SAPIENZA
UNIVERSITÀ DI ROMA

LLM-Powered Emotion Recognition from Music-Evoked EEG Signals

Faculty of Information Engineering, Informatics and Statistics
Master's Degree in Artificial Intelligence and Robotics

Antonello Giorgio
ID number 1836529

Advisor
Prof. Danilo Comminiello

Academic Year 2024/2025

Thesis defended on 22 July 2025
in front of a Board of Examiners composed by:

Prof. Riccardo Rosati (chairman)
Prof. Irene Amerini
Prof. Silvia Bonomi
Prof. Marco Console
Prof. Danilo Comminiello
Prof. Luca Di Giammarino
Prof. Nicola Galesi

LLM-Powered Emotion Recognition from Music-Evoked EEG Signals
Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: giorgio.1836529@studenti.uniroma1.it

Abstract

Emotion-aware technology is transforming adaptive gaming [1], personalized media, neuro-feedback [2], and mental-health monitoring [3], yet reliably decoding human emotions from brain signals remains difficult. Electroencephalography (EEG), which measures cortical electrical activity with millisecond precision [4], offers a rich view into neural dynamics, but its high dimensionality, low signal-to-noise ratio, and strong inter-subject variability challenge conventional models [5].

Therefore, in this thesis, I propose a dual-branch neural architecture that combines the temporal sensitivity of Large Language Models (LLMs) with the spatial reasoning of Dynamic Graph Convolutional Neural Networks (DGCNNs). I build on the DEAP dataset [6], which captures EEG responses to carefully curated music videos designed to evoke emotional reactions. After segmenting and normalizing the signals, I process each segment through two complementary paths. The temporal branch uses an EEG-Transformer coupled with an autoregressive LLM decoder to reconstruct masked signal windows, distilling temporal structure in a self-supervised fashion. In parallel, a DGCNN learns inter-electrode connectivity patterns. Their predictions are combined via logit averaging, with a confidence gate favouring the more reliable branch per sample.

Training is guided by cross-entropy objectives and a reconstruction loss on the LLM outputs. The experiments with GPT-2 and LLaMA variants show that this hybrid model achieves state-of-the-art accuracy and macro-F1 on arousal–valence classification, surpassing prior DEAP baseliness. These results demonstrate that regularized temporal modeling with LLM and spatial reasoning based on graphs are mutually renforcing in decoding affective states from EEG. I argue that LLMs represent a promising tool for analyzing EEG signals and predicting emotional states and their potential in affective computing deserves further exploration, particularly for generative applications such as emotion-driven music synthesis guided by real-time EEG decoding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Why emotion–aware AI matters

From adaptive videogames that modulate difficulty in real time [1] to clinical neurofeedback systems that monitor patient engagement [15], and personalized music-therapy tools that remix tracks to enhance well-being [16, 17], the ability to sense and interpret human emotions is reshaping the design of interactive technologies.

The growing availability of wearable sensors and the diffusion of affective computing toolkits confirm both the scientific interest and commercial potential of emotion-aware systems [3]. Yet, despite decades of progress, *reliable* automatic emotion recognition remains elusive. Visual cues, such as facial expressions, body posture or micro-expressions, can be occluded or deliberately masked. Peripheral physiological signals like galvanic skin response (GSR) or heart-rate variability (HRV) offer limited emotional coverage.

A more direct window into affective processing is offered by electroencephalography (EEG), whose millisecond-scale temporal resolution captures the dynamics of neural populations [4]. However, EEG data pose significant modelling challenges due to their high dimensionality, low signal-to-noise ratio and strong inter-subject variability [5], which often confound traditional machine-learning pipelines.

1.2 From handcrafted features to foundation models

Historically, EEG-based emotion classifiers relied on handcrafted features, such as spectral band power, differential asymmetry indices, or entropy measures, fed into shallow learners like support vector machines (SVMs) or k-NN classifiers [18]. While conceptually simple, these approaches require extensive preprocessing and expert knowledge, and often fail to generalize across subjects or conditions.

The rise of deep learning has led to the adoption of convolutional and recurrent architectures that learn hierarchical features directly from raw EEG signals [19, 20]. Nevertheless, two key limitations persist. First, purely *temporal* models, such as 1D-CNNs, LSTMs, and Transformer encoders, often neglect spatial dependencies across electrodes, overlooking important topological information. Second, most deep networks are trained from scratch on small EEG datasets, risking overfitting and ignoring the inductive biases embedded in large-scale pretraining.

Large Language Models (LLMs) such as GPT-2 and LLaMA have shown strong capabilities in learning long-range temporal dependencies via autoregressive training [21]. Although originally developed for natural language, their attention mechanisms operate independently of the data modality, requiring only a sequential embedding format. This makes them applicable not just to text, but also to signals like EEG [22, 23], audio, or time series, provided the input is suitably encoded. In parallel, Dynamic Graph Convolutional Neural Networks (DGCNNs) provide a principled framework for modeling spatial interactions across electrodes by learning data-driven graph topologies [24].

In this thesis, I combine these two approaches, proposing a dual-branch architecture that integrates temporal modeling with LLMs and spatial reasoning with DGCNNs into a single, end-to-end pipeline.

1.3 Research goal and challenges

I tackle the task of emotion classification on the publicly available DEAP dataset [6], focusing on binary and quaternary classification in the arousal–valence space [25]. This task entails several interconnected challenges:

1. **Noisy, high-dimensional input**: 32-channel EEG signals sampled at 128 Hz for 60 s yield over 240000 values per trial.
2. **Scarce data regime**: DEAP includes only 32 participants, exacerbating the risk of subject-specific overfitting.
3. **Temporal–spatial trade-off**: Most models specialize in either temporal pattern mining or spatial structure extraction, but not both.
4. **Scalability vs. adaptability**: Foundation models are powerful but resource-intensive; fine-tuning them on EEG data requires parameter-efficient techniques like LoRA [26].

1.4 Proposed solution

I address these challenges with a dual-branch neural architecture composed of two specialized encoders and a confidence-aware fusion head:

- *Temporal branch* — an `EEGTransformer` encodes short EEG chunks, which are then reconstructed by an autoregressive LLM (GPT-2, LLaMA-3.1 8B AWQ INT4 or LLaMA-3.2 1B) using a masked latent prediction task. The reconstruction loss regularises temporal dynamics.
- *Spatial branch* — a DGCNN operates on the same EEG segment, learning functional connectivity graphs via end-to-end optimization.
- *Fusion module* — logits from both branches are averaged and gated using a sample-wise confidence indicator to select the more reliable prediction stream.

Training strategy. Both branches are trained jointly using cross-entropy loss, while the LLM additionally minimizes a reconstruction loss on the masked temporal embeddings. Parameter-efficient fine-tuning via LoRA enables adaptation of billion-scale models using only a few megabytes of task-specific parameters [27].

1.5 Contributions

This thesis makes the following original **contributions**:

1. A Transformer-based architecture is introduced for EEG emotion recognition, where temporal representations are regularized through a Large Language Model (LLM) fine-tuned on masked latent embeddings. This represents, to the best of current knowledge, the first application of autoregressive LLM fine-tuning to biosignal analysis.

2. A dual-branch design is proposed, combining temporal (LLM-based) and spatial (graph-based) encoders in a parameter-efficient manner that leverages Low-Rank Adaptation (LoRA).

3. A confidence-aware logit fusion strategy is developed, enabling the model to dynamically prioritize the more informative prediction stream on a per-sample basis.

4. The system achieves new state-of-the-art performance on the DEAP dataset, outperforming recent multimodal and graph-based baselines in both accuracy and macro-F1 score.

1.6 Thesis roadmap

The rest of this thesis is organized as follows:

Chapter 2 introduces the theoretical and technical foundations of the work, covering neurophysiology, emotion modeling, EEG-based affect recognition, and recent advances in Transformers, Graph Neural Networks, and parameter-efficient fine-tuning.

Chapter 3 presents the proposed architecture in detail, describing the DEAP dataset, preprocessing pipeline, and model components.

Chapter 4 presents the experimental protocol, evaluation setup, and a comprehensive set of quantitative results. It includes comparisons with state-of-the-art baselines and ablation studies.

Chapter 5 summarizes the main contributions and outlines future directions, including the use of fine-tuned LLMs for EEG emotion recognition and potential multimodal extensions for real-time emotion-driven applications.

Chapter 2

Background

2.1 Preliminary Concepts

2.1.1 Electroencephalography (EEG)

EEG is a non-invasive technique for recording the brain's electrical activity using electrodes placed on the scalp. It measures the neural oscillations evoked by postsynaptic potentials within large groups of neurons, providing high temporal resolution (milliseconds) but relatively poor spatial resolution [4]. EEG has been widely applied in clinical diagnostics, cognitive neuroscience, and emerging fields like affective computing and brain-computer interfaces (BCIs) due to its non-invasive nature, cost-effectiveness compared to other neuroimaging methods, portability, and high temporal resolution [5].

EEG systems follow standardized protocols, such as the *10-20 International System* (see Figure 2.1), to ensure the reproducible placement of electrodes across different subjects and studies [28].

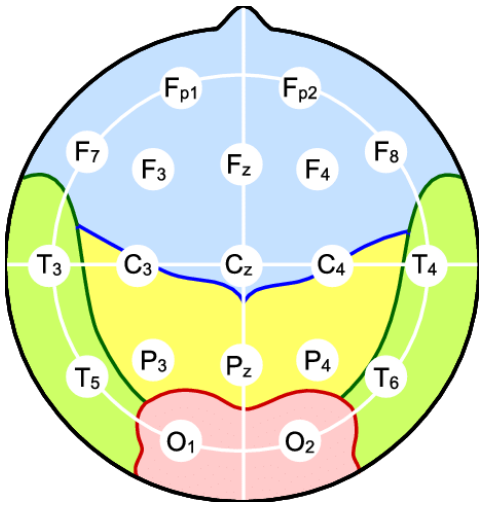


Figure 2.1. The 10-20 electrode configuration used for collecting EEG data. This system corresponds to specific regions of the scalp and helps localize brain activity during emotion processing.

EEG systems can vary in electrode density, ranging from inexpensive systems with fewer electrodes to high-density configurations with over 64 electrodes [29]. Recent research increasingly utilizes high-resolution spatial arrays to improve the localization of brain activity sources.

EEG signals are susceptible to a variety of physiological and environmental noise, such as eye movements (electrooculogram), muscle activity (electromyogram), and cardiac rhythms (electrocardiogram). Signal processing methods like band-pass filtering, artifact subspace reconstruction, and Independent Component Analysis (ICA) are crucial to remove these artifacts [30].

After preprocessing, feature extraction techniques transform raw EEG data into meaningful representations. Common methods include time-domain features (e.g., statistical moments), frequency-domain features (e.g., Power Spectral Density (PSD)), and time-frequency transformations (e.g., Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT)) [31, 32]. Feature selection often depends on the specific application, and emotional decoding can benefit from both frequency decomposition (including alpha, beta, and gamma bands) and spatial-temporal patterns [33].

EEG-based affective computing research typically relies on public datasets like DEAP [6] and SEED [19], which provide multi-channel EEG recordings annotated with emotional ratings. These datasets have become benchmarks for developing and evaluating new models, enabling fair comparisons across different approaches.

Recent advancements have explored real-time EEG applications using wearable and portable EEG devices. These systems enable applications in neurofeedback, immersive environments, and educational tools like the *NeuroHumanities Lab* [34]. Such systems demonstrate EEG's applicability beyond laboratory settings, providing context-sensitive emotional state detection.

2.1.2 Emotion Recognition Models

Emotion recognition from physiological signals, particularly EEG, has garnered significant attention in recent years due to its potential applications in healthcare, education, human-computer interaction, and entertainment. At the core of these systems lies the task of mapping complex biosignals to interpretable emotional states using computational models.

Emotion models are generally categorized into two primary frameworks: discrete and dimensional. Discrete models classify emotions into basic categories, such as happiness, sadness, anger, fear, surprise, and disgust, often inspired by Ekman's six universal emotions [12] (see Figure 2.12). On the other hand, dimensional models describe emotions within a continuous space, most commonly the Valence-Arousal (VA) or Valence-Arousal-Dominance (VAD) frameworks. Russell's Circumplex Model (see Figure 2.13) is a well-known 2D representation where valence measures pleasure-displeasure and arousal reflects the intensity of the emotion [25]. The VAD model introduces dominance, capturing the degree of control associated with the emotion.

To predict these emotional states from EEG signals, various machine learning techniques have been employed. Traditional methods include Support Vector Machines (SVM), k-Nearest Neighbors (kNN), Decision Trees (DT), and Random Forests (RF), which are typically paired with handcrafted feature sets [18, 32]. These

models have demonstrated robust performance, especially in scenarios with small datasets or when interpretability is important.

With the advent of deep learning, models capable of automatic feature extraction have gained prominence. Convolutional Neural Networks (CNNs) are well-suited for learning spatial patterns in EEG topographies, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, are effective in capturing temporal dependencies [20, 19]. More recent approaches leverage Graph Neural Networks (GNNs) to model spatial relationships between EEG channels, treating the brain as a functional connectivity graph [24, 35].

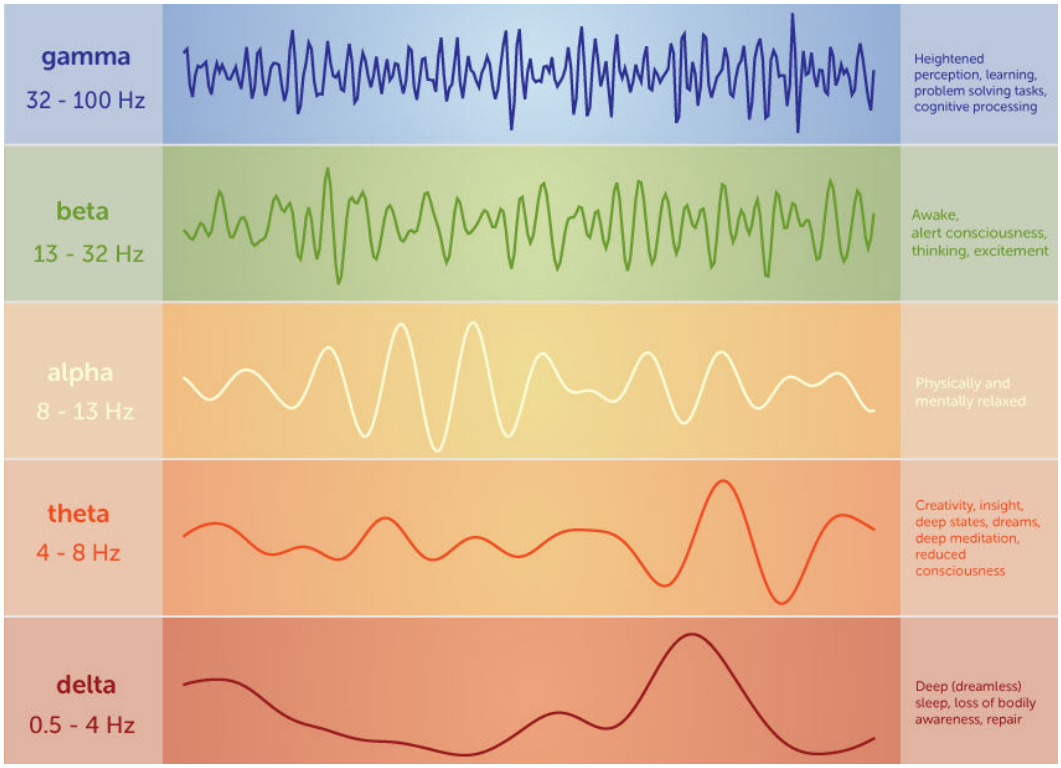


Figure 2.2. EEG frequency bands used in emotion recognition: Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–32 Hz) and Gamma (32–45 Hz). These bands are commonly associated with emotional and cognitive states.

Feature extraction remains a pivotal step, especially in hybrid systems that integrate both traditional and deep learning components. Spectral features, such as band powers (see Figure 2.2), are among the most informative for emotion recognition. Statistical features (mean, variance, standard deviation) and connectivity measures (e.g., Pearson correlation, phase-locking value, coherence) are also widely used [36, 37].

Modern models often combine spatial, temporal, and frequency-domain information. For instance, multimodal neural networks fuse EEG with other biosignals, and attention-based architectures dynamically weigh informative channels or time segments. These advances contribute to state-of-the-art results on benchmark datasets

like DEAP, SEED, and DREAMER.

2.1.3 Transformer Architectures and LLMs

Transformer models represent a paradigm shift in sequential data processing. Introduced by Vaswani et al. in 2017 [7], transformers eliminate the need for recurrence and convolutions, relying instead on a mechanism called self-attention to process input sequences in parallel and model long-range dependencies. The self-attention operation evaluates the relationships between all elements in a sequence, assigning context-aware weights that allow the model to capture intricate dependencies.

