprints202601.1212.v1

2. Parallel multi-scale convolutions. Three convolutional operations with different kernel sizes are applied in parallel:

$$F_{small} = \text{Conv}_{1\times1}(F_{initial}), \tag{19}$$
$$F_{medium} = \text{Conv}_{3\times3}(F_{initial}), \tag{20}$$
$$F_{large} = \text{Conv}_{7\times7}(F_{initial}), \tag{21}$$

where each branch captures patterns at a distinct temporal resolution. The parallel design ensures structural symmetry across scales, allowing fine-grained, intermediate, and coarse temporal dynamics to be modeled in a balanced manner.

3. Feature aggregation. The outputs from different scales are concatenated:

$$F_{concat} = F_{small} \oplus F_{medium} \oplus F_{large}, \tag{22}$$

preserving information from all temporal resolutions while maintaining their individual contributions.

4. Integrated transformation. The aggregated representation is further processed through convolutional and nonlinear transformations:

$$F_{conv} = \text{Conv}_{3\times3}(F_{concat}), \tag{23}$$
$$F_{relu} = \text{ReLU}(F_{conv}), \tag{24}$$

followed by a final MLP:

$$x' = \text{MLP}(F_{relu}), \tag{25}$$

which integrates information across scales into a unified feature representation.

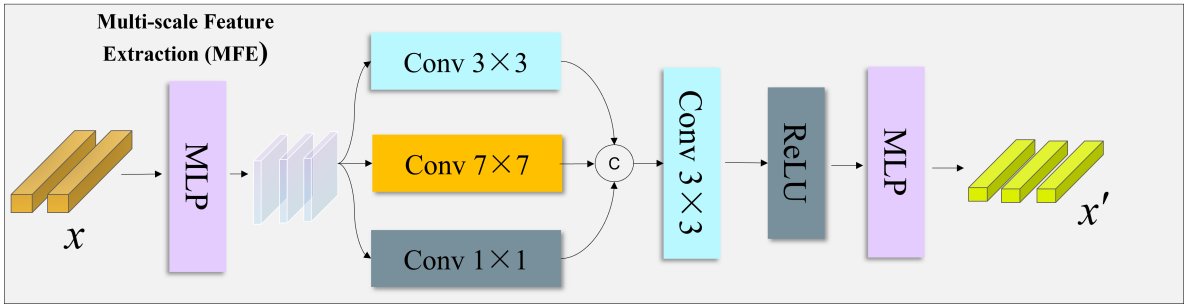

Figure 3. Architecture of the Multi-scale Feature Extraction (MFE) module. Input features are first transformed through an MLP, followed by parallel convolutional branches with different kernel sizes to capture patterns at multiple temporal scales. The resulting features are concatenated and further processed to produce a unified multi-scale representation.

By organizing feature extraction through parallel branches and coordinated aggregation, the MFE module captures temporal dynamics at multiple resolutions without favoring any single scale. This structurally balanced design enables MMEA-Net to represent both rapid emotional reactions and gradual changes in immersion in a consistent manner, providing robust multi-scale features for downstream multimodal fusion and dual-task prediction.

3.5. Training and Optimization

MMEA-Net is trained in an end-to-end manner using a unified multi-task objective that jointly optimizes emotion state classification and immersion level estimation. The overall loss function is formulated as a weighted combination of task-specific objectives and regularization:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{emotion} + \lambda_2 \mathcal{L}_{immersion} + \lambda_3 \mathcal{L}_{reg}, \tag{26}$$

 doi:10.20944/preprints202601.1212.v1

where $\mathcal{L}_{emotion}$ corresponds to the classification loss for emotion recognition, $\mathcal{L}_{immersion}$ measures regression error for immersion estimation, $\mathcal{L}_{reg}$ is a regularization term to prevent overfitting, and λ_1, λ_2, and λ_3 control the relative contribution of each component. This formulation enables balanced optimization across heterogeneous objectives while avoiding dominance by any single task.

The emotion classification loss is defined using categorical cross-entropy:

$$\mathcal{L}_{emotion} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}), \tag{27}$$

where N denotes the number of samples, C is the number of emotion categories, $y_{i,c}$ indicates whether sample i belongs to class c, and $p_{i,c}$ represents the predicted class probability. This loss encourages discriminative separation among emotional states while remaining compatible with joint optimization.

Immersion level estimation is modeled as a regression task using mean squared error:

$$\mathcal{L}_{immersion} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{28}$$

where y_i and $\hat{y}_i$ denote the ground truth and predicted immersion levels for sample i, respectively. This objective captures continuous variations in immersive experience and complements the discrete emotion classification task.

To improve generalization and stabilize training, L2 regularization is applied to all learnable parameters:

$$\mathcal{L}_{reg} = \frac{1}{2}\sum_{w\in\mathcal{W}} w^2, \tag{29}$$

where $\mathcal{W}$ denotes the set of model parameters. Regularization serves as a constraint that maintains symmetry in parameter updates across tasks during optimization.

Model training is performed using the Adam optimizer with an initial learning rate of 0.001. A learning rate decay strategy is employed to promote stable convergence:

$$\mathrm{lr} = \mathrm{lr}_0 \cdot \mathrm{decay\_rate}^{\lfloor \mathrm{epoch}/\mathrm{decay\_steps} \rfloor}, \tag{30}$$

where lr_0 is the initial learning rate, decay_rate is set to 0.95, and decay_steps is set to 2.

