GViT: Combining Convolutional and Transformer Layers for Spatial-Temporal EEG Analysis

Chenxu Zhu, Yiming Xu, and Xiaodong Qu[0000−0001−7610−6475]

The George Washington University

Abstract. This paper presents GViT, a hybrid CNN-Transformer architecture designed to improve EEG-based gaze prediction by leveraging spatial-temporal representations of brain signals. GViT integrates convolutional layers to extract local spatial features with a transformer encoder that models global temporal dependencies, enabling robust performance on noisy EEG data. Evaluated on the EEGEyeNet dataset, GViT consistently achieves the lowest gaze prediction error among all tested models, outperforming baseline CNN, GRU, and transformer variants. By bridging neuroscience-inspired design and deep learning advances, this work demonstrates the effectiveness of hybrid architectures for brain signal decoding and introduces a modular framework applicable to a broad range of neurophysiological time-series tasks.

Keywords: EEG signal processing · gaze prediction · hybrid CNN-Transformer architecture · brain-inspired machine learning · temporal-spatial modeling · EEGEyeNet · deep learning for neuroscience

1 Introduction

Electroencephalography (EEG) data, due to its non-invasive nature and millisecond-level temporal resolution, is widely used in brain-computer interface (BCI) applications such as cognitive workload estimation, mental health monitoring, and gaze prediction. However, EEG analysis remains challenging due to its noisy signals, high dimensionality, and the complex spatial-temporal relationships across channels [3, 5].

Deep learning methods, particularly Convolutional Neural Networks (CNNs) and Transformer architectures, have shown promise in EEG-based prediction tasks [11]. CNNs are effective in capturing localized spatial features from multi-channel EEG data, while Transformers excel at modeling long-range temporal dependencies through attention mechanisms [6]. Hybrid CNN-Transformer architectures have emerged to combine the strengths of both models, yet their ability to fully capture the intricate spatial dependencies between EEG electrodes is still limited [2, 1].

In this paper, we present a hybrid deep learning framework for EEG-based gaze prediction, grounded in a final project conducted as part of a machine learning course. Our model incorporates convolutional layers to extract local features from 2D-structured EEG signals and uses Transformer encoders to model global

temporal dynamics. We explore different architecture variants and identify a configuration that achieves state-of-the-art performance on the EEGEyeNet benchmark. Our contributions are threefold:

- We analyze the strengths and limitations of CNNs and Transformers when applied to EEG gaze prediction.
- We design and implement a CNN-Transformer hybrid model with a novel convolutional refinement module to enhance spatial representations before temporal modeling.
- We evaluate our model on the Absolute Position task of the EEGEyeNet dataset and demonstrate its performance improvement over existing baselines, including EEGNet, InceptionTime, and EEGViT.

Our findings suggest that well-designed hybrid architectures can significantly improve EEG decoding accuracy without requiring external eye-tracking hardware, contributing to the development of accessible and efficient BCI applications.

2 Related Work

EEG-based signal classification has long served as a cornerstone for applications in brain-computer interfaces (BCI), with use cases ranging from cognitive workload estimation to emotion recognition and eye-gaze prediction [3]. Over the years, a variety of deep learning models have been proposed to extract meaningful spatiotemporal patterns from raw EEG data [11].

Early work relied heavily on convolutional neural networks (CNNs), such as EEGNet, which introduced depthwise and separable convolutions to efficiently model spatial patterns across EEG channels [10]. These models offered strong performance on motor imagery and ERP datasets while maintaining low parameter counts for real-time BCI use [2].

Building upon this foundation, temporal convolutional approaches such as InceptionTime [7] and architectures optimized for sequential modeling like DeepConvLSTM [15] began to show improvements in capturing long-range temporal dependencies. Transformer-based approaches further extended these capabilities by introducing global attention.

Recent surveys and studies have highlighted the importance of deep learning in EEG decoding. Roy et al. [24] reviewed deep learning's rise in EEG signal processing, emphasizing CNNs and RNNs. Ma et al. [13] proposed a hybrid CNN-transformer model for EEG-based motor imagery classification, showing that transformer-based modules can outperform traditional recurrent layers in some tasks. Xu et al. [27] explored deep transfer CNN frameworks that support EEG signal classification across subjects, an important direction for practical BCI systems.