The transformer architecture consists of multi-head attention (MHA), feed-forward networks, layer normalization, and residual connections (see Figure 2.3). MHA enhances the model's capacity to attend to information from different subspaces simultaneously. However, since self-attention is permutation-equivariant, positional encodings are necessary to retain sequence order. These can be sinusoidal (as in the original model), learnable, or relative (e.g., ALiBi, RoPE).

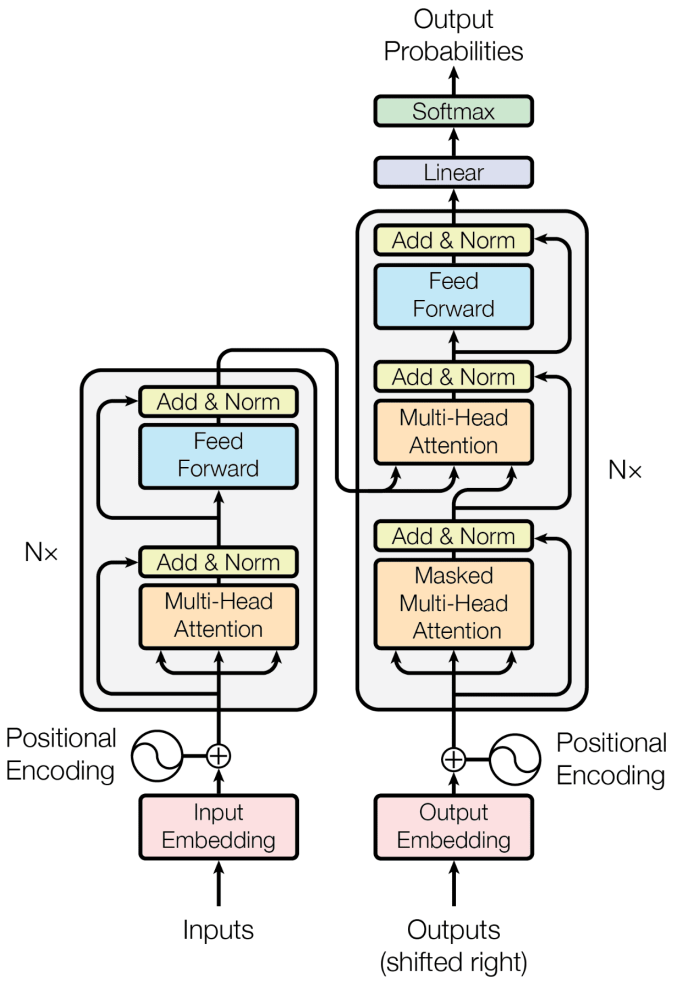


Figure 2.3. Standard Transformer encoder-decoder architecture as introduced in [7].

Compared to RNNs, transformers are more scalable and parallelizable, which has led to their dominance in NLP and expansion into other domains such as computer vision [13], audio, and graph-structured data. The Vision Transformer (ViT, see Figure 2.20), for instance, splits images into patches and treats them as tokens, applying standard transformer blocks for classification tasks.

Large Language Models (LLMs) like GPT-2 [38], and LLaMA 3.1 [39] (see comparison in Figure 2.21) are deep transformer networks trained on vast text corpora using self-supervised learning. These models yield contextual embeddings-word representations that change depending on context, enabling powerful transfer learning capabilities. LLMs scale exceptionally well with data and compute, exhibiting emergent abilities such as in-context learning, zero-shot generalization, and language reasoning.

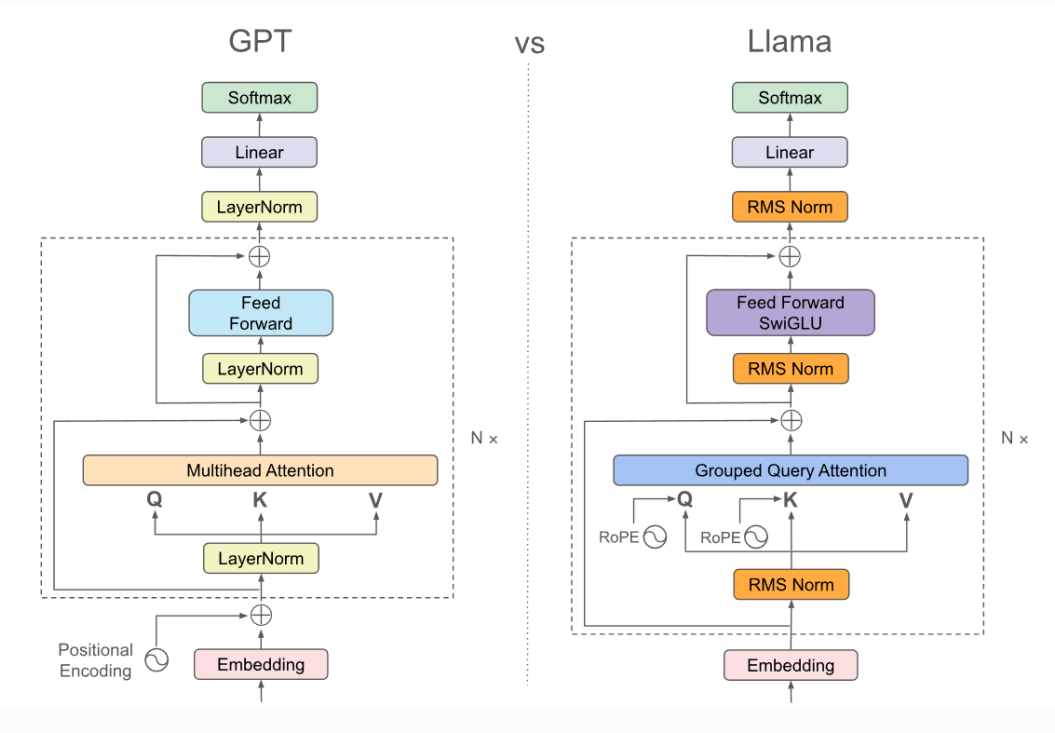


Figure 2.4. Architectural comparison between GPT and LLaMA. LLaMA replaces LayerNorm with RMSNorm, introduces SwiGLU in the feed-forward layers, and adopts Grouped Query Attention along with rotary positional encodings (RoPE), making it more efficient in both training and inference.

Adapting transformers to biosignals like EEG involves embedding multichannel temporal data into token sequences, applying positional encoding, and fine-tuning attention mechanisms to capture spatial-temporal structure [40, 24]. Some works use ViT-style approaches for EEG-based emotion recognition, converting EEG into 2D representations (e.g., wavelet-based images) [41]. Others apply the transformer directly to 1D raw EEG sequences with promising results.

Recent developments bridge LLMs and EEG processing through models such as EEG-GPT [22], which repurpose LLMs for EEG classification and interpretation, employing masked prediction and prompt tuning. These models show potential in

cross-subject generalization, cognitive state decoding, and interpretability. Transfer learning using contrastive EEG-text masked autoencoders is also gaining traction [42], enabling EEG-to-text decoding and multimodal applications.

Together, these advances suggest that transformers and LLMs provide a versatile foundation for modeling complex brain signals. Their ability to learn from raw data, capture long-range dependencies, and generalize across tasks makes them particularly well suited for EEG-based emotion recognition and other affective computing applications.

2.1.4 Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning technique designed to adapt large pretrained models, such as transformer-based LLMs, to downstream tasks without updating the full set of model parameters [26]. The key idea is to introduce a pair of low-rank trainable matrices into each attention and feed-forward layer, effectively approximating the weight update in a constrained subspace. This approach significantly reduces the number of parameters that need to be trained, resulting in lower memory and computational requirements while maintaining task performance.

LoRA modifies the original weight matrix W in a neural network layer by decomposing the update ΔW into a product of two low-rank matrices: $\Delta W = AB$, where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$ with $r \ll d, k$ (see Figure 2.5). The original weight is kept frozen, and only A and B are optimized during training. This leads to efficient fine-tuning even for extremely large models like GPT-3 or LLaMA [27].

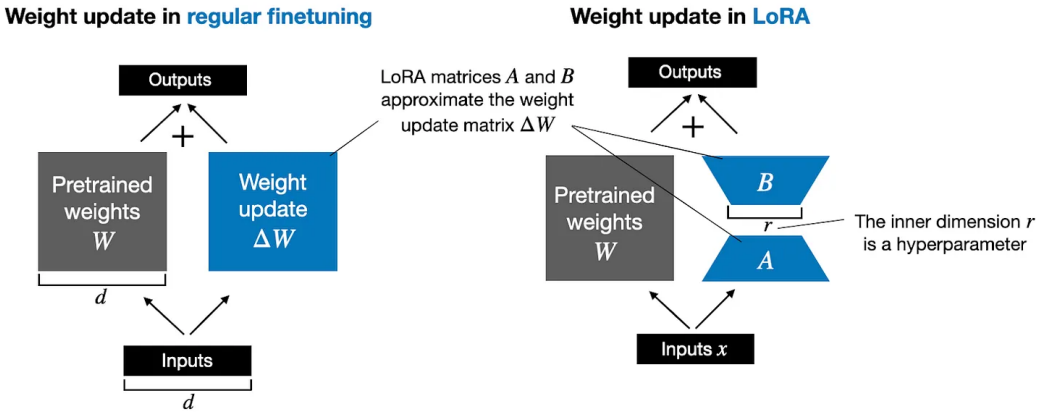


Figure 2.5. Low-Rank Adaptation (LoRA) inserts trainable rank-decomposed matrices into frozen transformer weights. This approach enables parameter-efficient fine-tuning of large models on limited-domain data.

In scenarios involving biosignals such as EEG, data availability is often limited due to the complexity and cost of acquiring labeled samples. LoRA offers an effective solution by enabling transfer learning from general-purpose language models to specialized domains, including affective computing and cognitive state decoding [43, 22]. When used with EEG data, LoRA facilitates domain adaptation by

finetuning only the most relevant model components, thus mitigating overfitting and reducing the computational cost of full model retraining.

Moreover, LoRA can be combined with other parameter-efficient strategies like prompt tuning, adapter modules, and quantization (as in QLoRA), further enhancing its utility in resource-constrained settings [44]. This makes it a compelling choice not only for language applications, but also for domains requiring lightweight adaptation of large models, including neuroscience and biomedical signal processing [26].

2.1.5 Convolutional Neural Networks for EEG

Convolutional Neural Networks (CNNs) are a foundational deep learning architecture originally developed for computer vision tasks, but widely adapted for time series, audio, and biomedical signal processing. CNNs operate by applying learnable convolutional filters across input data, enabling them to extract hierarchical spatial and temporal patterns [45, 46].

A convolutional layer (see Figure 2.6) applies cross-correlation operations over local regions of the input, using filters that are optimized during training to detect relevant patterns. Padding and stride parameters allow control over the spatial resolution of outputs. Pooling layers, such as max or average pooling, reduce the dimensionality and introduce translational invariance. Together, convolution and pooling layers form the core of CNN feature extractors.

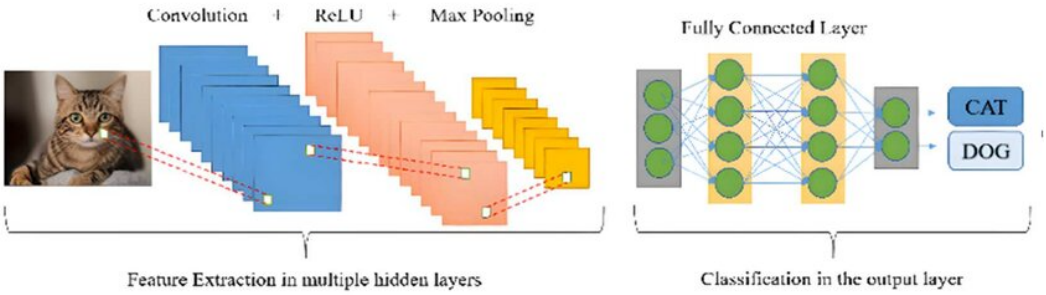


Figure 2.6. Architecture of a simple Convolutional Neural Network (CNN), consisting of convolutional layers, pooling layers, and fully connected layers. CNNs are widely used for extracting spatial features from structured inputs such as images and EEG topographies.

In EEG processing, CNNs are typically used to capture spatial dependencies between electrodes and temporal dynamics within signal windows. Depending on preprocessing, input representations can be raw 1D time series, 2D topographical maps, or even 3D tensors (time $\times$ channels $\times$ trials). CNNs applied to EEG are capable of learning frequency-specific activations and channel correlations without handcrafted features [31].

Several architectures have been explored in the context of EEG:

- **Temporal CNNs (TCNs)**: Use 1D convolutions over time to extract temporal dynamics from raw EEG signals [20].
- **2D CNNs**: Treat EEG as an image by projecting electrodes onto a 2D grid or topomap (e.g., based on 10-20 layout), then apply 2D filters [19].

- **Hybrid CNNs**: Combine convolutional layers with recurrent networks or graph-based models for richer spatial-temporal representations.

Deeper CNNs, such as VGG (see Figure 2.7) and ResNet (see Figure 2.8 and Figure 2.9), have been adapted to EEG by modifying kernel sizes, receptive fields, and downsampling strategies. Recent efforts also introduce dilated convolutions and residual connections to improve representation depth without increasing parameters.

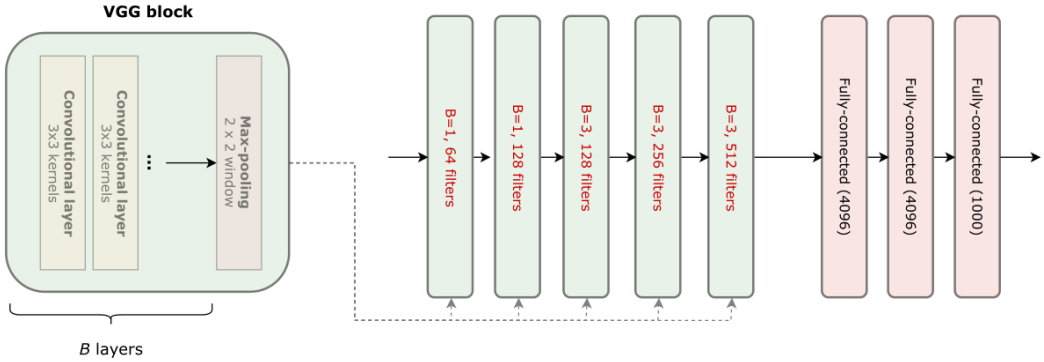


Figure 2.7. The VGG architecture, known for its simplicity and use of small 3×3 convolutional filters. It stacks multiple convolutional layers before downsampling via pooling and concludes with fully connected layers [8].

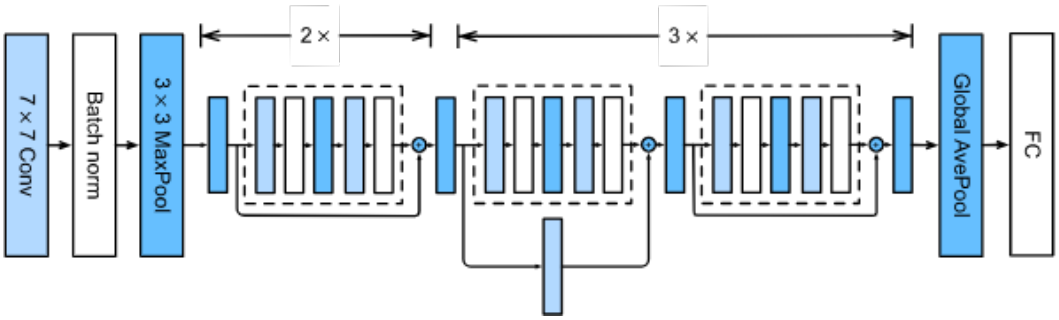


Figure 2.8. The overall ResNet architecture. ResNet introduces skip connections to allow gradients to propagate through deeper networks without vanishing, enabling very deep models such as ResNet-50 and ResNet-101 [9].

Regularization is especially important in CNN-based EEG systems to avoid overfitting due to limited data. Techniques such as dropout, batch normalization, and data augmentation (e.g., MixUp, CutMix) have been applied to improve generalization [47, 48, 49]. Furthermore, the modularity of CNNs makes them suitable for integration into larger multimodal architectures, such as CNN-GNN hybrids or CNN-Transformer stacks.

Overall, CNNs remain a powerful baseline and component for EEG-based emotion recognition pipelines due to their efficiency, flexibility, and ability to learn hierarchical representations from raw or preprocessed signals.