To further reduce overfitting, dropout with a rate of 0.3 is applied to fully connected layers, and weight decay with a coefficient of 0.0001 is used. Early stopping based on validation performance is adopted to prevent unnecessary overtraining. Mini-batch gradient descent is conducted with a batch size of 32, and training is performed for up to 100 epochs.

The dataset is partitioned into training (70%), validation (15%), and test (15%) subsets, with stratified sampling to preserve balanced class distributions. This training protocol ensures consistent evaluation and supports stable joint learning of emotion and immersion under a unified optimization framework.

3.5.1. VREED Dataset Overview

The VREED [12] (Virtual Reality for Emotion Elicitation Dataset) is a multimodal dataset designed to support emotion recognition research in immersive virtual reality environments. It provides synchronized recordings of physiological signals, behavioral measurements, and subjective self-reports collected from 25 participants exposed to a range of emotion-eliciting VR scenarios. The dataset comprises the following components:

- Visual and behavioral data, including eye-tracking metrics such as fixation duration, saccade amplitude, pupil dilation, and gaze coordinates.
- Physiological data from electrocardiogram (ECG) recordings sampled at 256 Hz, capturing heart rate dynamics and variability.

 doi:10.20944/preprints202601.1212.v1

- Physiological data from galvanic skin response (GSR) recordings sampled at 128 Hz, reflecting autonomic arousal through skin conductance changes.
- Self-reported emotion annotations, where participants evaluated their affective states using the Self-Assessment Manikin (SAM) scale along the valence, arousal, and dominance dimensions.
- Post-exposure questionnaire responses, including presence-related items used to assess immersion levels.

Participants experienced eight distinct 360° VR video stimuli selected to elicit representative emotional states across the valence–arousal space, such as joy, fear, sadness, and calmness. Each recording session lasted approximately 60 seconds, resulting in a total of 440 synchronized multimodal sequences. The structured alignment of behavioral signals, physiological responses, and subjective reports makes VREED particularly suitable for studying the coordinated relationship between emotion and immersion in VR environments.

3.5.2. Immersion Level Extraction

Although the VREED dataset was originally developed for emotion recognition, it also contains structured information relevant to immersion assessment through post-exposure presence questionnaires. These questionnaires consist of seven presence-related items derived from the Presence Questionnaire (PQ), each rated on a 7-point Likert scale ranging from 0 to 6.

To derive a continuous immersion level annotation, we designed a systematic aggregation procedure that integrates multiple PQ items while accounting for their relative contributions to the overall sense of presence. The extraction process involves the following steps:

1. Reverse coding. For negatively phrased items (POST_PQ2, POST_PQ3, and POST_PQ5), reverse coding is applied to ensure a consistent directional interpretation, such that higher scores uniformly correspond to higher immersion levels:

$$PQ_{\text{reversed}} = 6 - PQ_{\text{original}}. \tag{31}$$

2. Item selection. Based on factor analysis of the PQ responses, six items closely associated with immersion-related constructs are retained (POST_PQ1, POST_PQ2, POST_PQ3, POST_PQ4, POST_PQ5, and POST_PQ7). The item POST_PQ6, which primarily reflects attention to background music, is excluded due to its weak correlation with the remaining presence dimensions.
3. Weighted aggregation. A weighted average of the selected items is computed to obtain a unified immersion score:

$$\begin{aligned}\text{Immersion\_Score} = &\frac{w_1 \cdot PQ1 + w_2 \cdot (6 - PQ2) + w_3 \cdot (6 - PQ3) + w_4 \cdot PQ4}{\sum_{i \in \{1,2,3,4,5,7\}} w_i} \\ &+ \frac{w_5 \cdot (6 - PQ5) + w_7 \cdot PQ7}{\sum_{i \in \{1,2,3,4,5,7\}} w_i},\end{aligned} \tag{32}$$

where w_i denotes the weight associated with each PQ item. These weights are determined through principal component analysis (PCA), reflecting the relative contribution of each dimension to the overall immersion construct.

This aggregation strategy produces a continuous immersion label that preserves the balanced contribution of multiple presence-related dimensions, enabling consistent integration with physiological and behavioral features during joint emotion–immersion modeling.

3.5.3. Data Preprocessing

Prior to model training, a series of preprocessing steps were applied to ensure consistency and structural alignment across multimodal data streams:

- Temporal alignment. All modalities were synchronized and resampled to a unified sampling rate of 64 Hz, enabling consistent temporal correspondence across signals.

 doi:10.20944/preprints202601.1212.v1

- Normalization. Z-score normalization was applied to all features to standardize their scale:

$$x_{normalized} = \frac{x - \mu}{\sigma}, \tag{33}$$

  where μ and σ denote the mean and standard deviation of each feature, respectively.
- Segmentation. Continuous recordings were segmented into non-overlapping windows of 5 seconds, resulting in approximately 5,280 multimodal segments used for model training and evaluation.
- Feature extraction. For each modality, both statistical descriptors (e.g., mean and standard deviation) and temporal features (e.g., frequency-domain characteristics and rate-of-change measures) were extracted to capture complementary signal properties.
- Missing data handling. Occasional missing values in eye-tracking signals, such as those caused by blinks, were addressed using forward filling for short gaps and interpolation for longer gaps to preserve temporal continuity.

After preprocessing, the dataset was divided into training (70%), validation (15%), and test (15%) subsets using stratified sampling to maintain similar distributions of emotion labels and immersion scores across splits. To prevent data leakage, all segments originating from the same participant–video pair were assigned exclusively to a single subset.

3.6. Experimental Setup

3.6.1. Implementation Details

MMEA-Net was implemented using PyTorch 1.9.0. All experiments were conducted on a workstation equipped with an NVIDIA RTX 3090 GPU with 24GB memory, an Intel Core i9-10900K CPU, and 64GB RAM.