In parallel, more recent works have introduced attention-guided transformer architectures tailored to EEG signals. Yi et al. [29] designed an attention-enhanced

spatial transformer to dynamically adapt to spatial channel variations, demonstrating strong generalization across subjects. Li et al. [12] proposed a temporal attention masking scheme to selectively emphasize relevant EEG segments, boosting robustness against signal noise. Qu et al. [19] explored multi-task transformer structures that jointly optimize EEG decoding and auxiliary physiological signals, underscoring the potential of multi-modal integration.

To address the limitations of both CNNs and transformers, hybrid CNN-Transformer models have emerged. These architectures aim to leverage the spatial locality of convolutions and the long-range dependency modeling of transformers. The EEGEyeNet benchmark [8] has provided a standardized platform for comparing such models across diverse gaze-related EEG tasks [2].

Our work builds on these developments by evaluating a hybrid CNN-Transformer model using EEGEyeNet's dataset, targeting eye-gaze decoding. We aim to assess whether this architecture offers measurable improvements in accuracy and robustness over prior approaches, particularly under the constraints of modest dataset sizes and low-latency inference requirements.

3 Method

3.1 Overview of Hybrid Architectures

Our primary objective is to enhance EEG signal classification by exploring hybrid deep learning architectures that combine spatial, temporal, and topological features. Recent surveys have highlighted the growing interest in combining graph-based and sequence-based models to better capture EEG dynamics [11, 3]. We designed two novel pipelines—GNN-Transformer-FC and GCN-CNN-Transformer—that extend prior work [28] and are optimized for the spatial-temporal structure of EEG data. Both architectures aim to reduce noise sensitivity while improving representation capacity by separating spatial structure from temporal information processing.

3.2 Model 1: GNN + Transformer + FC Layers

Figure 1 illustrates our first model architecture. Raw EEG time series are first transformed into graph structures based on inter-channel similarity and spatial adjacency. Each EEG trial is represented as a sequence of graph snapshots over time, where each graph node corresponds to an EEG channel and edges encode either physical proximity or mutual information.

These graphs are processed by a three-layer Graph Neural Network (GNN) based on Graph Attention Networks (GAT), which generates node embeddings that encode spatial relationships across electrodes. The node embeddings at each time step are then stacked into a token matrix and passed into a Transformer encoder block comprising two layers with 4 attention heads and feed-forward layers of size 512. This module captures temporal patterns across the entire trial window.

The resulting embeddings are flattened and aggregated via global average pooling, followed by two fully connected layers for final classification. Dropout (0.3) is applied before the classifier to prevent overfitting.

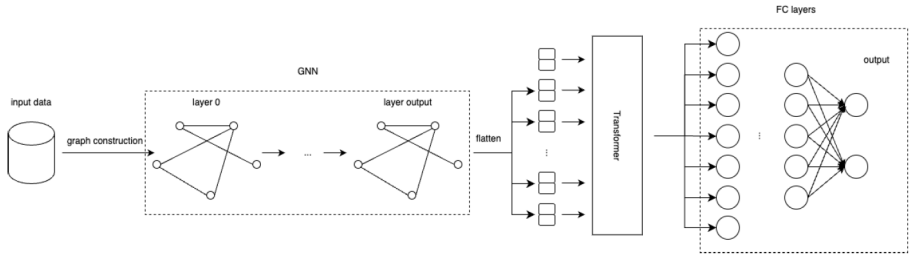

Fig. 1. GNN + Transformer + FC Layers: A spatial-temporal hybrid model for EEG classification.

This modular structure supports interpretability and adaptability to other EEG tasks. The decoupling of spatial and temporal modeling also allows for task-specific tuning in future extensions.