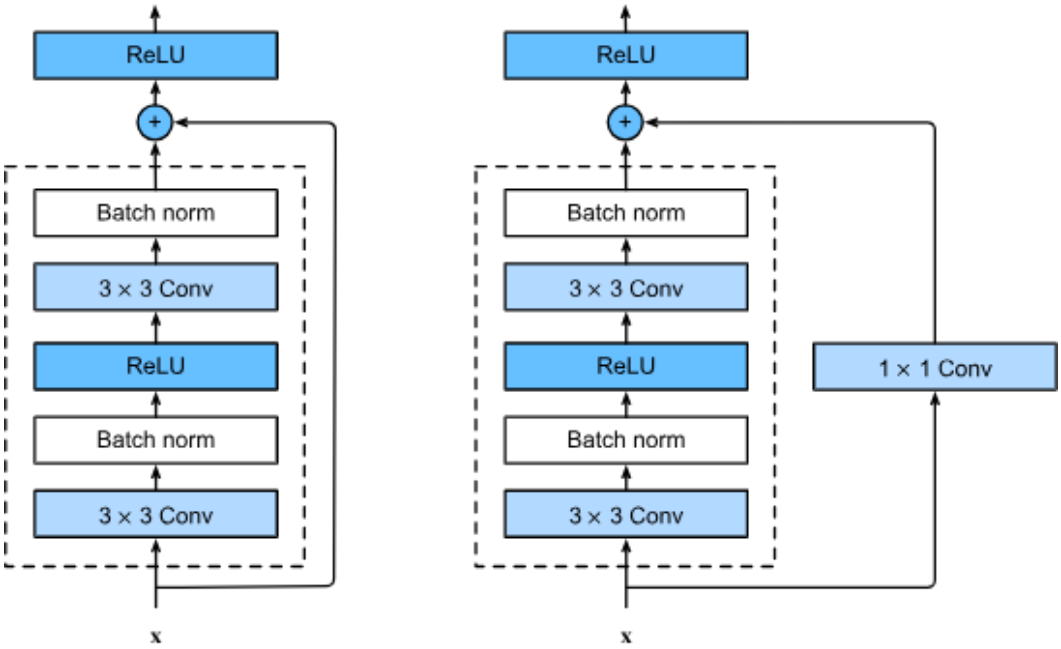


Figure 2.9. A ResNet residual block with a skip connection. The input x is added to the output of the convolutional layers, allowing the network to learn residual mappings instead of direct transformations.

2.1.6 Graph Neural Networks for EEG

Graph Neural Networks (GNNs) represent a class of deep learning models specifically designed to operate on graph-structured data. In the context of EEG signal analysis, GNNs have emerged as a powerful framework for modeling the complex spatial dependencies between electrodes and capturing topological characteristics of brain activity [50].

Unlike CNNs, which operate on regular grid data (e.g., images), GNNs generalize convolutional operations to non-Euclidean domains, such as graphs composed of nodes (e.g., EEG channels) and edges (e.g., functional or spatial connectivity). The core idea is to learn node embeddings by aggregating information from their neighbors, using message-passing or spectral filtering techniques [10, 51].

EEG as a Graph. EEG recordings from multiple electrodes naturally lend themselves to graph modeling. Each channel is treated as a node, and edges are defined based on spatial distance, correlation coefficients, coherence, or learned similarity metrics. Static graphs assume fixed adjacency matrices (e.g., from 10-20 layout, see Figure 2.18), while dynamic graphs adapt topology per input or training epoch.

Common GNN Variants in EEG. Several GNN architectures have been adapted for EEG processing:

- **Graph Convolutional Networks (GCNs)**: Apply spectral filters over

graphs using Laplacian-based convolutions [10] (see Figure 2.10).

- **Graph Attention Networks (GATs)**: Introduce attention mechanisms over neighbors to weigh their contributions [52].
- **Spatial-temporal GNNs**: Integrate temporal convolution with graph convolutions (e.g., ST-GCN, LGGNet) to model both dimensions jointly [24, 35].

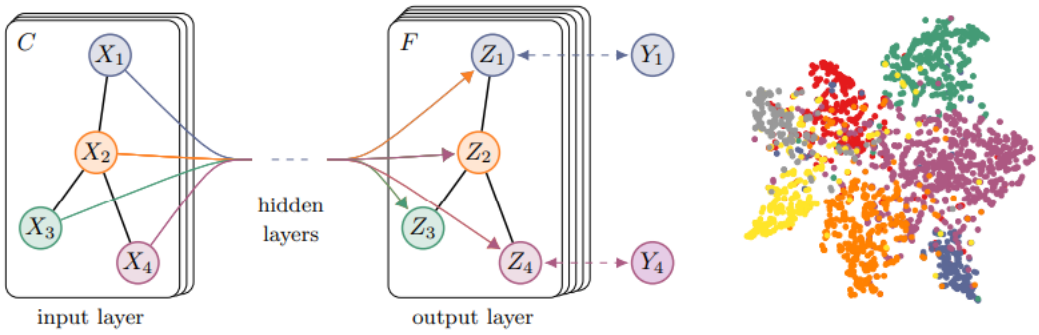


Figure 2.10. Schematic representation of a Graph Convolutional Network (GCN) adapted from [10]. The model aggregates information from neighboring nodes in the graph to learn node-level representations, which are then passed through hidden layers to predict class labels.

Applications in Emotion Recognition. GNNs are particularly suited for emotion recognition because they can encode inter-electrode interactions that vary across emotional states. Models such as LGGNet [35] (see Figure 2.19), AT-DGNN [53], and MT-LGSGCN [54] combine local-global graph modeling with CNN or transformer layers for superior spatial-temporal encoding.

In some designs, the graph is learned adaptively from data using neural attention or similarity-based edge construction, enabling the system to adjust to inter-subject variability and session noise. These techniques show improved generalization and robustness, especially in cross-subject emotion recognition settings.

GNNs also serve as building blocks for hybrid architectures, where they are combined with CNNs for low-level feature extraction and transformers or LLMs for higher-order reasoning, forming modular, end-to-end trainable pipelines.

2.2 Problem Formulation

2.2.1 Emotion Representation

Understanding how emotions are represented is a crucial aspect of affective computing and EEG-based emotion recognition. Two primary frameworks are widely used for modeling emotions: discrete and dimensional.

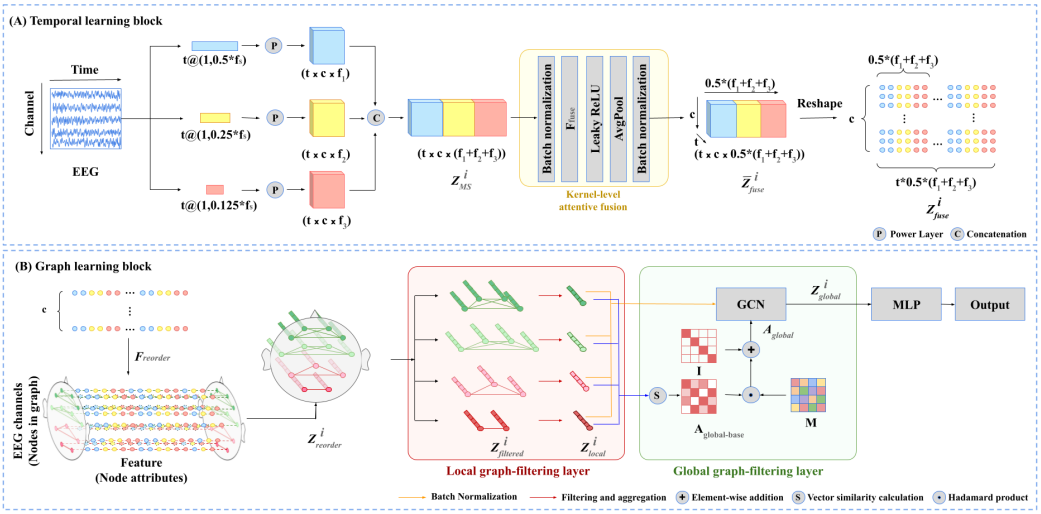


Figure 2.11. Overview of the LGGNet architecture [11]. The model includes a temporal learning block for multi-scale frequency feature extraction, and a graph learning block that combines local and global graph filtering over EEG channels to enhance spatial reasoning.

Discrete Representation. In this framework, emotions are categorized into basic classes such as happiness, sadness, anger, fear, disgust, and surprise. This approach is inspired by Ekman's theory of basic emotions (see Figure 2.12). While straightforward to implement, discrete emotion labels can be overly simplistic and may not capture the complexity or ambiguity of affective states, especially when physiological signals like EEG are involved.

Dimensional Representation. Dimensional models conceptualize emotions as points in a continuous space. The most prominent example is Russell's Circumplex Model (see Figure 2.13), which defines emotions along the axes of Valence (pleasant–unpleasant) and Arousal (calm–excited) [25]. This representation is particularly suitable for EEG-based studies, as it aligns well with regression-based learning frameworks and allows for more fine-grained emotion classification.

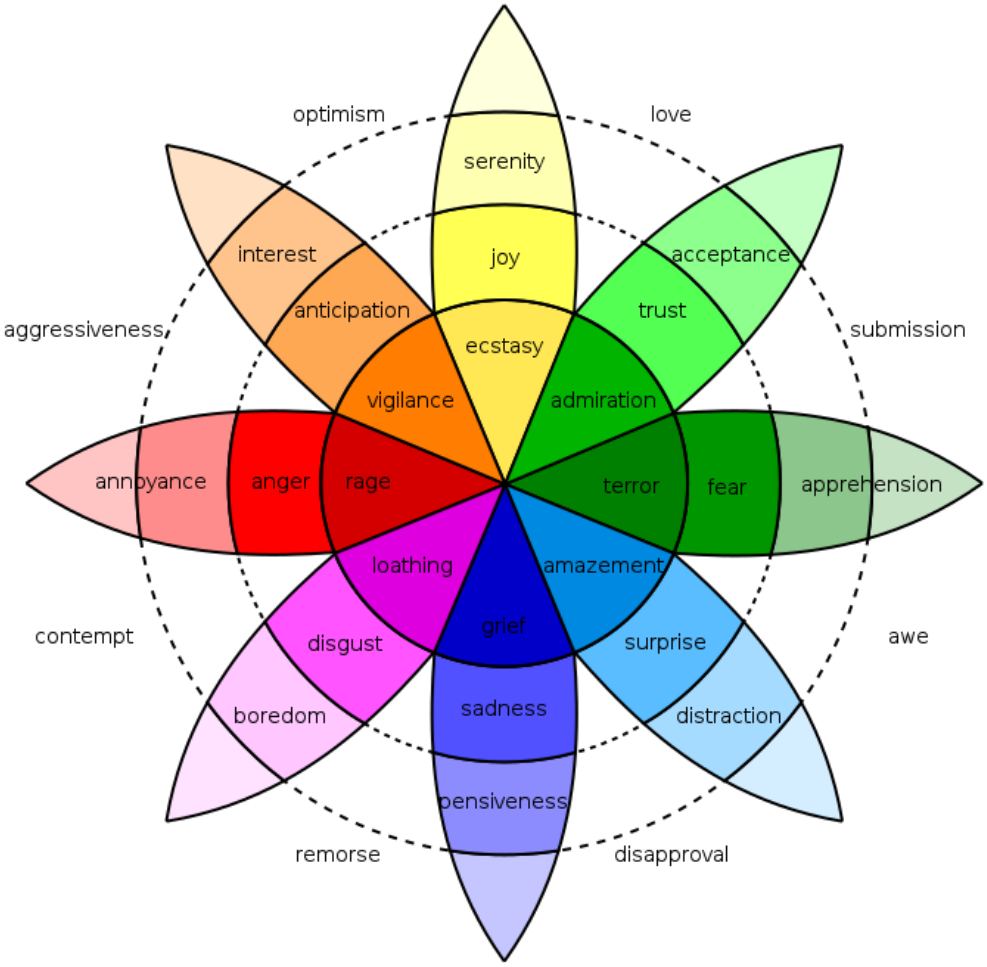


Figure 2.12. Visual representation of discrete emotions arranged in a wheel-like structure. The diagram shows primary emotions at the center (e.g., joy, fear, anger) and their nuanced variations as concentric layers, highlighting how complex emotions derive from basic categories [12].

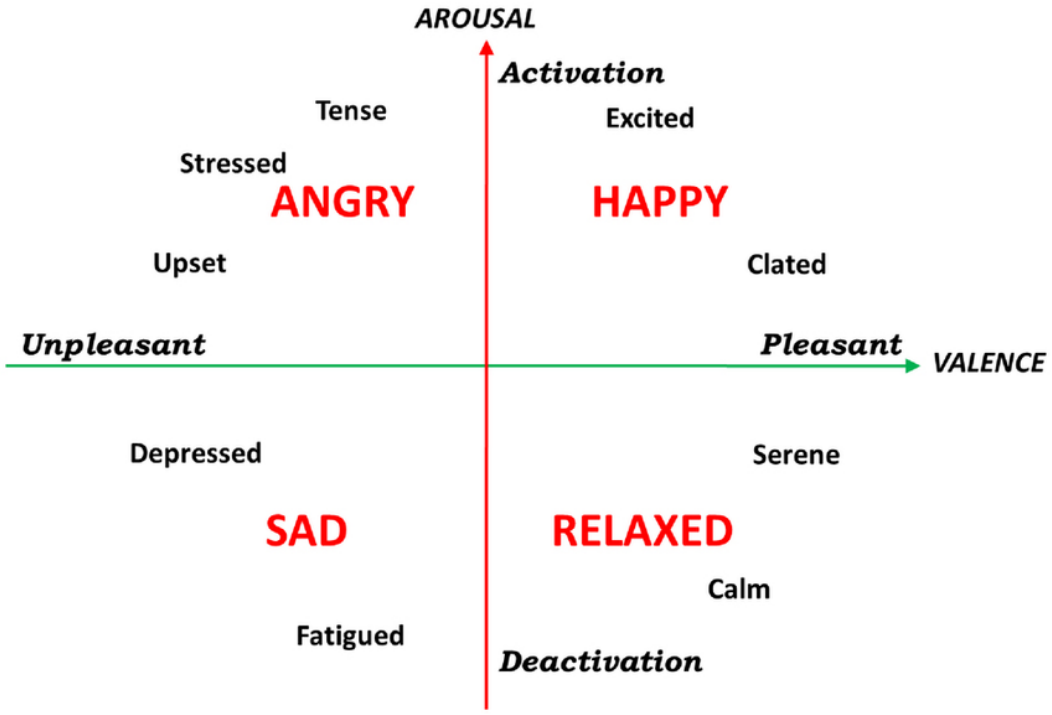


Figure 2.13. Russell's Circumplex Model of Emotion, depicting the relationship between emotional valence (pleasant-unpleasant) and arousal (activation-deactivation).

Mapping EEG to Emotion. Mapping EEG signals to emotional states requires constructing a supervised learning pipeline. This typically involves:

- **Preprocessing:** Filtering, artifact removal, and normalization.
- **Segmentation:** Dividing the continuous EEG stream into epochs aligned with stimulus presentation or emotional annotation (e.g., using sliding windows).
- **Label Alignment:** Associating each EEG segment with corresponding emotional scores. In dimensional settings, these are often obtained from self-assessment questionnaires like the Self-Assessment Manikin (SAM) [14].
- **Modeling:** Training classifiers or regressors to predict emotional states.

This pipeline is used in widely known datasets such as DEAP [6], SEED [19], and DREAMER [55], where each EEG trial is labeled with valence and arousal scores collected from participants. These datasets support both classification (e.g., high vs. low arousal) and regression approaches (e.g., predicting a continuous score).

In multimodal settings, EEG-derived features are often fused with other physiological or behavioral signals (e.g., GSR, facial expressions, speech) to obtain a richer and more robust emotional representation. The dimensional approach provides flexibility for this integration, allowing models to learn cross-modal relationships in a continuous space [56].

2.2.2 Challenges in EEG-based Emotion Recognition

EEG-based emotion recognition is a promising yet complex area of affective computing. Despite its non-invasiveness and high temporal resolution, EEG presents several challenges that must be addressed to achieve reliable emotion classification and regression. These challenges span across signal properties, annotation procedures, and model generalizability.

- **Signal Variability:** EEG signals are inherently noisy, non-stationary, and subject-dependent. Differences in scalp impedance, brain morphology, and electrode positioning introduce variability across subjects and sessions [5]. Moreover, the same emotional stimulus may elicit different neural patterns in different individuals, making inter-subject generalization difficult [57].
- **Low Signal-to-Noise Ratio (SNR):** Emotion-related EEG activity is typically subtle and often obscured by unrelated brain processes and external noise. Artifacts due to eye blinks, muscle movement, and environmental interference further reduce SNR. Advanced preprocessing (e.g., ICA, artifact rejection) and robust feature extraction methods are necessary to mitigate these issues [30, 58].
- **Label Ambiguity:** Emotional labels, especially in dimensional models, are derived from subjective self-assessments. Variability in emotional perception and reporting across participants can introduce inconsistencies in the dataset [14]. Moreover, emotions can rapidly fluctuate, challenging the temporal alignment between EEG segments and affective labels.
- **Data Scarcity:** Collecting high-quality, annotated EEG data for emotion recognition is expensive and time-consuming. Many public datasets (e.g., DEAP, SEED) are limited in terms of subject diversity, recording length, and ecological validity. The scarcity of large-scale, diverse datasets hinders the training of deep learning models and affects generalization [6, 55].
- **Cross-Session and Cross-Device Generalization:** EEG patterns vary not only between subjects but also across different recording sessions and hardware setups. This complicates transfer learning and real-world deployment, where models must adapt to new users or devices with minimal calibration [59].