Model hyperparameters were selected through grid search on the validation set and configured as follows:

- Number of transformer layers in Hybrid-M: 4
- Number of sparse attention layers in Cross-Domain Fusion: 3
- Number of attention heads: 8
- Hidden dimension: 256
- Dropout rate: 0.3
- Learning rate: 0.001 with cosine annealing schedule
- Batch size: 32
- Maximum epochs: 100 with early stopping (patience = 10)
- Loss weights: $\lambda_1 = 1.0$ for emotion classification, $\lambda_2 = 0.5$ for immersion estimation, and $\lambda_3 = 0.0001$ for regularization

To improve robustness, data augmentation techniques were applied during training, including random temporal shifting within ± 0.5 seconds, magnitude scaling within $\pm 10\%$, and additive Gaussian noise with standard deviation $\sigma = 0.01$.

3.6.2. Evaluation Metrics

For the emotion state classification task, the following evaluation metrics are employed:

- Accuracy. The proportion of correctly classified samples:

$$\text{Accuracy} = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}}. \tag{34}$$

- F1-score. The harmonic mean of precision and recall:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{35}$$

 doi:10.20944/preprints202601.1212.v1

where Precision = TP/(TP+FP) and Recall = TP/(TP+FN).

- Cohen's Kappa. A measure of agreement between predicted and true labels that accounts for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{36}$$

where p_o denotes the observed agreement and p_e represents the expected agreement by chance.

For the immersion level estimation task, the following regression metrics are used:

- Root Mean Square Error (RMSE). The square root of the mean squared prediction error:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{37}$$

- Mean Absolute Error (MAE). The average absolute difference between predicted and ground-truth immersion levels:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \tag{38}$$

- Coefficient of Determination (R^2). The proportion of variance in the target variable explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{39}$$

where $\bar{y}$ denotes the mean of the ground-truth immersion scores.

3.6.3. Baseline Models

To evaluate the effectiveness of the proposed MMEA-Net architecture, several baseline models are constructed for comparison. These baselines are obtained by replacing the Hybrid-M modules with alternative sequence modeling approaches and substituting the Cross-Domain Fusion mechanism with direct feature concatenation. This design enables a controlled assessment of the contributions of each architectural component.

The baseline models include:

- LSTM-based model [43]. Uses two-layer LSTM networks with 128 hidden units for each modality, followed by direct feature concatenation.
- BiLSTM-based model [44]. Employs bidirectional LSTMs to model forward and backward temporal dependencies, with simple concatenation for multimodal fusion.
- GRU-based model [45]. Replaces Hybrid-M modules with GRU networks (two layers, 128 hidden units) and applies direct feature concatenation.
- 1D-CNN-based model [46]. Uses one-dimensional convolutional networks with multiple kernel sizes (3, 5, and 7) for temporal feature extraction, followed by feature concatenation.
- XGBoost-based model [47]. Extracts handcrafted statistical and frequency-domain features from each modality and performs prediction using XGBoost, with concatenated outputs.
- CNN-LSTM-based model [48]. Combines convolutional layers for feature extraction with LSTM layers for temporal modeling, using direct concatenation for fusion.
- Transformer-based model [49]. Applies Transformer encoder blocks with multi-head self-attention to each modality, followed by simple feature concatenation.
- MLP-based model [50]. Processes modality-specific features using multilayer perceptrons and concatenates the resulting representations.
- MMEA-Net (Simple Fusion). Retains the Hybrid-M modules while replacing the Cross-Domain Fusion mechanism with direct feature concatenation.

 doi:10.20944/preprints202601.1212.v1

All baseline models follow the same overall processing pipeline, training strategy, and evaluation protocol as MMEA-Net. Each model operates on the same set of modalities, namely eye-tracking, ECG, and GSR signals, and performs joint emotion classification and immersion estimation. Identical preprocessing procedures, including temporal alignment, normalization, and segmentation into 5-second windows, are applied across all methods.

Hyperparameters for each baseline are optimized via grid search on the validation set, focusing on learning rate, hidden dimensions, and dropout rates. The loss formulation and task weighting remain consistent for all models, ensuring a fair and balanced comparison. This experimental design allows for a systematic evaluation of both the Hybrid-M modules and the Cross-Domain Fusion mechanism under a unified dual-task setting.

3.7. Experimental Results and Analysis

In this section, we report the experimental results of the proposed MMEA-Net and compare its performance with various baseline models on the VREED dataset. All models are evaluated on both validation and test sets using the classification and regression metrics described earlier for emotion recognition and immersion estimation.

3.7.1. Comparison with Baseline Models

Table 1 summarizes the quantitative performance of MMEA-Net and baseline methods on the validation and test sets. The baselines cover a broad range of modeling paradigms, including recurrent architectures (LSTM [43], BiLSTM [44], and GRU [45]), convolutional approaches (1D-CNN [46] and CNN-LSTM [48]), tree-based models (XGBoost [47]), and attention-driven methods such as Transformer [49] and MLP-based [50] models.

As shown in Table 1 and Figure 4, MMEA-Net achieves the best overall performance across both tasks on the validation and test sets. On the emotion classification task, the proposed model attains the highest test accuracy, F1 score, and Cohen's Kappa, indicating improved discriminative capability and classification stability. For immersion estimation, MMEA-Net yields the lowest prediction errors (RMSE and MAE) and the highest coefficient of determination, demonstrating superior regression performance.