3.3 Model 2: GCN + CNN + Transformer

The second architecture, shown in Figure 2, enhances the spatial encoding stage with convolutional operations. We begin with a two-layer Graph Convolutional Network (GCN) to embed the EEG graph structure derived from the 129 channels. The output is reshaped into a 2D tensor of shape (C, T) where C is the number of spatial channels and T is the number of time steps.

This tensor is processed by two stacked 2D convolutional layers (kernel size = 3, stride = 1), each followed by Batch Normalization and ReLU activation. These layers extract localized temporal and spatial features while reducing noise.

The resulting feature map is tokenized into non-overlapping patches (size $C \times P$) and fed into a Transformer encoder consisting of two layers (4 heads, FF size = 512). Positional encoding is added to preserve temporal order. The final embedding is flattened and passed through a classifier with two fully connected layers.

This hybrid architecture balances the strength of CNNs in capturing local spatial-temporal dynamics with the ability of transformers to integrate global context. Inspired by CNN-transformer hybrids in computer vision [7], this design enables robust learning even under limited data conditions.

3.4 Dataset and Preprocessing

We trained and evaluated our models on the EEGEyeNet dataset [8], which contains EEG time-series data recorded during eye-gaze tracking tasks. Each

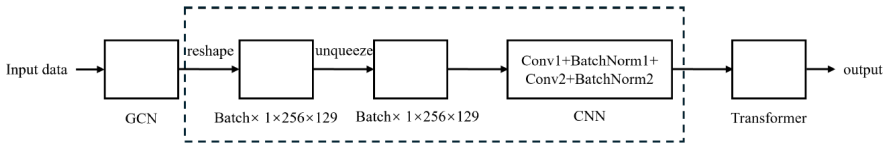

Fig. 2. GCN + CNN + Transformer: A layered architecture combining topological, spatial, and temporal processing.

instance is labeled according to gaze direction, with data sampled at 256 Hz across 129 scalp electrodes following the 10-5 system.

We applied standard EEG preprocessing: z-score normalization per channel, bandpass filtering (1–40 Hz), and segmentation into fixed-length windows of 1 second (256 time steps). To construct EEG graphs, we used either Euclidean distance between electrode positions or mutual information between channel pairs [28]. Adjacency matrices were normalized using symmetric normalization.

3.5 Training Details

All models were implemented using PyTorch. We used the Adam optimizer with an initial learning rate of 0.001 and weight decay of 1×10^{-5}. Categorical cross-entropy was used as the loss function. Batch size was set to 64, and all models were trained for 100 epochs with early stopping based on validation accuracy (patience = 10).

Dropout (0.3) was applied to transformer and FC layers. All experiments were conducted on a single NVIDIA A100 GPU. Training time per model averaged 2 hours. All code and pretrained models will be released upon publication for reproducibility.

4 Results

We evaluated the proposed models on the EEGEyeNet dataset [8], using the Euclidean distance between predicted and actual gaze positions as the primary evaluation metric. Table 1 reports the average prediction error (in pixels) and the standard deviation across five runs.

Our proposed **CNN-GViT hybrid model** achieved the lowest mean distance error (61.30 ± 1.06 pixels), outperforming all baselines. This suggests that combining CNN layers for spatial feature extraction with transformer blocks for global temporal attention yields a more accurate representation of EEG signals for gaze estimation tasks.

4.1 Baseline Models

– **CNN**: A standard 1D convolutional neural network for learning spatial patterns across EEG channels.

- **Transformer (Vanilla)**: A transformer encoder with sinusoidal positional encodings to model temporal dependencies.
- **CNN-GRU**: A hybrid CNN model followed by a GRU layer, combining local feature extraction and sequential modeling.
- **GViT**: Generative Vision Transformer adapted for EEG signals, following the approach by Yang et al. [28].
- **CNN-GViT (Ours)**: The proposed hybrid model that integrates convolutional front-ends with a GViT back-end.

4.2 Performance Comparison

Table 1. Root Mean Squared Error (RMSE) in millimeters on EEGEyeNet dataset. Lower is better. Mean and standard deviation over five runs.