Addressing these challenges requires a combination of methodological improvements, including data augmentation, domain adaptation, transfer learning, and the integration of multimodal data sources.

2.2.3 Multimodal Approaches

To overcome the limitations of unimodal EEG-based emotion recognition, recent research has increasingly focused on multimodal approaches. These systems aim to fuse complementary signals, such as EEG, facial expressions, speech, galvanic skin response (GSR), and heart rate variability (HRV), to enhance robustness and accuracy [56, 60]. Since emotional states manifest across multiple physiological and

behavioral channels, integrating these modalities provides a more comprehensive representation of affective responses.

Multimodal Fusion Strategies. Fusion can occur at different stages of the pipeline:

- **Early Fusion:** Concatenating raw or low-level features from different modalities before feeding them into a classifier. This requires temporal alignment and normalization but allows for joint feature learning [61].
- **Late Fusion:** Each modality is processed independently, and the outputs (e.g., class probabilities) are combined through voting, averaging, or meta-classifiers. This method is modular and resilient to missing modalities.
- **Hybrid Fusion:** Combines early and late fusion via hierarchical architectures, attention mechanisms, or graph-based integration [62].

EEG in Multimodal Systems. EEG serves as a rich but complex signal, typically capturing internal, cognitive-affective states. It is often fused with more observable modalities like facial expression and other biosignals, which are easier to interpret but potentially less reliable in passive scenarios (see Figure 2.14). For example, in the DEAP dataset, EEG is collected alongside peripheral physiological signals like EOG, GSR, and temperature [6].

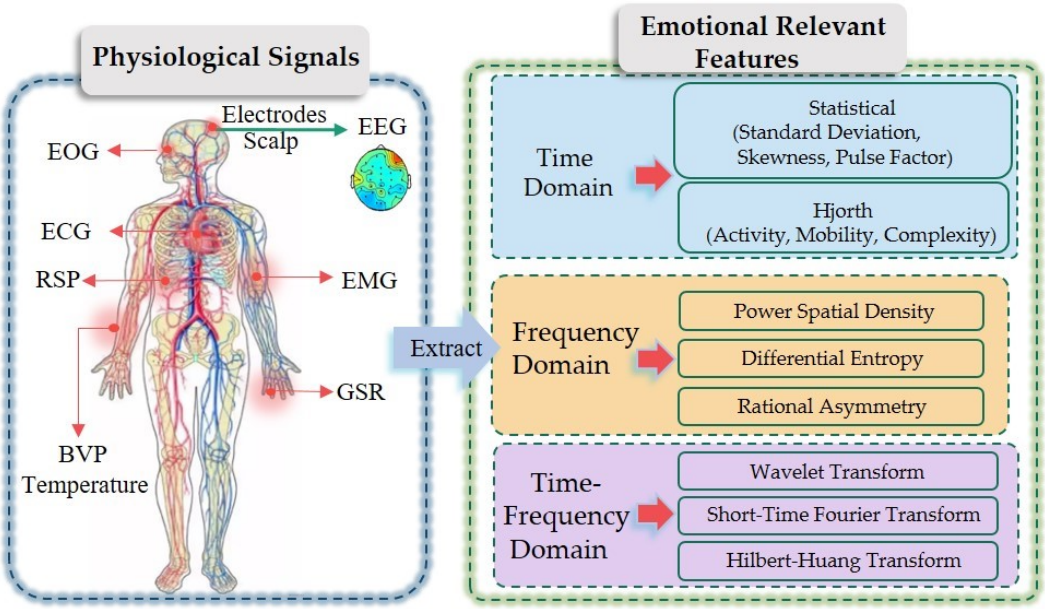


Figure 2.14. Multimodal emotion recognition architecture combining EEG with behavioral and physiological signals. Different fusion strategies (data-level, feature-level, and decision-level) are used to integrate signals before final prediction. Adapted from `https://encyclopedia.pub/entry/2545`.

Advanced methods use deep learning to align feature spaces from different modalities. Techniques include:

- **Cross-modal attention:** Learns relevance weights across modalities.
- **Contrastive learning:** Maximizes agreement between paired modalities (e.g., EEG-audio) and separates mismatched pairs [63].
- **Knowledge distillation:** Teacher networks trained on rich modalities guide student models trained on one modality alone [64].

Recent work integrates large pretrained models (e.g., LLMs or ViTs) across modalities, using them to extract embeddings or guide learning in low-resource domains. These systems show promise in real-world applications such as mental health monitoring, personalized feedback in learning environments (see Figure 2.15), and immersive multimedia experiences [42, 34].

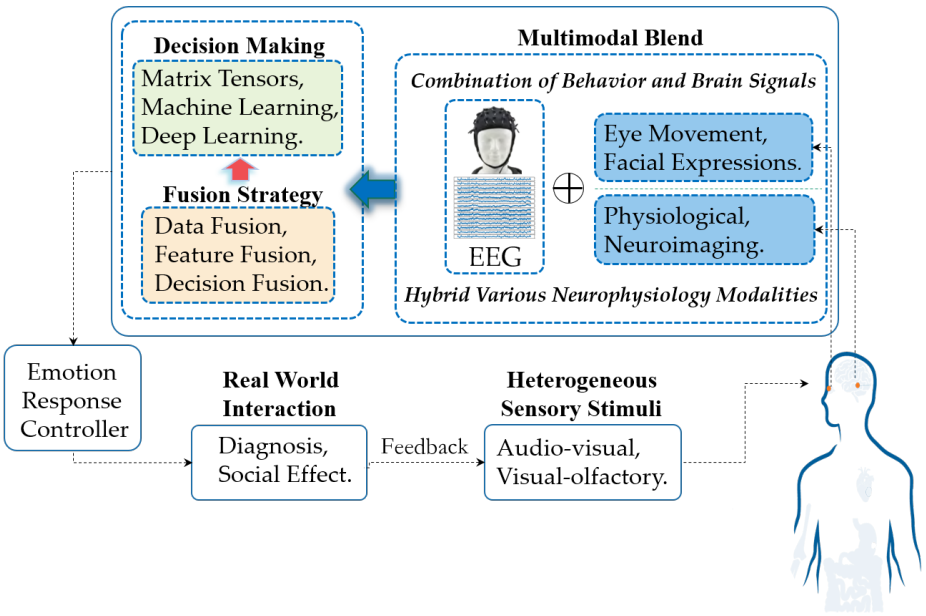


Figure 2.15. Overview of physiological signal types and their corresponding emotion-relevant features. Features are typically extracted in time, frequency, and time-frequency domains using various statistical, spectral, and transformation-based methods. Adapted from `https://encyclopedia.pub/entry/2545`.

Despite the promise of multimodal systems, challenges remain. These include modality synchronization, varying data quality, sensor heterogeneity, and the complexity of designing effective fusion mechanisms. Nevertheless, multimodal emotion recognition is a key direction for making affective computing systems more accurate and human-aligned.

2.3 State of the Art

2.3.1 Machine Learning Approaches

Early EEG-based emotion recognition systems relied heavily on classical machine learning (ML) methods and handcrafted features. Algorithms such as Support Vector

Machines (SVM), k-Nearest Neighbors (kNN), Decision Trees, and Random Forests demonstrated the feasibility of decoding affective states from EEG data [19, 18].

These pipelines typically involved pre-processing steps like band-pass filtering and artifact removal, followed by feature extraction using statistical or frequency-based methods. For instance, Hjorth parameters (activity, mobility, complexity), fractal dimensions, and power spectral density in alpha, beta, and gamma bands were commonly employed [36, 32]. These features were then passed to classifiers—SVMs being particularly effective in high-dimensional, low-sample regimes.

Despite their computational efficiency and interpretability, these models lacked the capacity to learn hierarchical representations and were limited by their reliance on expert-designed features. Furthermore, they often struggled with generalization, especially in subject-independent settings, and could not capture the nonlinear spatiotemporal patterns inherent in EEG.

Nonetheless, their low resource requirements make them attractive for embedded and real-time systems [34], and they continue to serve as strong baselines.

2.3.2 Deep Neural Networks for EEG

Deep learning models have transformed EEG-based emotion recognition by enabling end-to-end architectures that learn directly from raw signals. These systems automatically extract spatial, temporal, and spectral representations through layered structures [58] (see Figure 2.16).

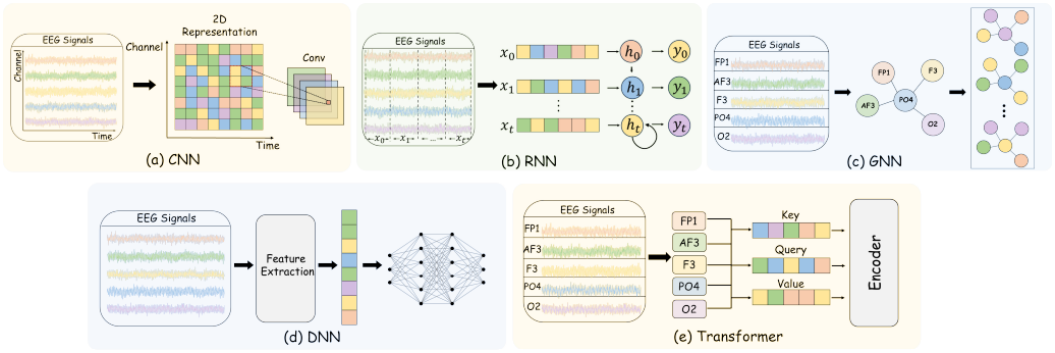


Figure 2.16. Overview of typical deep neural network architectures for EEG-based emotion recognition: (a) CNN, (b) RNN, (c) GNN, (d) DNN, and (e) Transformer-based models.

Convolutional Neural Networks (CNNs). CNNs are widely used to capture spatial relationships between EEG channels. Some models use 1D convolutions over temporal sequences, while others reshape EEG channels into 2D grids and apply 2D filters. Early architectures like DeepConvNet and ShallowConvNet demonstrated the effectiveness of CNNs in extracting frequency-specific filters [65]. EEGNet further optimized CNNs using depthwise separable convolutions for better parameter efficiency [66], and TSception employed multi-scale convolutions to enhance temporal feature extraction [67].

Recurrent Neural Networks (RNNs). RNNs, including LSTM and GRU variants, are designed to model sequential dependencies in EEG signals. Their gating mechanisms allow them to capture temporal dynamics over extended windows, which is beneficial for decoding emotions that evolve gradually [20]. Figure 2.17 illustrates the architectural differences between RNN, LSTM, and GRU units.

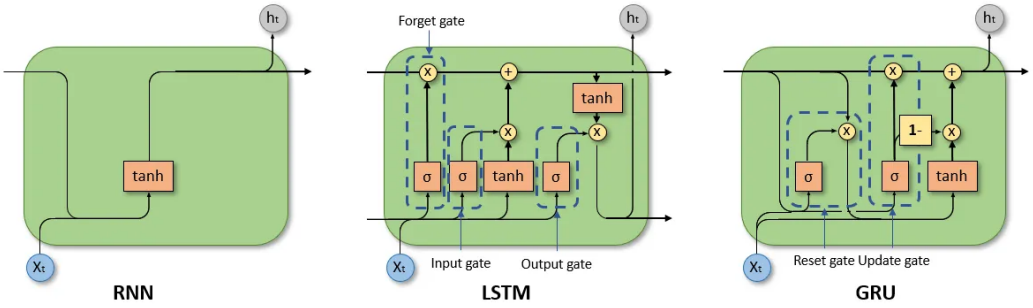
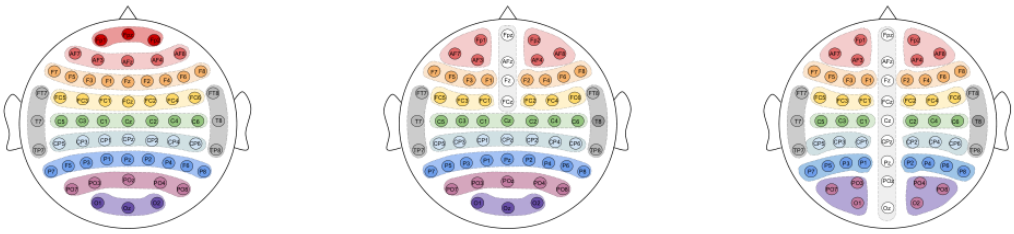


Figure 2.17. Architectural comparison between a vanilla Recurrent Neural Network (RNN), a Long Short-Term Memory unit (LSTM), and a Gated Recurrent Unit (GRU). LSTM and GRU incorporate gating mechanisms that mitigate the vanishing gradient problem and better capture long-term dependencies.

Graph Neural Networks (GNNs). CNNs and RNNs treat EEG as Euclidean signals, overlooking the spatial topology among electrodes. GNNs address this by representing EEG channels as graph nodes and inter-channel relationships as edges. GCNs [10] and GATs [52] learn these structures explicitly. Dynamic variants like DGCNNs update the graph structure based on inter-channel similarities computed from the input data [24].

Recent GNN-based models such as LGGNet [11] introduce neurophysiologically inspired local-global graphs (see Figure 2.18), where electrode groupings reflect cortical functional regions. These are further enriched by symmetry-based adjacency structures in frontal and hemispheric domains.



(a) General (b) Frontal (c) Hemisphere

Figure 2.18. Three types of local-global-graph definitions proposed in LGGNet [11]. (a) General: each local graph reflects the activity of a brain functional area. (b) Frontal: symmetric frontal graphs are added based on known asymmetry in frontal regions. (c) Hemisphere: symmetric graphs are defined across the left and right hemispheres. Colored nodes indicate local subgraphs; dotted lines denote functional groupings.

Hybrid and Modular Architectures. Hybrid architectures that integrate temporal and spatial reasoning have gained traction in recent EEG-based emotion recognition research. Notable examples include LGGNet and MT-LGSGCN. LGGNet [11] constructs local-global functional graphs based on neurophysiological priors, while MT-LGSGCN [54] employs multi-teacher knowledge distillation to unify spatial and temporal information through a student network. These approaches reflect a growing emphasis on modularity and interpretability in neural pipelines.

Figure 2.19 illustrates the LGGNet pipeline, which combines multi-scale temporal convolutions with graph-based filtering blocks to enhance spatial learning over EEG channels.

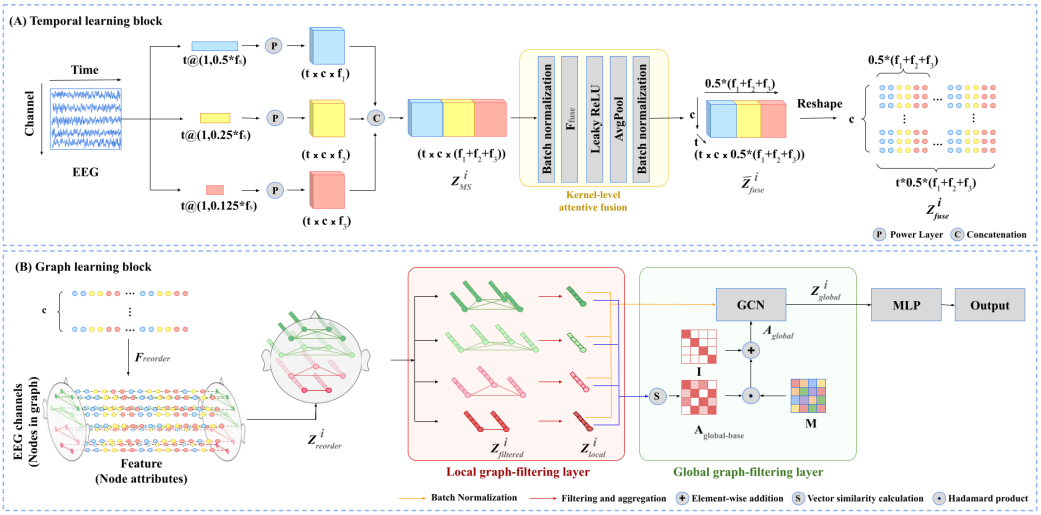


Figure 2.19. Overview of the LGGNet architecture [11]. The model includes a temporal learning block for multi-scale frequency feature extraction, and a graph learning block that combines local and global graph filtering over EEG channels to enhance spatial reasoning.