Table 1. Performance comparison of MMEA-Net and baseline models on the validation and test sets of the VREED dataset.

Model	Validation Set						Test Set					
	Emotion Classification			Immersion Prediction			Emotion Classification			Immersion Prediction		
	Acc (%)	F1 (%)	Kappa	RMSE	MAE	R^2	Acc (%)	F1 (%)	Kappa	RMSE	MAE	R^2
LSTM-based [43]	67.43	65.87	0.53	1.24	0.97	0.46	65.91	64.22	0.51	1.31	1.02	0.43
BiLSTM-based [44]	69.21	67.95	0.56	1.18	0.92	0.49	68.07	66.54	0.55	1.23	0.95	0.47
GRU-based [45]	67.85	66.32	0.54	1.21	0.94	0.47	66.23	64.89	0.52	1.28	0.99	0.44
1D-CNN-based [46]	70.14	68.76	0.58	1.14	0.88	0.52	68.92	67.41	0.56	1.19	0.93	0.49
XGBoost-based [47]	65.37	63.54	0.51	1.29	1.03	0.41	64.21	62.38	0.49	1.35	1.08	0.38
CNN-LSTM-based [48]	71.53	70.24	0.60	1.08	0.84	0.55	70.19	68.75	0.58	1.15	0.89	0.52
Attention-based	72.68	71.42	0.62	1.06	0.82	0.57	71.34	70.05	0.60	1.12	0.87	0.54
Transformer-based [49]	74.12	72.98	0.64	0.99	0.76	0.61	72.87	71.65	0.62	1.05	0.81	0.58
MLP-based [50]	63.79	62.14	0.48	1.33	1.07	0.38	62.45	60.89	0.46	1.39	1.13	0.35
MMEA-Net (Simple Fusion)	73.65	72.31	0.63	1.02	0.79	0.59	72.14	70.88	0.61	1.08	0.84	0.56
MMEA-Net (Ours)	76.93	75.47	0.68	0.91	0.70	0.65	75.42	74.19	0.66	0.96	0.74	0.63

Several observations can be drawn from the comparative results. First, the comparison between MMEA-Net and its simple fusion variant indicates that structured cross-domain fusion contributes substantially to performance gains, particularly for immersion prediction. Second, recurrent and convolutional baselines exhibit competitive but consistently lower performance, suggesting limitations in capturing complex temporal and cross-modal interactions. Third, attention-based models improve over traditional sequence models but still fall short of MMEA-Net, highlighting the importance of jointly modeling multi-scale temporal features and coordinated multimodal fusion. Finally, the relatively weaker performance of the XGBoost-based model underscores the difficulty of representing multimodal physiological dynamics using handcrafted features alone.

 doi:10.20944/preprints202601.1212.v1

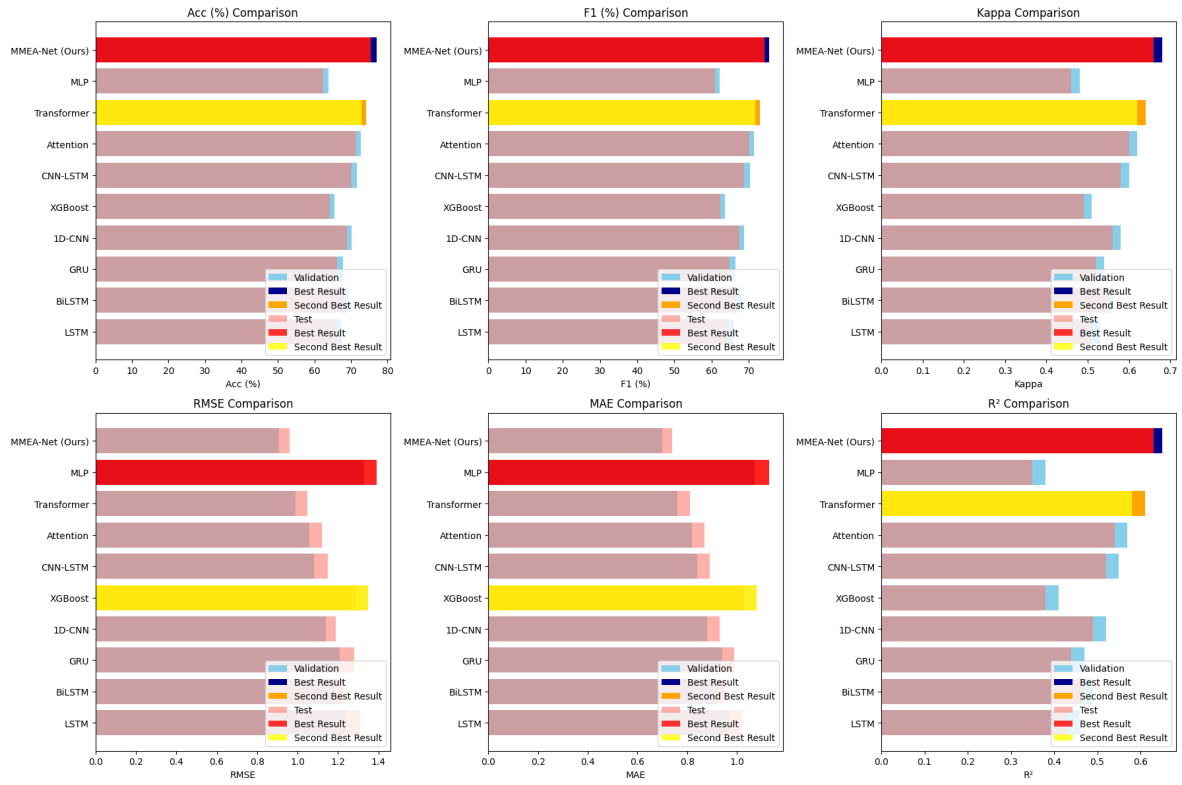

Figure 4. Performance comparison of MMEA-Net and baseline models on emotion classification and immersion prediction tasks.

Overall, these results demonstrate that the proposed framework effectively integrates multi-modal signals and jointly optimizes emotion and immersion objectives within a unified and balanced modeling strategy.

3.7.2. Model Component Analysis

To examine the contribution of individual components in MMEA-Net, we conducted ablation experiments by selectively removing or replacing key modules. Table 2 reports the corresponding results for both emotion classification and immersion prediction tasks.

Table 2. Ablation study on model components. Each row corresponds to a variant of MMEA-Net with a specific component removed or replaced.

Model Variant	Accuracy (%)	F1 (%)	Kappa	RMSE	MAE	R^2
MMEA-Net (Full Model)	75.42	74.19	0.66	0.96	0.74	0.63
w/o Hybrid-M	71.35	70.21	0.59	1.14	0.89	0.54
w/o Cross-Domain Fusion	72.14	70.88	0.61	1.08	0.84	0.56
w/o MFE	72.93	71.56	0.62	1.03	0.80	0.58

The ablation results indicate that each component contributes to the overall performance of MMEA-Net. Removing the Hybrid-M module leads to a noticeable degradation in both emotion classification accuracy and immersion prediction quality, reflecting its role in preserving modality-specific temporal characteristics. Excluding the Cross-Domain Fusion module also results in consistent performance drops, underscoring the importance of structured multimodal integration for jointly modeling affective and immersive states. In addition, the removal of the Multi-scale Feature Extraction (MFE) component negatively affects performance, confirming the necessity of capturing temporal patterns at multiple resolutions, particularly for physiological signals exhibiting diverse dynamics.

 doi:10.20944/preprints202601.1212.v1

3.7.3. Multimodal Contribution Analysis

To further analyze the influence of different data modalities, we evaluated MMEA-Net under single-modality, dual-modality, and full-modality settings. Table 3 summarizes the corresponding results.

Table 3. Ablation study on different modalities and their combinations.

Modalities Used	Accuracy (%)	F1 (%)	Kappa	RMSE	MAE	R^2
Eye-tracking only	65.37	63.92	0.51	1.28	1.02	0.44
ECG only	62.14	60.53	0.47	1.35	1.09	0.40
GSR only	59.86	58.12	0.45	1.41	1.15	0.36
Eye-tracking + ECG	71.29	69.87	0.60	1.09	0.86	0.56
Eye-tracking + GSR	70.45	68.92	0.59	1.12	0.89	0.54
ECG + GSR	68.73	67.18	0.56	1.18	0.94	0.51
All modalities	75.42	74.19	0.66	0.96	0.74	0.63

The results demonstrate that models relying on a single modality exhibit substantially lower performance than those using multimodal inputs, highlighting the benefit of integrating heterogeneous physiological and behavioral signals. Among individual modalities, eye-tracking yields the strongest performance, likely due to its direct association with user attention and interaction patterns. Dual-modality configurations consistently outperform single-modality setups, with the combination of eye-tracking and ECG achieving the best overall balance between classification and regression performance. Incorporating all three modalities further improves results across both tasks, indicating that complementary information from visual, cardiac, and autonomic signals jointly contributes to a more comprehensive representation of emotional and immersive states.

3.7.4. Single-Task vs. Multi-Task Learning

To further examine the impact of joint optimization, we compared single-task learning settings, where emotion classification or immersion prediction is performed independently, with the multi-task learning configuration used in MMEA-Net. The quantitative results are reported in Table 4.

Table 4. Comparison between single-task and multi-task learning approaches.

Learning Approach	Accuracy (%)	F1 (%)	Kappa	RMSE	MAE	R^2
Single-task (Emotion Classification)	73.81	72.54	0.63	–	–	–
Single-task (Immersion Prediction)	–	–	–	1.02	0.80	0.59
Multi-task Learning	75.42	74.19	0.66	0.96	0.74	0.63

The results show that the multi-task learning strategy yields consistent improvements over single-task learning for both objectives. Emotion classification accuracy increases by 1.61% compared to the single-task setting, while immersion prediction exhibits a reduction of 5.88% in RMSE. These gains suggest that jointly learning emotion and immersion enables the model to exploit shared and complementary information between the two tasks, leading to more informative representations. In addition, the multi-task formulation contributes to improved generalization and reduced overfitting, particularly under limited data conditions.

3.7.5. Summary

Through a comprehensive set of ablation studies, we evaluated the contribution of individual components, data modalities, and learning strategies in MMEA-Net. The experimental findings can be summarized as follows. First, the Hybrid-M and Cross-Domain Fusion modules play a central role in effectively encoding and integrating multimodal physiological and behavioral signals. Second, multi-scale feature extraction enhances the model's ability to capture temporal dynamics at different

 doi:10.20944/preprints202601.1212.v1

resolutions, which is especially important for physiological data. Third, multimodal configurations consistently outperform single-modality settings, demonstrating the benefit of leveraging complementary information from heterogeneous sources. Finally, the multi-task learning framework achieves better overall performance than single-task learning, indicating that emotion recognition and immersion estimation can mutually reinforce each other when modeled jointly. These observations highlight the effectiveness of multimodal, multi-scale, and multi-task learning strategies for emotion and immersion assessment in virtual reality environments.