Transformer Models. Transformers are increasingly used to model long-range temporal dependencies in EEG. ViT-style models (see ViT architecture in Figure 2.20) tokenize EEG signals into patches or frequency windows, learning attention weights across time and channels [41]. These approaches avoid recurrence and allow better scaling with sequence length.

2.3.3 Multimodal Emotion Recognition

Although this thesis focuses solely on EEG signals, it is important to acknowledge the broader research landscape where multimodal emotion recognition is becoming standard. Studies combine EEG with facial expressions, speech, GSR, ECG, and other physiological signals to enhance performance and robustness [56, 60].

Fusion strategies vary across early, late, and intermediate stages. Intermediate methods using attention and shared embeddings are especially effective in real-world contexts where modalities can be noisy or incomplete [61, 62]. While multimodal

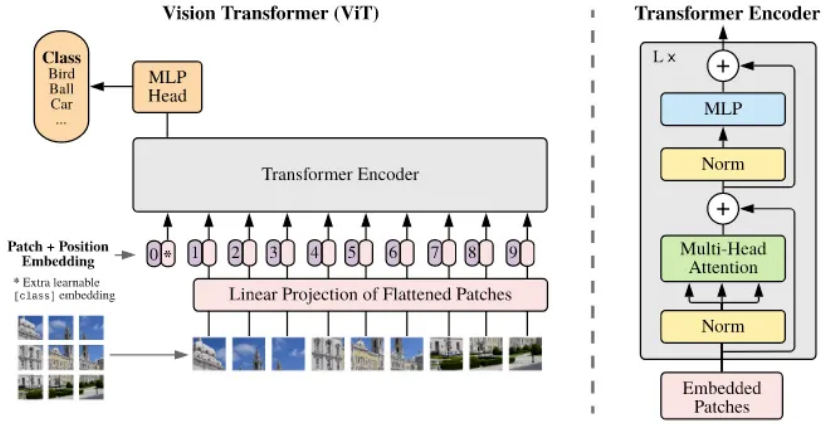


Figure 2.20. Overview of the Vision Transformer (ViT) architecture. An input image is split into fixed-size patches, linearly embedded and combined with positional encodings, then passed through a Transformer encoder. A learnable classification token is prepended to the sequence to enable class prediction via an MLP head. Adapted from [13].

models improve generalization, they often require complex synchronization and infrastructure, which limits their use in portable or low-latency applications.

2.3.4 LLMs in Biosignal Processing

Large Language Models (LLMs), like GPT-2 [38] and LLaMA 3.1 [39] (see comparison in Figure 2.21), have demonstrated significant potential in biosignal modeling. Their transformer-based self-attention layers allow for flexible modeling of long-range dependencies, making them suitable for temporal biosignals like EEG.

EEG-to-Token Conversion and Finetuning. A recent trend involves converting EEG data into token-like representations (e.g., via quantization or patch embeddings) and feeding them to LLMs. These models are then fine-tuned using masked prediction tasks, as shown in EEG-GPT [22], which enables robust learning from sparse supervision and improves generalization to unseen subjects.

Self-Supervised Representation Learning. Pretraining strategies such as masked autoencoding and contrastive learning allow LLMs to reconstruct or align EEG segments. For example, CMAEs (Contrastive Masked AutoEncoders) enforce consistency between EEG embeddings and emotion labels, improving interpretability and domain transfer [42].

Applications Beyond Classification. LLMs have been used for more than classification. They are being explored for emotion reasoning, EEG-to-text generation, and mental state decoding in clinical diagnostics [43].

In the following chapters, I build upon these advancements to propose a novel

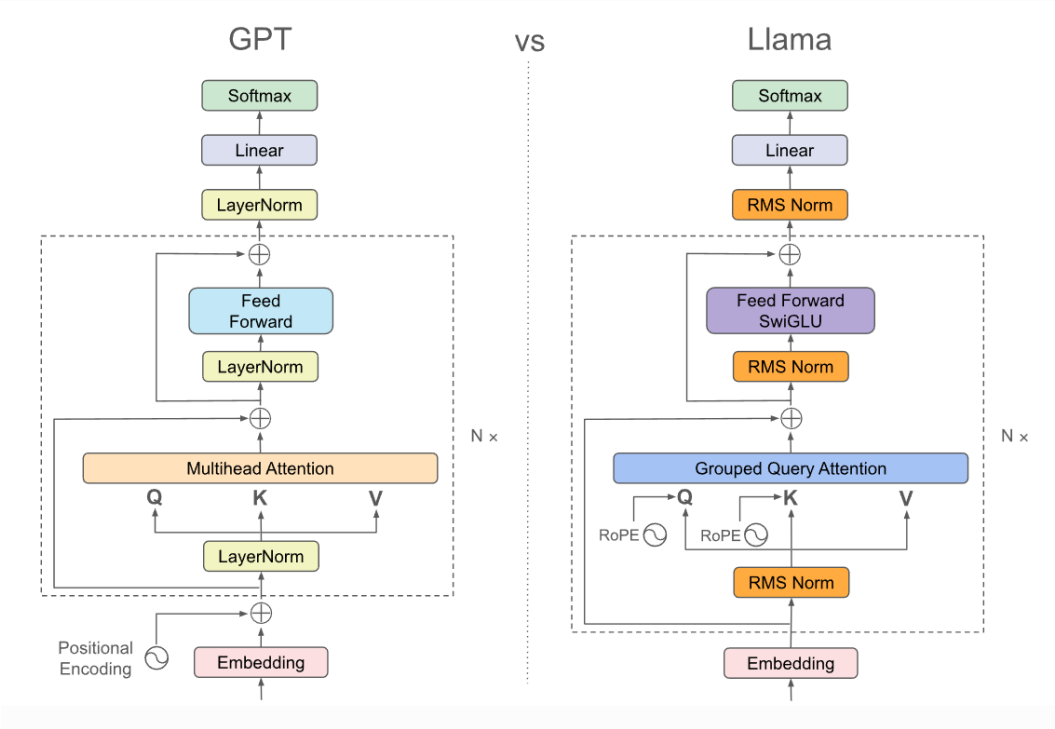


Figure 2.21. Architectural comparison between GPT and LLaMA. LLaMA replaces LayerNorm with RMSNorm, introduces SwiGLU in the feed-forward layers, and adopts Grouped Query Attention along with rotary positional encodings (RoPE), making it more efficient in both training and inference.

architecture for emotion recognition from EEG signals, leveraging large language models fine-tuned on masked temporal embeddings and fusing them with graph-based spatial reasoning. Although prior work has shown the potential of multimodal systems, this study focuses solely on EEG signals to isolate and enhance temporal-spatial modeling capacity using foundation models.

Chapter 3

Proposed Methodology

3.1 Dataset and Preprocessing

3.1.1 DEAP Dataset Overview

The DEAP dataset is a publicly available multimodal corpus designed to support research in emotion recognition from physiological signals, with a particular focus on EEG [6]. It contains recordings from 32 participants who each watched 40 one-minute music video excerpts intended to elicit a wide range of affective responses (see Figure 3.1).

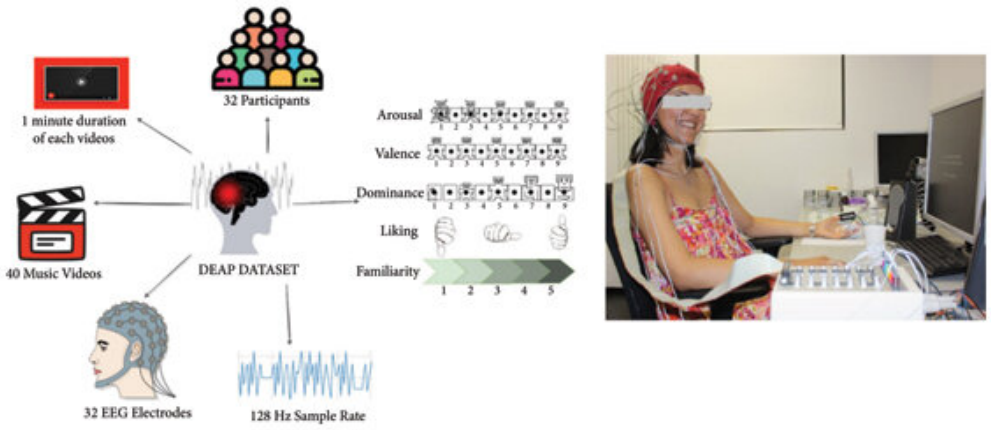


Figure 3.1. Overview of the DEAP dataset experimental setup [6]. The dataset includes EEG recordings from 32 participants while they watched 40 one-minute music videos, rated on arousal, valence, dominance, liking, and familiarity.

These video clips were carefully selected using a hybrid approach combining Last.fm affective tags and manual curation to ensure coverage of different emotional quadrants within the valence-arousal space [25] (see Figure 3.2).

For each trial, participants provided self-assessments along five affective dimensions: valence, arousal, dominance, liking, and familiarity. These ratings were given on continuous 9-point (valence, arousal, dominance, liking) or 5-point (familiarity)

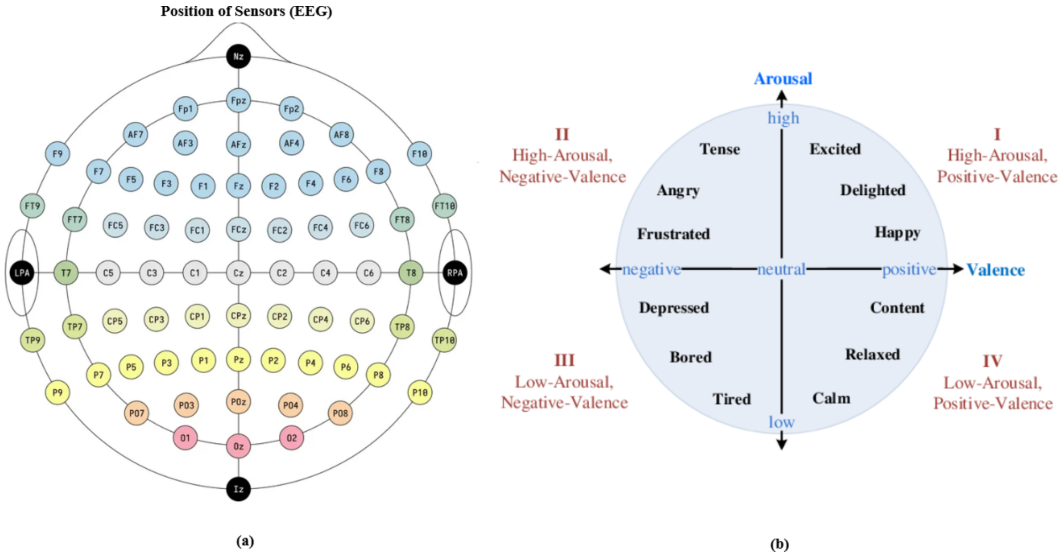


Figure 3.2. (a) 10–20 electrode placement system used in EEG data acquisition. (b) The valence–arousal circumplex model of emotion, dividing the affective space into four quadrants based on the emotional activation and valence levels.

Likert scales, using the Self-Assessment Manikin (SAM) framework [14] (see Figure 3.3), enabling both regression and classification paradigms in downstream tasks.

Physiological signals were acquired at a sampling rate of 512 Hz, including:

- **EEG:** 32-channel scalp recordings following the international 10-20 system [28] (see Figure 3.2).
- **Peripheral signals:** electrooculogram (EOG), electromyogram (EMG), galvanic skin response (GSR), respiration, temperature, and blood volume pressure (BVP).
- **Video recordings:** frontal facial video data were collected for 22 participants, though not used in this work.

In this study, only the EEG modality was utilized, specifically the 32 channels corresponding to scalp electrodes. Prior to model forwarding, the raw EEG signals underwent a standardized preprocessing pipeline. The first step consisted in removing the initial 3-second baseline period from each trial, keeping only the 60 seconds of stimulus-driven EEG. These 60-second segments were then downsampled to 128 Hz to reduce computational complexity and improve model training efficiency.

Following downsampling, the EEG data were reshaped into the form $(32, 40, 32, 60, 128)$, where:

- 32 corresponds to the number of participants,
- 40 is the number of trials per subject,
- 32 is the number of EEG channels,
- 60 is the duration in seconds per trial,

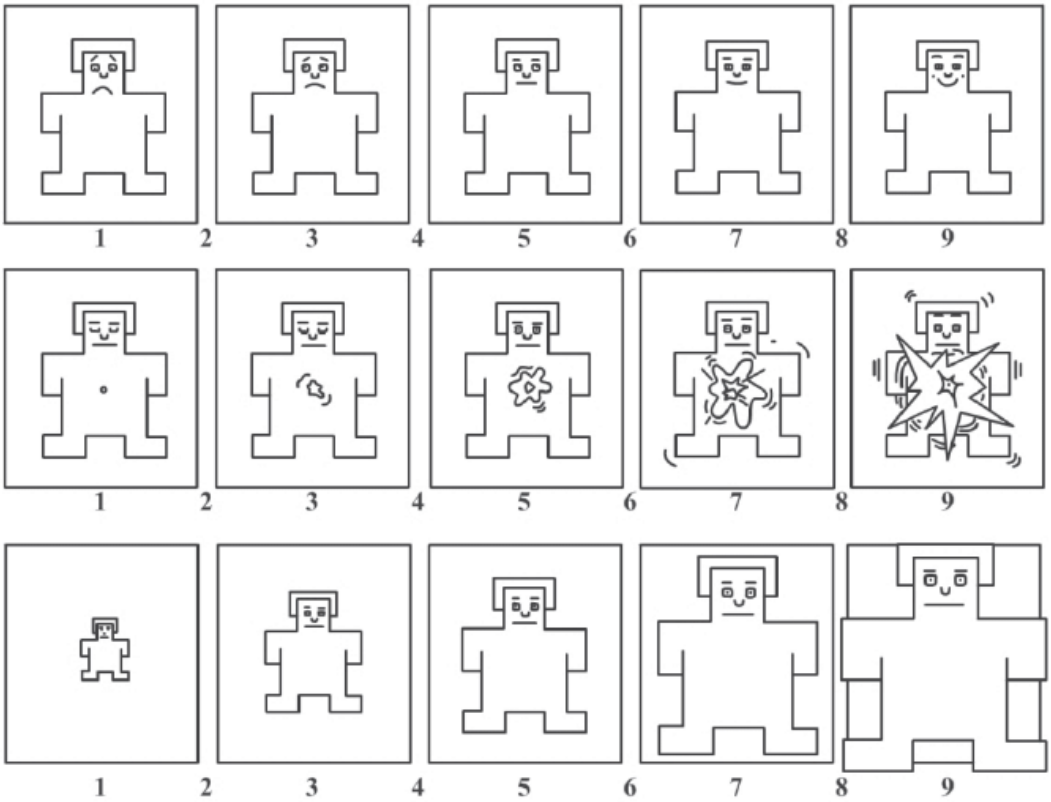


Figure 3.3. Self-Assessment Manikin (SAM) used for reporting emotional state in terms of valence (top row), arousal (middle row), and dominance (bottom row). Participants select one figure from each row to rate their emotional response [14].

- 128 is the number of samples per second (sampling rate).

This structure supports fine-grained temporal analysis and facilitates various segmentation and chunking strategies in downstream processing. The segment-level annotation structure also makes DEAP particularly suitable for training neural architectures that operate on fixed-length EEG segments.

Furthermore, the continuous annotations in the valence-arousal space are commonly binarized around a threshold of 5.0 for classification tasks. In this work, both valence and arousal labels were discretized into two classes, high (rating ≥ 5) and low (rating < 5), following standard practice in the literature.

3.1.2 Signal Segmentation and Chunking

To effectively capture the temporal dynamics of emotion-related EEG activity, the continuous signals were first divided into overlapping segments (see Figure 3.4).

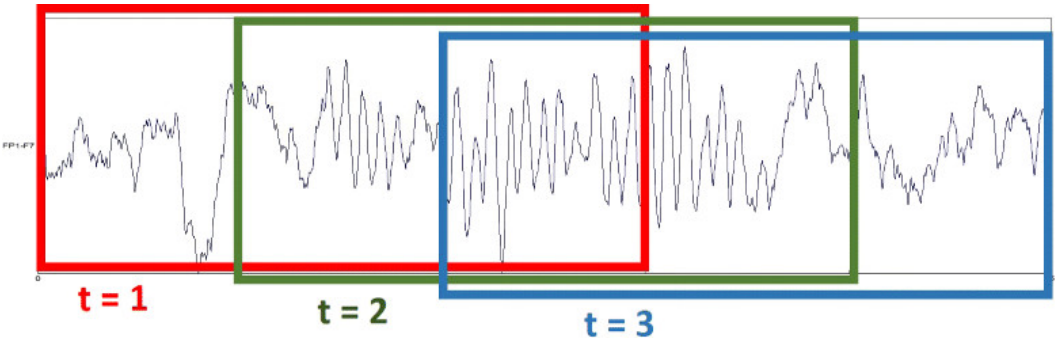


Figure 3.4. Example of EEG signal segmentation, where the continuous signal is divided into smaller, overlapping segments.

The segments were then passed to both the temporal and spatial branches of the network. While both branches process the segment data, an additional operation was applied to the temporal path, where the segments were further split into fixed-length chunks. This step was necessary to prepare the segments for input into the EEGTransformer, with each chunk acting as a masked input for the Large Language Model (LLM) used as a reconstruction module.

Each segment, after being preprocessed and downsampled to 128 Hz, was divided into smaller *chunks* of 64 samples (e.g., 0.5 seconds with 1s segments) using a stride of 32 samples (e.g., 0.25 seconds with 1s segments). These chunks are then used as the basic units of input for the Transformer-based temporal encoder, which allows the model to capture fine-grained temporal dependencies while maintaining continuity between overlapping windows.

Formally, given an EEG segment $S \in \mathbb{R}^{C \times T}$, where $C = 32$ is the number of EEG channels and $T = 128$ is the number of time samples per segment, a 1D sliding window of length $L = 64$ and stride $s = 32$ is applied along the temporal dimension to extract $N = \lfloor (T - L)/s \rfloor + 1 = 3$ chunks:

$$\text{Chunks} = \{S[:, i : i + L] \mid i = 0, 32, 64\}.$$

Each resulting chunk has the shape 32×64 and is processed independently by the EEGTransformer.

This hierarchical segmentation, trial → segment → chunk, enables the model to operate on local patterns while preserving global temporal structure. It also facilitates data augmentation through overlapping windows, increasing the effective number of training samples. This approach is particularly useful for training large-capacity models on limited EEG datasets (like DEAP).

Labels associated with each original 60-second trial are inherited by all derived segments and chunks. For each segment extracted from a trial, the corresponding valence/arousal label is replicated. This strategy assumes that emotional state remains relatively stable within the duration of each stimulus video, a common simplification in EEG-based affective computing [18].

3.2 Model Architecture

3.2.1 Temporal Branch: EEG Transformer Encoder

The temporal encoding of EEG signals in this work is performed using a dedicated Transformer-based architecture, referred to as *EEGTransformer*. Transformers have shown remarkable success in modeling sequential data due to their self-attention mechanism, which enables capturing long-range dependencies without the need for recurrence [7]. Their recent adoption in EEG-based tasks has demonstrated superior performance in capturing temporal patterns over traditional RNNs or CNNs [68].

In my architecture, each chunk of shape $[B, L, C]$, where B is the batch size, L the chunk length (64 samples), and C the number of EEG channels (32), is first projected through a linear layer to a higher-dimensional space of d_{model} features per timestep. Positional embeddings are then added to maintain temporal ordering, a crucial component when dealing with self-attention models that are inherently permutation-invariant.

A Transformer encoder layer is then applied across the time dimension (see Figure 3.5). The EEGTransformer consists of a single ‘nn.TransformerEncoderLayer‘ (from PyTorch), configured with multi-head self-attention ($n_{\text{head}} = 8$), layer normalization, dropout, and GELU activation (see Figure 3.6). This structure allows the model to learn temporal dependencies and contextual representations for each EEG chunk.

The encoded chunk representations are subsequently aggregated via average pooling over the time axis, resulting in a fixed-size vector representation per chunk. A final linear layer then maps the aggregated representation to the desired embedding dimension $d_{\text{emb}} = 768$, matching the embedding size expected by the downstream LLM module.

This temporal encoder is designed to serve two purposes: (i) provide meaningful representations of EEG sequences for classification, and (ii) support the autoregressive learning objective used for reconstructing masked tokens via the LLM branch. It operates independently on each chunk, enabling high parallelization and efficient training across segments.

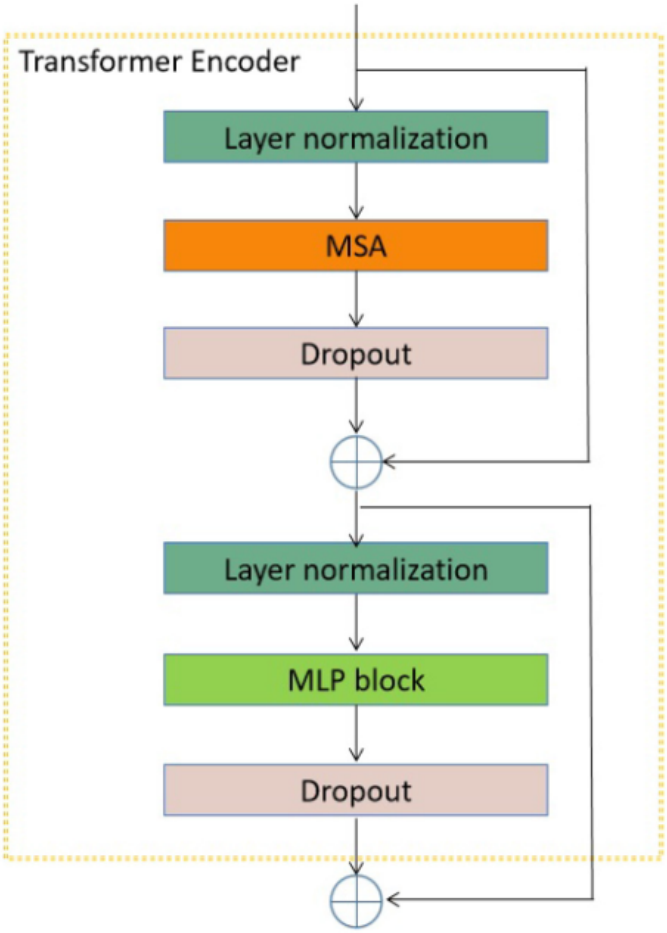


Figure 3.5. Detailed architecture of a Transformer encoder. It includes layer normalization, multi-head self-attention (MSA), dropout, and MLP block (see Figure 3.6) to process EEG data.

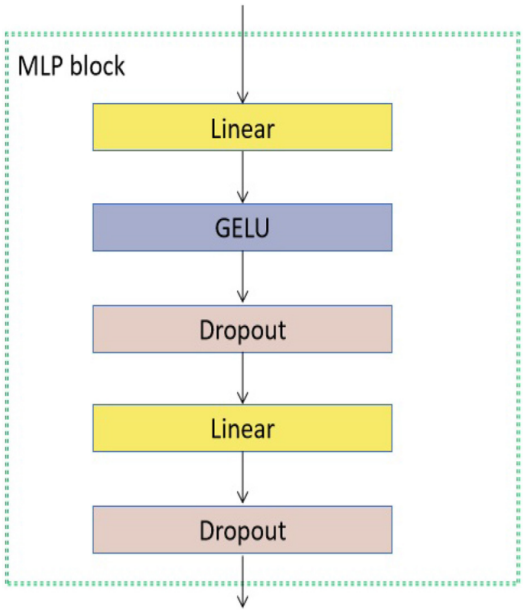


Figure 3.6. MLP block architecture, consisting of two linear layers interspersed with GELU activation and dropout regularization.

3.2.2 Spatial Branch: Dynamic Graph CNN Encoder

While temporal modeling is crucial for capturing dynamic patterns in EEG signals, the spatial configuration of electrodes also carries important information about brain activity distribution. To capture this spatial dependency, a *Dynamic Graph Convolutional Neural Network (DGCNN)* is adopted as the spatial encoder branch (see Figure 3.7).

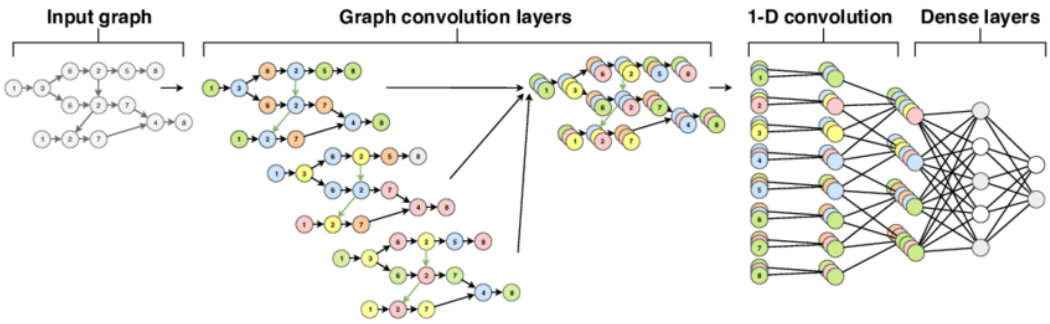


Figure 3.7. The DGCNN architecture used in the spatial branch of the model. This architecture encodes dynamic inter-electrode dependencies, allowing for effective spatial reasoning on EEG signals. The model applies graph convolution layers to capture topological relationships between EEG channels.

The DGCNN models the EEG electrodes as nodes in a graph, where the edges represent pairwise functional or anatomical relationships. Unlike fixed-graph approaches that rely on pre-defined adjacency matrices, the DGCNN dynamically learns the adjacency matrix $\mathbf{A} \in \mathbb{R}^{C \times C}$ during training, where $C = 32$ is the number

of EEG channels. This allows the model to flexibly adapt to subject-specific and task-specific spatial dependencies, a feature particularly important for emotion recognition tasks [35, 24].

The network performs spectral graph convolutions using Chebyshev polynomials of order k to efficiently approximate the graph Laplacian, as introduced in [69]. Each graph convolutional layer updates the node features based on localized neighborhoods in the learned graph, and is followed by a bias-ReLU nonlinearity. The output from multiple stacked graph layers is then flattened and passed through a fully connected layer to produce a d_{emb}-dimensional representation, matching the embedding size of the temporal branch.

Formally, given an EEG segment $X \in \mathbb{R}^{B\times C\times T}$ (batch size B, channels $C = 32$, time steps $T = 128$), the DGCNN treats each electrode as a graph node with a feature vector of length T. The spatial representation is thus constructed by operating over the electrode dimension, aggregating temporal features in a spatially-aware manner.

The use of a dynamic adjacency matrix enables the encoder to learn implicit spatial attention patterns, which can reflect hemispheric asymmetries, inter-region synchrony, or localized activations.

3.2.3 Reconstruction Module (LLM with Finetuning)

A distinctive component of the proposed methodology is the integration of a generative reconstruction module based on large-scale language models (LLMs), specifically GPT-2 [38], LLaMA 3.1 8B AWQ-INT4, and LLaMA 3.2 1B [39]. These models are adapted to the EEG modality through a sequence-level embedding reconstruction task. The module is designed to model the latent dynamics of EEG sequences in a self-supervised manner, leveraging the autoregressive capabilities of the LLMs to predict masked intermediate representations. This enables the model to enforce consistency, continuity, and plausibility across temporally adjacent segments, improving the inductive bias of the overall system.

Each EEG segment is first encoded by the temporal branch into a sequence of chunk-level embeddings $\{e_1, e_2, ..., e_N\}$, where $e_i \in \mathbb{R}^{d_{\text{emb}}}$. The reconstruction task is framed as a causal prediction problem over these embeddings. Specifically, for each position i, a learnable token replaces e_i, and the LLM receives as input the truncated prefix $\{e_1, ..., e_{i-1}, [\text{learnable}]\}$. The model is then tasked with predicting $\hat{e}_i$, evaluated against the original e_i using a mean squared error (MSE) loss:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N-1}\sum_{i=2}^{N} \|\hat{e}_i - e_i\|^2.$$

This formulation is conceptually similar to masked language modeling or next-token prediction in NLP, but operates directly in a continuous latent space derived from EEG. Unlike discrete tokens, the embeddings capture rich temporal features, allowing the model to extrapolate local dynamics from preceding contexts.

To make the computation of this paradigm efficient, the LLaMA models are loaded with 4-bit quantization (NF4) and adapted via `Low-Rank Adaptation (LoRA)` [26, 27], which targets a subset of key transformer submodules (e.g., $q/k/v/o$ projections, feedforward layers). LoRA introduces trainable low-rank matrices $\mathbf{A}, \mathbf{B}$ such that the

adapted weight becomes $\mathbf{W} + \mathbf{BA}$, with $\mathbf{W}$ kept frozen. This approach drastically reduces the number of parameters that need to be updated, while maintaining performance.

The reconstruction loss is modulated through a linear warm-up schedule, where $\mathcal{L}_{\text{rec}}$ is suppressed during the early epochs and progressively incorporated into the joint training objective. Its contribution is controlled by a hyperparameter λ_{rec}, and is combined with the classification losses from both the temporal and spatial branches.

This reconstruction module draws inspiration from recent work integrating LLMs into EEG-based pipelines [22, 64]. It extends the traditional classification task by introducing a generative auxiliary task, allowing the system to benefit from the LLM's pretraining on text while remaining grounded in the continuous nature of neurophysiological signals.

3.2.4 Fusion and Classification Module

The outputs from the temporal and spatial encoding branches are integrated through a dual-path classification framework (see Figure 3.8), allowing for independent processing of each modality and flexible fusion at the decision level. Let $\mathbf{z}_{\text{temp}} \in \mathbb{R}^{d_{\text{emb}}}$ represent the temporal representation obtained from the averaged chunk embeddings generated by the EEGTransformer, and let $\mathbf{z}_{\text{spat}} \in \mathbb{R}^{d_{\text{emb}}}$ denote the spatial embedding derived from the DGCNN-based encoder. These representations are then processed independently by two dedicated classification heads, FC_T and FC_S, each composed of a multi-layer perceptron (MLP) with ReLU activations and dropout for regularization:

$$\hat{\mathbf{y}}_\text{T} = \text{FC}_\text{T}(\mathbf{z}_{\text{temp}}), \quad \hat{\mathbf{y}}_\text{S} = \text{FC}_\text{S}(\mathbf{z}_{\text{spat}}),$$

where $\hat{\mathbf{y}} \in \mathbb{R}^C$ represents the output logits for the C emotion classes (with $C = 2$ for binary valence/arousal classification).

Multiple fusion strategies are supported to combine the predictions from both classifiers:

1. **Soft-voting**: This method averages the class probabilities from both branches:

$$\hat{\mathbf{y}} = \frac{1}{2}\left(\text{softmax}(\hat{\mathbf{y}}_\text{T}) + \text{softmax}(\hat{\mathbf{y}}_\text{S})\right).$$

2. **Confidence-based selection**: This strategy selects the prediction corresponding to the highest maximum softmax score between the two heads, ensuring that the most confident branch contributes to the final prediction.

From a training perspective, each classification head is supervised individually using a weighted cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = \alpha \cdot \mathcal{L}_{\text{CE}}(\hat{\mathbf{y}}_\text{T}, \mathbf{y}) + (1 - \alpha) \cdot \mathcal{L}_{\text{CE}}(\hat{\mathbf{y}}_\text{S}, \mathbf{y}),$$

where $\alpha \in [0, 1]$ is a balancing coefficient. In practice, α is typically set to 0.5 unless otherwise specified.

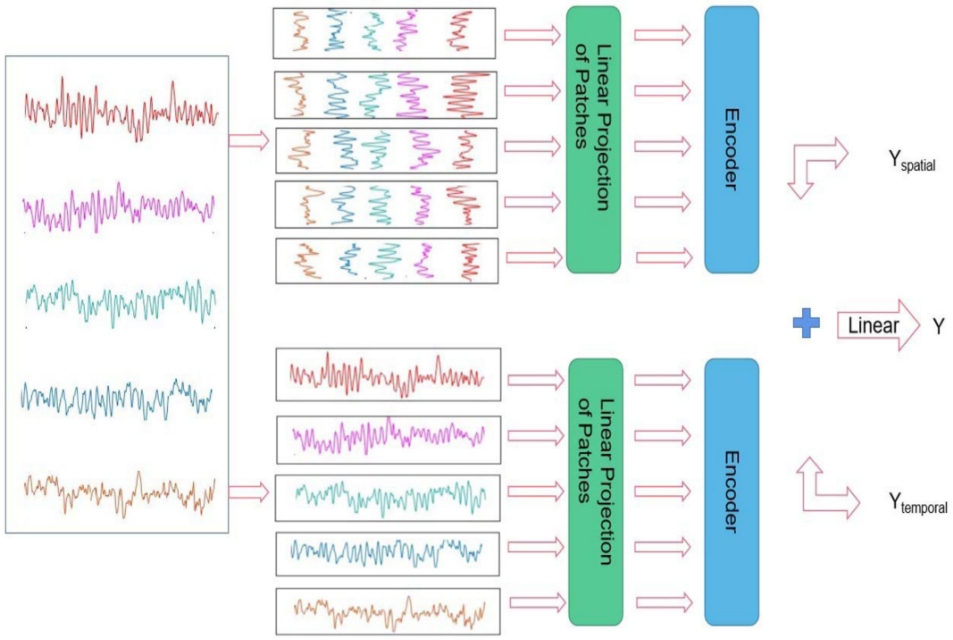


Figure 3.8. An approximate representation of the architecture, where both the temporal and spatial branches are processed independently and then combined through fusion techniques, including average and confidence-based selection.