3.8. Loss Weight Analysis

To investigate the balance between emotion classification and immersion prediction, we conducted experiments with different combinations of loss weights. The regularization weight λ_3 was fixed at 0.0001, while λ_1 and λ_2 were varied from 0 to 1 in increments of 0.2, subject to the constraint $\lambda_1 + \lambda_2 = 1$ to maintain a consistent overall loss scale. The experimental results are summarized in Table 5.

Table 5. Performance comparison under different loss weight settings for emotion classification (λ_1) and immersion prediction (λ_2).

λ_1	λ_2	Accuracy (%)	F1 (%)	Kappa	RMSE	MAE	R^2
0.8	0.2	74.65	73.28	0.64	1.15	0.92	0.49
0.6	0.4	75.21	73.93	0.65	1.02	0.81	0.58
0.4	0.6	75.42	74.19	0.66	0.96	0.74	0.63
0.2	0.8	74.83	73.65	0.65	0.98	0.76	0.62

As shown in Table 5, the configuration with $\lambda_1 = 0.4$ and $\lambda_2 = 0.6$ yields the best overall performance, achieving a balanced improvement in both emotion classification accuracy and immersion prediction error. When the loss function places full emphasis on a single task, performance on the other task degrades substantially, indicating that independent optimization fails to capture shared information between emotion and immersion.

A moderate emphasis on immersion prediction appears to benefit emotion classification performance, suggesting that immersion-related signals provide complementary contextual information that supports affective inference. This observation is consistent with findings in immersive experience research, where immersion often influences the intensity and expression of emotional responses.

These results demonstrate the importance of balanced task weighting in multi-task learning and confirm that appropriate coordination between emotion and immersion objectives leads to more stable and effective optimization. The loss weights $\lambda_1 = 0.4$ and $\lambda_2 = 0.6$ are therefore adopted in all subsequent experiments.

4. Downstream Task Extensions Based on the MMEA-Net Model

Beyond quantitative evaluation, the proposed MMEA-Net framework can be naturally extended to a range of downstream tasks that rely on joint emotion recognition and immersion estimation. These extensions demonstrate how multimodal affective and experiential representations can support adaptive decision-making and interaction design in virtual reality environments.

One representative extension is personalized art content recommendation driven by emotional and immersion analysis. By continuously estimating users' emotional states and immersion levels, the system can adapt recommended content to users' momentary affective conditions and engagement patterns. For example, when low emotional valence or declining immersion is detected, the recommendation strategy may shift toward more stimulating or interactive content to restore engagement. This task illustrates how joint emotion–immersion modeling can enhance personalization by incorporating experiential context rather than relying solely on static user preferences.

Another extension involves emotion- and immersion-aware dynamic art presentation. Using real-time predictions from MMEA-Net, presentation parameters such as pacing, scene transitions, or

 doi:10.20944/preprints202601.1212.v1

interaction intensity can be adjusted to respond to changes in users' emotional states or immersion levels. When decreasing engagement or affective responses are observed, the system can adapt the presentation flow to maintain continuity and user involvement. This extension highlights the applicability of the proposed framework to adaptive presentation systems that respond to experiential feedback.

MMEA-Net also supports real-time art content switching and optimization based on affective feedback. Continuous monitoring of emotional and immersion signals enables timely identification of disengagement or emotional shifts, triggering automatic content transitions or adjustments in visual and auditory elements. Such feedback-driven adaptation allows immersive systems to respond promptly to user state changes, providing a mechanism for maintaining sustained engagement during interactive experiences.

In addition, the multimodal representations learned by MMEA-Net can be leveraged to support art creation and curation processes. By analyzing aggregated emotional responses and immersion patterns across users, the system can provide feedback regarding how different artistic elements influence affective and experiential outcomes. This information can assist artists and curators in refining presentation strategies or content composition, translating multimodal physiological and behavioral data into actionable insights for creative optimization.

Finally, emotion and immersion analysis enables broader user experience enhancement strategies. By jointly modeling affective and experiential dimensions, interactive systems can dynamically adjust interface design, guidance mechanisms, or interaction modes to support emotional regulation and sustained engagement. This extension emphasizes the role of multimodal affective computing in experience-aware system design, where emotional and immersion feedback informs adaptive interaction strategies.

Collectively, these downstream task extensions demonstrate that MMEA-Net is not limited to isolated prediction tasks but can serve as a general affective–experiential perception module for adaptive virtual reality systems. They further illustrate how joint modeling of emotion and immersion can support personalized interaction, dynamic presentation, and experience-aware optimization in immersive environments.

5. Conclusion and Future Work

This paper presented a multimodal learning framework for joint emotion recognition and immersion estimation in virtual reality environments. By integrating eye-tracking, electrocardiogram, and galvanic skin response signals within a unified multi-task architecture, the proposed approach enables simultaneous modeling of users' affective states and immersive experiences. The experimental results demonstrate that jointly learning these two related tasks improves performance compared to single-task baselines, highlighting the benefit of shared representations for multimodal affective analysis.

The study further illustrated how joint emotion–immersion modeling can support a range of downstream tasks, including adaptive content recommendation, dynamic presentation adjustment, real-time content switching, creative feedback analysis, and user experience optimization in immersive environments. These extensions show that the learned multimodal representations are not limited to prediction tasks but can serve as a general perception component for experience-aware virtual reality systems.