This architecture decouples the representation learning paths for the temporal and spatial branches while maintaining a shared training objective. The temporal branch captures sequential dependencies using attention mechanisms, while the spatial branch focuses on topological relationships via graph convolutions.

In the following chapter, the results of the experiments conducted to evaluate the performance of the proposed model are presented. These experiments focus on assessing the model's ability to perform emotion recognition from EEG signals in a subject-dependent setting. The results are evaluated based on key metrics such as accuracy and macro-averaged F1 score, computed across different subjects in the DEAP dataset. The chapter also includes a comparison with baseline models, as well as an analysis of how various hyperparameters and model configurations impact performance. Additionally, ablation studies are conducted to understand the contribution of different components, such as the reconstruction loss and the use of LoRA for fine-tuning large language models. The findings highlight the strengths and limitations of the proposed approach, offering insights for future improvements.

Chapter 4

Experimental Results

4.1 Experimental Setup

4.1.1 Hardware and Computational Constraints

All experiments were conducted using an NVIDIA A100 GPU with 40GB of RAM. This setup enabled the training and inference of large-scale transformer-based architectures, including the 8-billion parameter LLaMA 3.1 in INT4 precision. The high memory capacity was particularly beneficial for batch processing and for loading multiple model components, including LoRA adapters and quantized LLM backbones.However, due to limited access to the A100 GPU, the training runs were constrained by time, which occasionally impacted the number of experiments completed.

4.1.2 Subject-Dependent Approach

To assess the generalization capability of the proposed model, subject-dependent training and testing were performed. For each of the 32 subjects in the DEAP dataset, a model instance was trained on data from the selected subject and tested on that same subject. This subject-dependent approach simulates real-world scenarios where emotion recognition models must adapt to individual differences in EEG data.

4.1.3 Input Segmentation and Preprocessing

Each EEG trial in the DEAP dataset consists of 60 seconds of stimulus-evoked neural activity recorded at 128 Hz over 32 channels. To enable fine-grained temporal modeling, each trial was divided into 1-second non-overlapping segments, resulting in 60 segments per trial and 2400 segments per subject. Each segment was then split into overlapping chunks of 64 samples (0.5 seconds) with a stride of 32 samples (0.25 seconds), yielding 3 chunks per segment. This segmentation scheme allows the temporal transformer to operate on short contextual windows while preserving continuity within the segments. The EEG signals were z-score normalized across time for each segment to reduce inter-subject and inter-trial variance.

Labels for arousal and valence were binarized using a threshold of 5.0, converting both dimensions into binary classes (low/high). These binarized labels were

propagated from the original trial to all derived segments and chunks.

4.1.4 Training Strategy and Hyperparameters

Each experiment was trained for up to 150 epochs using the `AdamW optimizer`, with a base learning rate of 10^{-3} and a weight decay of 10^{-4}. A linear warm-up phase was applied over the first 5% of training steps, followed by a cosine annealing schedule to improve convergence stability.

Batch sizes were set to 32 during training and 64 during evaluation. Dropout was applied throughout the network with a fixed probability of 0.1, including in self-attention layers, MLPs, and fusion modules. The loss function combined cross-entropy terms for the temporal and spatial classifiers, with a reconstruction loss for masked temporal embeddings when the LLM component was active. The reconstruction loss was weighted by a scaling factor λ_{rec} (default = 2.0) to balance its contribution against classification objectives.

For fine-tuning large language models, their learning rate was scaled down by a factor of 0.5 relative to the base components. `Low-Rank Adaptation (LoRA)` was applied to both GPT-2 and LLaMA models, with a rank of 16 and scaling factor of 32. Only specific transformer submodules (e.g., attention projections) were updated, keeping the rest of the LLM frozen to reduce memory usage and mitigate overfitting. `Early stopping` was optionally applied after 15 epochs without improvement in validation accuracy, though in most cases, full training cycles were completed.

4.1.5 Evaluation Metrics

To assess model performance, I adopted two widely used metrics in EEG-based classification: accuracy and macro-averaged F1 score.

Accuracy. Accuracy measures the proportion of correctly classified EEG segments over the total number of predictions. It is reported per subject and as an average across all 32 subjects.

Macro F1 Score. The macro-averaged F1 score is used as the primary metric to address the frequent class imbalance in valence and arousal labels. It averages the F1 scores of each class independently, ensuring that minority classes contribute equally to the final score:

$$\text{F1}_{\text{macro}} = \frac{1}{2}(\text{F1}_{\text{low}} + \text{F1}_{\text{high}})$$

This metric is particularly relevant in affective computing tasks, where class distributions are often skewed.

Head-wise Metrics. In addition to global metrics, the accuracy and F1 score of the temporal classifier, the spatial classifier, and their fused outputs are reported. The fused outputs are obtained using two strategies: **average**, where the logits from both branches are averaged, and **confidence**, where the prediction with the highest confidence (softmax score) is selected. This allows us to analyze the contribution of each modality and evaluate the impact of different fusion strategies.

4.2 Quantitative Results

4.2.1 Baseline Comparisons

To contextualize the performance of the proposed architecture, it was compared with a diverse range of baseline models, including traditional classifiers, convolutional and recurrent networks, and recent graph- and transformer-based approaches. These baselines span from shallow learning methods like SVM and KNN to structured deep architectures such as TSCeption, LGGNet, and MT-LGSGCN [54], providing a comprehensive overview of performance on the DEAP dataset.

Table 4.2 summarizes key results reported in the literature for EEG-based emotion recognition on the DEAP dataset, comparing performance across multiple architectures for both arousal and valence classification tasks. Each entry includes the model type and its corresponding accuracy and macro-averaged F1 score.

Table 4.1. Comparison of classification performance across methods for both arousal and valence tasks on the DEAP dataset.

Method	Arousal		Valence	
	ACC	**F1**	**ACC**	**F1**
SVM	60.37%	57.33%	55.19%	57.87%
KNN	59.48%	57.49%	53.03%	55.12%
ShallowConvNet	61.19%	61.19%	59.42%	62.26%
DeepConvNet	61.03%	62.58%	59.92%	62.04%
TSCeption	61.57%	63.24%	59.14%	62.33%
DGCNN	60.80%	60.34%	53.97%	56.27%
LGGNet-Fro	61.19%	63.96%	58.95%	63.89%
LGGNet-Hem	61.52%	63.79%	59.18%	64.34%
LGGNet-Gen	61.81%	64.49%	59.14%	64.58%
MT-LGSGCN	63.59%	65.11%	61.69%	65.23%

As shown in Table 4.2, traditional machine learning models such as SVM and KNN achieve modest results, with accuracy and F1 scores typically below 60% on both arousal and valence tasks. Shallow neural networks like ShallowConvNet and DeepConvNet provide slight improvements by capturing local temporal features, while more structured approaches such as TSCeption and DGCNN incorporate deeper temporal and spatial reasoning.

Graph-based architectures like LGGNet variants further enhance performance by explicitly modeling inter-channel relationships, with LGGNet-Gen achieving consistent gains over baseline convolutional methods. Among the evaluated models, MT-LGSGCN delivers the highest classification scores, demonstrating the benefit of combining spatial graph representations with temporal modeling in a unified architecture.

4.2.2 LLM-powered Model Performance

Large language models (LLMs) have only recently been explored in the context of EEG-based emotion recognition. Their ability to model sequential dependencies

and learn from limited supervision makes them particularly well-suited for neural decoding tasks under data-scarce conditions.

To date, few works have attempted to apply LLMs to EEG signals in an end-to-end classification setting. Among these, LM-KD represents a prominent example [64]. However, it relies on multimodal teacher supervision, including physiological signals such as GSR, and performs knowledge distillation rather than direct emotion classification. Moreover, it does not adapt the LLM to the EEG domain via finetuning on the DEAP dataset itself.

In contrast, the approach proposed in this thesis constitutes, to the best of my knowledge, the first instance of fine-tuning a large language model directly on the DEAP dataset for the joint classification of arousal and valence. This is achieved by combining a temporal LLM decoder, optimized via Low-Rank Adaptation (LoRA) and trained with a masked reconstruction loss, with a spatial graph encoder that captures inter-electrode dependencies.

Importantly, the entire training pipeline is conducted solely on EEG data, without resorting to auxiliary modalities such as GSR or visual input, and without the use of synthetic augmentation techniques. This fully self-contained setup demonstrates that LLM-based architectures can be adapted to EEG-based emotion recognition in a resource-efficient manner.

A comparison with prior state-of-the-art models on the DEAP dataset is reported in Table 4.2. The proposed method achieves the best performance on both arousal and valence classification, establishing a new benchmark in the EEG-only setting.

Table 4.2. Comparison of classification performance across methods for both arousal and valence tasks on the DEAP dataset, with respect to our model.

Method	**Arousal**		**Valence**	
	ACC	**F1**	**ACC**	**F1**
SVM	60.37%	57.33%	55.19%	57.87%
KNN	59.48%	57.49%	53.03%	55.12%
ShallowConvNet	61.19%	61.19%	59.42%	62.26%
DeepConvNet	61.03%	62.58%	59.92%	62.04%
TSCeption	61.57%	63.24%	59.14%	62.33%
DGCNN	60.80%	60.34%	53.97%	56.27%
LGGNet-Fro	61.19%	63.96%	58.95%	63.89%
LGGNet-Hem	61.52%	63.79%	59.18%	64.34%
LGGNet-Gen	61.81%	64.49%	59.14%	64.58%
MT-LGSGCN	63.59%	65.11%	61.69%	65.23%
Ours (LLM-Powered)	**66.99%**	**65.78%**	**64.69%**	64.53%

4.3 Ablation Studies

4.3.1 Effect of LLM Fine-tuning

To isolate the contribution of the LLM component, several variants of the model are compared : (i) a version where the language model is kept frozen and used only for inference; (ii) a version where the LLM is fine-tuned via Low-Rank Adaptation (LoRA)

on the masked reconstruction objective; and (iii) a version where the LLM module is removed, but the temporal branch is retained using only the EEGTransformer encoder, without masked reconstruction or autoregressive decoding.

Table 4.3 reports the performance differences across these settings, evaluated exclusively on the arousal classification task. When LoRA fine-tuning is enabled, the LLM significantly improves both accuracy and F1-score, confirming its role not only as a temporal encoder but also as a powerful regularizer through self-supervised dynamics modeling.

Table 4.3. Impact of LLM fine-tuning on arousal classification performance.

Model Variant	ACC	F1
No LLM (EEGTransformer only)	57.48%	53.24%
Frozen LLM (no fine-tuning)	63.94%	61.55%
Fine-tuned LLM (w/ LoRA)	**66.99%**	**65.78%**

These results suggest that simply including a pretrained LLM is not sufficient: without fine-tuning, the model offers limited gains. The self-supervised reconstruction loss $\mathcal{L}_{\text{rec}}$ acts as a crucial training signal that enhances the temporal branch's alignment with EEG dynamics.

Moreover, by updating only a small subset of parameters (e.g., LoRA matrices in the attention blocks), it is maintained computational efficiency while capturing domain-specific patterns, avoiding overfitting on the small EEG dataset. This validates the use of parameter-efficient tuning strategies in neuro-symbolic applications of LLMs.

4.3.2 Impact of Reconstruction Loss

One of the distinctive elements of the proposed architecture is the inclusion of a self-supervised reconstruction loss $\mathcal{L}_{\text{rec}}$, applied to chunk-level embeddings generated by the temporal encoder and predicted autoregressively by the LLM. This loss encourages the model to capture local dynamics and continuity within the EEG stream, even without direct access to labels.

To quantify its impact, I compare models trained with and without the reconstruction objective, keeping all other components (LoRA configuration, LLM weights, training schedule) fixed. Results are shown in Table 4.4.

Table 4.4. Effect of the reconstruction loss $\mathcal{L}_{\text{rec}}$ on LLM-powered model performance in the arousal classification task.

Model Configuration	ACC	F1
No reconstruction	64.02%	63.34%
LLM (fine-tuned), with $\mathcal{L}_{\text{rec}}$	66.99%	65.78%

The addition of $\mathcal{L}_{\text{rec}}$ consistently improves both classification accuracy and macro-averaged F1 score by approximately 1.8 percentage points. This suggests that the reconstruction loss provides a valuable auxiliary signal, helping the LLM to

learn more robust temporal embeddings aligned with the intrinsic structure of EEG activity.

Furthermore, the reconstruction loss acts as an implicit regularizer: by forcing the model to predict latent representations of future segments, it learns to encode continuity and context, rather than simply memorizing class boundaries. This is particularly beneficial in noisy, low-data environments like EEG, where labels alone may be insufficient to drive generalizable learning.

4.3.3 Segment Length and Temporal Resolution

EEG signals are inherently continuous, non-stationary, and multi-scale in nature. Choosing an appropriate segmentation strategy is therefore critical to capturing emotionally relevant neural patterns while avoiding temporal drift or dilution of signal quality. In this study, I evaluated both short (1-second) and long (4-second) segmentation strategies, applied to each 60-second DEAP trial. Each segment was further divided into overlapping 0.5-second chunks with a stride of 0.25 seconds to enable fine-grained temporal modeling. This chunking design allows the transformer-based temporal encoder to process local dynamics while preserving contextual information. The comparison between the two segment lengths reflects a trade-off between resolution, label stability, and computational efficiency.

Rationale for Temporal Segmentation. Segment duration plays a pivotal role in EEG-based emotion recognition, especially in models that integrate temporal encoders or autoregressive decoders. In this work, I explored two primary segment lengths: 1.0 second and 4.0 seconds. Shorter segments, such as 1.0 s, offer finer temporal granularity, allowing the model to respond to rapid neural fluctuations. They also generate a larger number of training examples, which can help regularize the learning process and reduce overfitting. However, because DEAP emotion labels are assigned at the video level, these fine slices may introduce label noise, assuming affective stationarity at sub-second scales.

Longer segments, such as 4.0 s, tend to average out transient fluctuations and may better reflect the participant's emotional state over a stable period. They reduce the number of training samples, but in the experiments they led to slightly better classification accuracy, suggesting a trade-off between temporal precision and semantic coherence.

From a computational perspective, 1-second segments enabled slightly faster training and larger batch sizes, especially in the LLM-driven pipeline, due to their shorter effective sequence lengths. This is particularly relevant in architectures where attention complexity scales quadratically with input length.

Experimental Setup. To assess the impact of temporal resolution, I compared the following configurations:

- **1.0 s segments:** default configuration used in most experiments; each segment split into 3 overlapping 0.5 s chunks (stride 0.25 s);
- **4.0 s segments:** split into 7 overlapping 0.5 s chunks (stride 0.25 s).

Both variants used the same temporal encoder (EEGTransformer), the same LLM backbone (LLaMA 3.1 8B AWQ), and identical training schedules. Table 4.5 reports the classification results.

Table 4.5. Impact of segment duration and chunking on model performance on the arousal classification task.

Segment Configuration	ACC	F1
1.0 s (3 × 0.5 s chunks)	65.09%	64.38%
4.0 s (7 × 0.5 s chunks)	**66.99%**	**65.79%**

Observations. The 4-second segments slightly outperformed the 1-second configuration in terms of both accuracy and F1-score. This suggests that longer segments may better encapsulate the affective signal while filtering out high-frequency noise. However, the 1-second setting still remains competitive, offering benefits in terms of training speed, memory efficiency, and sample count. This makes it suitable for prototyping, ablation studies, or low-resource deployments.

These findings align with earlier EEG literature, where 2–4 s windows have been commonly used for capturing affect-related rhythms and event-related potentials [19, 67]. Our results confirm that segment length remains a key hyperparameter in EEG modeling, especially in LLM-augmented systems.

Implications for LLM-driven EEG Modeling. The choice of segment duration impacts not only input formatting but also the learning dynamics of the autoregressive LLM decoder. Longer sequences increase the LLM's receptive field and may enable better continuity modeling. However, they also increase training time and require more memory due to the $\mathcal{O}(n^2)$ cost of self-attention.

In this trade-off space, the 4-second segments may be preferred when accuracy is the top priority, while 1-second segments offer a leaner setup for broader experimentation. Future work could explore hybrid schemes that adjust segment length based on signal entropy or stimulus dynamics, allowing more adaptive temporal granularity.