Future work will focus on extending the framework to larger and more diverse multimodal datasets, as well as improving generalization across users, scenarios, and application domains. Incorporating long-term temporal modeling and contextual memory mechanisms may further enhance the understanding of sustained emotional and immersion trends. In addition, integrating the framework with real-world VR and AR platforms will allow investigation of its effectiveness in interactive and deployment-oriented settings. Overall, this work provides a structured approach to jointly modeling emotion and immersion, offering a basis for experience-aware system design in immersive computing.

 doi:10.20944/preprints202601.1212.v1

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

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. The extended VREED dataset with immersion labels used in this study is not publicly available due to participant privacy protection agreements.

Acknowledgments: The authors would like to thank Shandong University of Arts for providing institutional support and a supportive academic environment for this research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Lønne, T.; Karlsen, H.; Langvik, E.; Saksvik-Lehouillier, I. The effect of immersion on sense of presence and affect when experiencing an educational scenario in virtual reality: A randomized controlled study. *Heliyon* **2023**, *9*, e16928. https://doi.org/10.1016/j.heliyon.2023.e16928.
2. Yang, X.; Cheng, P.; Liu, X.; Shih, S. The impact of immersive virtual reality on art education: A study of flow state, cognitive load, brain state, and motivation. *Education and Information Technologies* **2024**, *29*, 6087–6106. https://doi.org/10.1007/s10639-023-12168-2.
3. Marín-Morales, J.; Llinares, C.; Guixeres, J.; Alcañiz, M. Emotion Recognition in Immersive Virtual Reality: From Statistics to Affective Computing. *Sensors* **2020**, *20*, 5163. https://doi.org/10.3390/s20185163.
4. Chen, Z.; Han, Z.; Wu, L.; Huang, J. Multisensory Imagery Enhances the Aesthetic Evaluation of Paintings: A Virtual Reality Study. *Empirical Studies of the Arts* **2026**, p. 02762374251412761.
5. Cai, Y.; Li, X.; Li, J. Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review. *Sensors* **2023**, *23*, 2455. https://doi.org/10.3390/s23052455.
6. Moin, A.; Aadil, F.; Ali, Z.; Kang, D. Emotion recognition framework using multiple modalities for an effective human-computer interaction. *Journal of Supercomputing* **2023**. https://doi.org/10.1007/s11227-023-05233-6.
7. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Zhang, W. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion* **2022**, *83*, 19–52. https://doi.org/10.1016/j.inffus.2022.01.007.
8. Lin, W.; Li, C. Review of studies on emotion recognition and judgment based on physiological signals. *Applied Sciences* **2023**, *13*, 2573. https://doi.org/10.3390/app13042573.
9. Ezzameli, K.; Mahersia, H. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion* **2023**, *99*, 101847. https://doi.org/10.1016/j.inffus.2023.101847.
10. Fu, Z.; Zhang, B.; He, X.; Li, Y.; Wang, H.; Huang, J. Emotion recognition based on multi-modal physiological signals and transfer learning. *Frontiers in Neuroscience* **2022**, *16*, 1000716. https://doi.org/10.3389/fnins.2022.1000716.
11. Lee, Y.; Pae, D.; Hong, D.; Lim, M.; Kang, T. Emotion recognition with short-period physiological signals using bimodal sparse autoencoders. *Intelligent Automation and Soft Computing* **2022**, *32*, 657–673. https://doi.org/10.32604/iasc.2022.026243.
12. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.; Yazdani, A.; Ebrahimi, T.; Patras, I. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing* **2011**, *3*, 18–31. https://doi.org/10.1109/T-AFFC.2011.15.
13. Tabbaa, L.; Searle, R.; Bafti, S.; Hossain, M.; Intarasisrisawat, J.; Glancy, M.; Ang, C. VREED: Virtual reality emotion recognition dataset using eye tracking and physiological measures. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **2021**, *5*, 1–20. https://doi.org/10.1145/3494986.
14. Yahyaeian, A.A.; Sabet, M.; Zhang, J.; et al. Enhancing Immersive Learning: An Exploratory Pilot Study on Large Language Model-Powered Guidance in Virtual Reality Labs. *Computer Applications in Engineering Education* **2026**, *34*, e70127.

 doi:10.20944/preprints202601.1212.v1

15. De Giglio, V.; Evangelista, A.; Giannakakis, G.; Konstantaras, A.; Kamarianakis, Z.; Uva, A.E.; Manghisi, V.M. Assessing the Impact of Cinematic Virtual Reality Simulations on Young Drivers: Behavior and Physiological Responses. *Virtual Reality* **2026**, *30*, 11.
16. García-Batista, Z.E.; Guerra-Peña, K.; Jurnet, I.A.; Cano-Vindel, A.; Álvarez-Hernández, A.; Herrera-Martinez, S.; Medrano, L.A. Design and Preliminary Evaluation of AYRE: A Virtual Reality-Based Intervention for the Treatment of Emotional Disorders. *Journal of Behavioral and Cognitive Therapy* **2026**, *36*, 100560.
17. Wei, L.; Liu, L.; Faridniya, H. Promoting Mental Health and Preventing Emotional Disorders in Vulnerable Adolescent Girls through VR-Based Extreme Sports. *Acta Psychologica* **2026**, *262*, 106088.
18. Baker, N.A.; Polhemus, A.H.; Baird, J.M.; Kenney, M. Embodied Fully Immersive Virtual Reality as a Therapeutic Modality to Treat Chronic Pain: A Scoping Review. *Virtual Worlds* **2026**, *5*, 3.
19. Hernandez-Melgarejo, G.; Luviano-Juarez, A.; Fuentes-Aguilar, R. A framework to model and control the state of presence in virtual reality systems. *IEEE Transactions on Affective Computing* **2022**, *13*, 1854–1867. https://doi.org/10.1109/TAFFC.2021.3100679.
20. Ochs, C.; Sonderegger, A. The interplay between presence and learning. *Frontiers in Virtual Reality* **2022**, *3*, 742509. https://doi.org/10.3389/frvir.2022.742509.
21. Liu, B.; Liu, X.; Jin, X.; Stone, P.; Liu, Q. Conflict-averse gradient descent for multi-task learning. In Proceedings of the Advances in Neural Information Processing Systems, 2021, Vol. 34, pp. 18878–18890.
22. Liu, D.; Yu, Y. MT2ST: Adaptive Multi-Task to Single-Task Learning. *arXiv preprint* **2024**, [2406.18038].
23. 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. https://doi.org/10.1007/s11042-022-12325-3.
24. Katada, S.; Okada, S. Biosignal-based user-independent recognition of emotion and personality with importance weighting. *Multimedia Tools and Applications* **2022**, *81*, 30219–30241. https://doi.org/10.1007/s11042-022-12103-1.
25. Dissanayake, V.; Seneviratne, S.; Rana, R.; Wen, E.; Kaluarachchi, T.; Nanayakkara, S. SigRep: Toward robust wearable emotion recognition with contrastive representation learning. *IEEE Access* **2022**, *10*, 18105–18120. https://doi.org/10.1109/ACCESS.2022.3151106.
26. Yan, J.; Zheng, W.; Xu, Q.; Lu, G.; Li, H.; Wang, B. Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech. *IEEE Transactions on Multimedia* **2016**, *18*, 1319–1329. https://doi.org/10.1109/TMM.2016.2557721.
27. Wang, M.; Lin, Y.; Wang, S. The Nature Diagnosability of Bubble-Sort Star Graphs under the PMC Model and MM Model. *Int. J. Eng. Appl. Sci.* **2017**, *4*.
28. Wang, S.; Wang, Z.; Wang, M.; Han, W. g-Good-Neighbor Conditional Diagnosability of Star Graph Networks under PMC Model and MM Model. *Frontiers of Mathematics in China* **2017**, *12*, 1221–1234.
29. Wang, M.; Lin, Y.; Wang, S.; Wang, M. Sufficient Conditions for Graphs to Be Maximally 4-Restricted Edge Connected. *Australas. J. Comb.* **2018**, *70*, 123–136.
30. Wang, M.; Wang, S. Connectivity and Diagnosability of Center k-Ary n-Cubes. *Discrete Applied Mathematics* **2021**, *294*, 98–107.
31. Wang, M.; Xiang, D.; Wang, S. Connectivity and Diagnosability of Leaf-Sort Graphs. *Parallel Processing Letters* **2020**, *30*, 2040004.
32. Xiang, D.; Hsieh, S.Y. G-Good-Neighbor Diagnosability under the Modified Comparison Model for Multiprocessor Systems. *Theoretical Computer Science* **2025**, *1028*, 115027.
33. 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.
34. Hu, Z.; Chen, L.; Luo, Y.; Zhou, J. EEG-based emotion recognition using convolutional recurrent neural network with multi-head self-attention. *Applied Sciences* **2022**, *12*, 11255. https://doi.org/10.3390/app122111255.
35. Xiao, G.; Shi, M.; Ye, M.; Xu, B.; Chen, Z.; Ren, Q. 4D attention-based neural network for EEG emotion recognition. *Cognitive Neurodynamics* **2022**, pp. 1–14. https://doi.org/10.1007/s11571-022-09860-4.
36. Alharbi, H. Explainable feature selection and deep learning based emotion recognition in virtual reality using eye tracker and physiological data. *Frontiers in Medicine* **2024**, *11*, 1438720. https://doi.org/10.3389/fmed.2024.1438720.
37. Souza, V.; Maciel, A.; Nedel, L.; Kopper, R. Measuring presence in virtual environments: A survey. *ACM Computing Surveys* **2021**, *54*, 1–37. https://doi.org/10.1145/3466817.

 doi:10.20944/preprints202601.1212.v1

38. Liu, X.; Zhou, H.; Liu, J. Deep learning-based analysis of the influence of illustration design on emotions in immersive art. *Mobile Information Systems* **2022**, *2022*, 3120955. https://doi.org/10.1155/2022/3120955.
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, 2017, Vol. 30, pp. 5998–6008.
40. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Liu, P. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **2020**, *21*, 1–67.
41. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* **2020**, [2010.11929].
42. Li, G.; Bai, L.; Zhang, H.; Xu, Q.; Zhou, Y.; Gao, Y.; Wang, M.; Li, Z. Velocity Anomalies around the Mantle Transition Zone beneath the Qiangtang Terrane, Central Tibetan Plateau from Triplicated P Waveforms. *Earth and Space Science* **2022**, *9*, e2021EA002060.
43. Gers, F.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Computation* **2000**, *12*, 2451–2471. https://doi.org/10.1162/089976600300015015.
44. Siami-Namini, S.; Tavakoli, N.; Namin, A. 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.
45. Rana, R. Gated Recurrent Unit (GRU) for Emotion Classification from Noisy Speech. *arXiv preprint* **2016**, [1612.07778].
46. 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.
47. Nielsen, D. Tree Boosting with XGBoost—Why Does XGBoost Win "Every" Machine Learning Competition? Master's thesis, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, 2016.
48. 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.
49. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in Transformer. In Proceedings of the Advances in Neural Information Processing Systems, 2021, Vol. 34, pp. 15908–15919.
50. Valanarasu, J.; Patel, V. 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.