4.3.4 Backbone Comparison: GPT vs LLaMA

While the proposed architecture is designed to be backbone-agnostic with respect to the underlying large language model (LLM) decoder, the choice of LLM plays a critical role in shaping the temporal representation power, reconstruction fidelity, and generalization behavior of the model. To evaluate this sensitivity, three LLMs integrated into the pipeline are compared:

- **GPT-2 Small (124M)** – lightweight, fully open-source, and easy to fine-tune;
- **LLaMA 3.2 1B** – a compact, modern LLM with RoPE embeddings and strong pretraining;
- **LLaMA 3.1 8B AWQ INT4** – quantized 4-bit version of a high-capacity model using AWQ for efficiency.

Motivation and Constraints. GPT-2 was adopted both for direct comparison with prior LLM-based EEG architectures, such as LM-KD [64], and as a fallback under constrained GPU access. Although the training was performed on an A100 GPU with 40GB of VRAM, session availability was limited, preventing extended training of large backbones. Quantized models such as LLaMA 3.1 8B AWQ (INT4) made possible to leverage modern architectures while remaining within feasible memory budgets.

LoRA-Based Parameter Efficiency. All LLMs in the pipeline were fine-tuned using Low-Rank Adaptation (LoRA), a parameter-efficient technique that injects trainable low-rank matrices into selected attention and feedforward layers. This drastically reduces the number of parameters updated during training: for instance, only ~1.2M parameters were fine-tuned in GPT-2 Small (124M). This approach minimizes overfitting, lowers memory usage, and makes training feasible even under constrained GPU resources. The frozen weights of the base LLM remain untouched, ensuring that general-domain linguistic knowledge is retained while allowing for domain-specific temporal adaptation. A comparison of parameter efficiency and classification performance is reported in Table 4.6.

Table 4.6. Performance and parameter efficiency of different LLM backbones with LoRA fine-tuning, on the arousal classification task.

LLM Backbone	ACC	F1
GPT-2 Small	66.99%	65.78%
LLaMA 3.2 1B	64.02%	63.59%
LLaMA 3.1 8B AWQ INT4	64.50%	64.28%

Efficiency Trade-offs. Among the evaluated LLM backbones, GPT-2 Small achieved the highest accuracy and F1-score on the arousal classification task, despite being the smallest model by parameter count. Its compact size enabled full-precision training, larger batch sizes, and faster convergence, making it particularly suitable for low-resource scenarios, rapid iteration, and ablation studies. Notably, its performance outpaced both larger LLaMA variants, highlighting that model capacity alone does not guarantee superior results in EEG-based emotion recognition.

The LLaMA 3.2 1B model, although more modern and expressive, delivered lower performance than GPT-2 in this context. However, it supported mixed-precision training without requiring quantization, and remained manageable in terms of GPU memory and training duration. This made it a viable middle-ground option when balancing expressiveness with resource constraints.

By contrast, the LLaMA 3.1 8B AWQ INT4 model, despite its scale, showed only marginal improvements over the 1B variant and underperformed compared to GPT-2. Training was conducted using 4-bit quantization to reduce memory usage, which introduced limitations in gradient fidelity and required smaller batch sizes. Moreover, due to the model's size, multiple training sessions had to be resumed from checkpoints because of GPU session limits on the A100 (40GB), increasing overall complexity and reducing reproducibility.

These results emphasize that in EEG pipelines, especially under constrained settings, the best-performing model may not be the largest. Instead, parameter-efficient backbones such as GPT-2, when combined with LoRA fine-tuning and self-supervised objectives, can yield state-of-the-art results without incurring excessive computational costs.

Logits Fusion Strategies. To assess how different output fusion strategies affect model performance, I conducted an ablation using the GPT-2 backbone under four configurations: (i) using only the temporal branch logits, (ii) using only the spatial branch logits, (iii) averaging the logits from both branches, and (iv) selecting the prediction with the highest softmax confidence (confidence-based fusion). Results are summarized in Table 4.7.

The spatial-only configuration slightly outperformed the temporal-only variant in terms of accuracy (65.10% vs. 65.23%), although the temporal branch maintained a higher F1-score (65.01%). This reflects the complementary nature of spatial and temporal modeling in EEG data, where spatial representations may be more stable, while temporal branches better capture phase and dynamic dependencies.

Among fusion strategies, both average and confidence-based fusion performed similarly, reaching the highest overall accuracy (66.99%) and F1-score (65.78%). These results confirm that combining both branches is beneficial and that confidence-based fusion does not introduce additional instability or bias.

Table 4.7. Comparison of prediction strategies using GPT-2 for arousal classification.

LLM Backbone	**ACC**	**F1**
Temporal Branch Only	65.23%	65.01%
Spatial Branch Only	65.10%	63.46%
Average Fusion	**66.99%**	**65.78%**
Confidence Fusion	66.97%	65.78%

These findings suggest that multi-branch fusion is essential for optimal decoding performance. In particular, the average strategy appears to offer a good balance between accuracy and generalization, making it the default fusion method used in subsequent experiments.

Chapter 5

Conclusions

5.1 Summary of Contributions

This thesis introduced a dual-branch neural architecture for EEG-based emotion recognition, combining a temporal encoder based on a pretrained Large Language Model (LLM) and a spatial encoder based on a Dynamic Graph Convolutional Neural Network (DGCNN). The key novelty lies in the integration of LLMs as temporal sequence processors, fine-tuned directly on EEG data from the DEAP dataset. To the best of my knowledge, this represents the first instance of adapting a foundation model to EEG-based affective computing through direct end-to-end finetuning on DEAP.

To enhance temporal representation learning, a masked reconstruction objective was employed on latent EEG embeddings. This self-supervised auxiliary task regularizes the LLM decoder and encourages it to capture richer temporal dynamics beyond what categorical supervision provides. Parameter-efficient fine-tuning was achieved through Low-Rank Adaptation (LoRA), allowing the use of backbones up to 8 billion parameters (LLaMA 3.1 8B) within the constraints of a single A100 GPU. Experiments show that even lightweight models like GPT-2 can outperform larger LLMs in this domain, confirming that scale does not always correlate with downstream EEG performance.

On the spatial side, the DGCNN encodes electrode interactions through learnable adjacency matrices, capturing localized spatial dependencies in a data-driven fashion. The integration of this spatial path improves classification robustness and complements the temporal stream by leveraging anatomical structure in the EEG layout.

Fusion is performed through two strategies: averaging the logits from both branches, and confidence-based selection of the most reliable prediction on a per-sample basis. Empirical results indicate that the average fusion strategy consistently offers the best trade-off between stability and performance.

Comprehensive experiments on the DEAP dataset demonstrate that the proposed model achieves new state-of-the-art results in subject-dependent arousal and valence recognition, outperforming previous EEG-only models. Extensive ablation studies highlight the impact of each design choice, including the benefit of the reconstruction loss, the contribution of each modality, and the role of segment duration. These

results support the claim that temporal modeling with pretrained LLMs,when guided by EEG-specific objectives and combined with spatial reasoning, provides a promising direction for affective computing.

5.2 Limitations

Although the method seems to show some promising results, there are some limitations in the proposed approach:

- **Data constraints.** The DEAP dataset is limited to just 32 subjects and also contains poor semantic descriptions of internal affect. It will be interesting for future work to generalize to more naturalistic, real-world settings which will necessitate a more diverse and ecologically valid dataset.
- **Inference cost.** While quantization and LoRA lower the resource requirement, big LLMs such as LLaMA-3.1 8B are still computationally expensive for real-time usages. Some further compression techniques or distillation may be required, in order to deploy to wearable systems.
- **Unimodal focus.** It addresses EEG signals as the only available input modality, which allows us to investigate the fine-tuning of LLM in a controlled manner when modeling temporal biosignals. However, affect is inherently multimodal in nature, and using the additional modalities of facial, GSR, or audio sources would likely lead to even better performance and generalizability.

5.3 Future Directions

This work suggests several interesting directions for future work:

- **Multimodal integration utilizing GSR and Mixture of Experts.** The future work may consider additional physiological signals e.g., Galvanic Skin Response (GSR) through the Mixture of Experts (MoE) model. Here, each expert is an expert processing a different modality (e.g., EEG, GSR, audio), and the outputs of the experts are adapted to be combined together. Such modular architectures may lead to a more robust feature extraction as well as the fact that each branch can specialize on merely modality-specific patterns.
- **Direct EEG-to-token conversion in LLMs.** Instead of a Transformer encoder, the EEGs could potentially be quantized to discrete levels and tokenized directly, such that they could serve as native input sequences to an autoregressive LLM. A similar model formulation may be more in line with the language-modeling paradigm, and might be able to elicit richer sequential dynamics and internal representations.
- **EEG-guided music recommendation.** Building on the current framework, the architecture could be extended to support real-time emotion-driven music recommendation. By decoding affective states from EEG signals and mapping them to musical features such as tempo, key, or genre, the system could

dynamically curate playlists that align with or regulate the user's emotional state, paving the way for EEG-based adaptive music therapy or immersive multimedia experiences.

Bibliography

[1] Thomas Bjørner. Using eeg data as dynamic difficulty adjustment in a serious game about the plastic pollution in the oceans. In *Proceedings of the ACM International Conference on Information Technology for Social Good (GoodIT '23)*, pages 1–10, New York, NY, USA, 2023. ACM.

[2] Giuseppe Ciccarelli, Giusy Federico, Giulia Mele, Andrea Di Cecca, Marianna Migliaccio, Carla Rita Ilardi, Vincenzo Alfano, Marco Salvatore, and Carmela Cavaliere. Simultaneous real-time eeg–fmri neurofeedback: A systematic review. *Frontiers in Human Neuroscience*, 17:1123014, 2023.

[3] Eleni Gkintoni, Anastasios Aroutzidis, Helen Antonopoulou, and Charis Halkiopoulos. From neural networks to emotional networks: A systematic review of eeg-based emotion recognition in cognitive neuroscience and real-world applications. *Brain Sciences*, 15(3):220, 2025.

[4] Donald L. Schomer and Fernando H. Lopes da Silva, editors. *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Oxford University Press, New York, 7 edition, 2017. Online edition, accessed 13 July 2025.

[5] Fabien Lotte, Ludovic Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Andriamiarinjatovo Rakotomamonjy, and Florian Yger. A review of classification algorithms for eeg-based brain–computer interfaces: A 10-year update. *Journal of Neural Engineering*, 15(3):031005, 2018.

[6] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[11] Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. Lggnet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9773–9786, 2024.

[12] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Margaret M Bradley and Peter J Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[15] Sangyong Park, Hanmoi Sim, and Wonhyung Lee. Dynamic game difficulty control by using eeg-based emotion recognition. *International Journal of Control and Automation*, 7(3):267–272, 2014.

[16] Shuang Ran, Wei Zhong, Lin Ma, Danting Duan, Long Ye, and Qin Zhang. Mind to music: An eeg signal-driven real-time emotional music generation system. *International Journal of Intelligent Systems*, 2024:Article ID 9618884, 17 pages, 2024.

[17] B. Madhubala, A.V. Senthil Kumar, Kyla L. Tennin, and R.V. Suganya. Personalized music recommendation system for athletes using eeg signals. In *Coaching in Communication Research*, pages 141–168. IGI Global, 2025.

[18] Sara M. Alarcão and Manuel J. Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Trans. Affective Comput.*, 2019.

[19] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Trans. Autonomous Mental Development*, 2015.

[20] M. S. K. Hosseini, S. M. Firoozabadi, K. Badie, and P. Azadfallah. Personality-based emotion recognition using eeg signals with a cnn-lstm network. *Brain Sciences*, 2023.

[21] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[22] Jonathan W. Kim, Ahmed Alaa, and Danilo Bernardo. Eeg-gpt: Exploring capabilities of large language models for eeg classification and interpretation. *arXiv preprint arXiv:2401.18006*, 2024.

[23] Naseem Babu, Jimson Mathew, and A. P. Vinod. Large language models for eeg: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2506.06353*, 2025.

[24] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2020.

[25] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Weizhu Wang, and Zichao Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[27] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[28] Valer Jurcak, Daisuke Tsuzuki, and Ippeita Dan. 10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems. *NeuroImage*, 34(4):1600–1611, 2007.

[29] Yuhong M. Chi, Yijun Wang, Ying Wang, Charles Maier, Tzyy-Ping Jung, and Gert Cauwenberghs. Dry and noncontact eeg sensors for mobile brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(2):228–235, Dec 2011.

[30] Arnaud Delorme and Scott Makeig. Eeglab: An open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, Mar 2004.

[31] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, 2015.

[32] Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. Single-trial analysis and classification of erp components: A tutorial. *NeuroImage*, 2011.

[33] Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. Identifying stable patterns over time for emotion recognition from eeg. *IEEE Transactions on Affective Computing*, 10(3):417–429, 2019.

[34] Miguel A Blanco-Ríos, Milton O Candela-Leal, et al. Real-time eeg-based emotion recognition model using principal component analysis and tree-based models for neurohumanities. *arXiv preprint arXiv:2401.15743*, 2024.

[35] Jing Zhang, Yu Wang, and Bao-Liang Lu. Graph convolutional networks for eeg-based emotion recognition. *IEEE Transactions on Affective Computing*, 2022.

[36] Stefano Valenzi, Md Kamrul Islam, Peter Jurica, and Andrzej Cichocki. Individual classification of emotions using eeg. *Journal of Biomedical Science and Engineering*, 7:604–620, 2014.

[37] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. Eeg-based emotion recognition using frequency domain features and support vector machines. In *Proc. ICONIP*, 2011.

[38] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[39] Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[40] Xin Zhang et al. Spatial-temporal transformers for eeg emotion recognition. *arXiv preprint arXiv:2110.06553*, 2022.

[41] Arjun Arjun, Aniket Singh Rajpoot, and Mahesh Raveendranatha Panicker. Introducing attention mechanism for eeg signals: Emotion recognition with vision transformers. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5723–5726, 2021.

[42] Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. Enhancing eeg-to-text decoding through transferable representations from pre-trained contrastive eeg-text masked autoencoder. *arXiv preprint arXiv:2402.17433*, 2024.

[43] Dong Hyeok Lee and Chun Kee Chung. Enhancing neural decoding with large language models: A gpt-based approach. In *2024 12th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–4, 2024.

[44] Xiao Liu, Yicheng Ji, Yixin Fu, Yanan Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2022.

[45] Yann LeCun et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[47] Nitish Srivastava et al. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[48] Hongyi Zhang et al. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[49] Sangdoo Yun et al. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[50] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(1):4–24, 2021.

[51] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[52] Petar Veličković et al. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[53] Minghao Xiao, Zhengxi Zhu, Kang Xie, and Bin Jiang. Meeg and at-dgnn: Improving eeg emotion recognition with music introducing and graph-based learning. *arXiv preprint arXiv:2407.05550*, 2024.

[54] Yi-Chun Huang, Yong Lu, Xin-Yu Si, and Jing Yang. Mt-lgsgcn: Eeg-based emotion recognition using multi-scale temporal and local-global spatial graph convolution network. In *Proc. IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM)*, 2023.

[55] Stamos Katsigiannis and Naeem Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2018.

[56] Mohammad Soleymani, Johan Lichtenauer, Thierry Pun, and Maja Pantic. A survey of multimodal sentiment and emotion analysis. *Image and Vision Computing*, 65:3–14, 2017.

[57] Wei-Long Zheng et al. Multichannel eeg-based emotion recognition via group sparse canonical correlation analysis. *IEEE Transactions on Cognitive and Developmental Systems*, 9(3):281–290, 2017.

[58] Alex Craik, Yuan He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of Neural Engineering*, 16(3):031001, 2019.

[59] Qingshan She, Xinsheng Shi, Feng Fang, Yuliang Ma, and Yingchun Zhang. Cross-subject eeg emotion recognition using multi-source domain manifold feature selection. *Computers in Biology and Medicine*, 159, 2023.

[60] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[61] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[62] Masahiro Suzuki and Yutaka Matsuo. A survey of multimodal deep generative models. *arXiv preprint arXiv:2207.02127*, 2022.

[63] Sheng Wang et al. Contrastive learning for multimodal emotion recognition. In *Proc. AAAI Conf. Artif. Intell.*, 2022.

[64] Yuzhe Zhang, Huan Liu, Yang Xiao, Mohammed Amoon, Dalin Zhang, Di Wang, Shusen Yang, and Chai Quek. Llm-enhanced multi-teacher knowledge distillation for modality-incomplete emotion recognition in daily healthcare. *IEEE Journal of Biomedical and Health Informatics*, pages 1–11, 2024.

[65] Robin T Schirrmeister, Jost Tobias Springenberg, Lukas D J Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

[66] Vernon J Lawhern, Alex J Solon, Nicholas R Waytowich, Steven M Gordon, Catherine P Hung, and Brent J Lance. Eegnet: A compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

[67] Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *arXiv preprint arXiv:2104.02935*, 2022.

[68] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2023.

[69] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852, 2016.