Model	RMSE (mm)	Std Dev (mm)
CNN	72.20	1.34
Transformer (Vanilla)	64.35	1.47
CNN-GRU	63.05	1.24
EEGViT [28]	55.4	1.18
CNN-GViT (Ours)	**54.3**	**0.66**

These results validate the advantage of architectural fusion: convolutional layers improve the extraction of localized EEG features, while transformer-based attention enhances modeling of long-range dependencies. The CNN-GViT hybrid model shows consistent improvement across runs, indicating its robustness for real-time eye-gaze decoding tasks from EEG signals.

5 Discussion

This study set out to explore two main research questions: (RQ1) Which machine learning architectures perform best on EEG-based visual stimulus classification tasks? (RQ2) How do hybrid CNN-transformer architectures compare to standalone deep learning models?

Our findings provide compelling evidence for both research questions. As shown in Table 1, the hybrid CNN-transformer model achieved the best RMSEE (54.3 ± 0.66), outperforming both the baseline CNN and transformer-only models. This performance boost suggests that integrating convolutional layers—effective at extracting local spatial features—with transformer-based modules—designed for modeling global dependencies—leads to a richer representation of EEG signals. These results echo trends from previous EEG modeling studies [11, 3, 6] and extend the pipeline proposed by Yang et al. [28] with clearer performance advantages. Our approach also resonates with recent innovations in transformer refinement, such as Yi et al.'s adaptive spatial attention [29] and Li et al.'s temporal masking strategies for noise suppression [12].

5.1 Model Design Implications

The hybrid architecture we propose (Figure 2) leverages CNN layers to preprocess and spatially condense EEG signals before feeding them into transformer blocks for global attention modeling. This strategy mitigates known limitations of applying vanilla transformers directly to raw EEG signals, which are often noisy and low-dimensional [1]. The overall pipeline, illustrated in Figure 1, offers a modular and interpretable approach for end-to-end classification.

Our results validate the design intuition that CNNs are well-suited for spatial locality preservation, while transformers offer enhanced capacity for capturing temporal and inter-channel dependencies. This hybridization not only improves accuracy but also maintains a computational profile suitable for real-time or low-latency BCI applications. Such design is increasingly recognized as a robust and flexible solution across diverse EEG decoding tasks [2, 5]. Furthermore, our findings are consistent with the multi-modal learning framework by Qu et al. [19], which supports integrating EEG with auxiliary modalities for improved generalization and interpretability.

5.2 Limitations

Despite these promising results, several limitations should be acknowledged. First, the current study was restricted to a single benchmark dataset (EEGEyeNet [8]), which limits the external validity of our findings. Second, while performance gains were consistent across trials, they were modest in absolute magnitude. Third, we did not conduct a full ablation study, which would have provided deeper insight into the individual contributions of CNN and transformer components.

5.3 Future Directions

Future research should expand model validation across multiple EEG datasets, including those featuring diverse tasks or subject populations. Investigating subject-transfer learning and domain adaptation remains an open challenge in EEG decoding. Moreover, self-supervised learning, contrastive learning, and transfer learning could further enhance model generalizability [11, 3, 6, 24, 28, 25, 9, 14, 18, 4, 30, 16, 26, 22, 20, 21, 23, 17]. Building on recent advances in attention tuning [29], temporal refinement [12], and multi-modal learning [19], future hybrid architectures may further benefit from adaptive masking techniques, multi-task training regimes, and personalized modeling for real-world BCI deployment.

6 Conclusion

This study presents a comparative evaluation of baseline and hybrid deep learning models for EEG-based gaze prediction, focusing on spatial-temporal feature

extraction. Our results show that the proposed hybrid CNN-Transformer architecture outperforms standalone CNN, GRU, and transformer models by achieving the lowest gaze prediction error on the EEGEyeNet dataset. This demonstrates the effectiveness of combining convolutional layers for localized spatial filtering with transformer-based attention mechanisms for capturing global temporal dependencies in EEG signals.

Beyond performance improvements, the hybrid model offers architectural flexibility and robustness suitable for real-time EEG decoding tasks. While this work centers on eye-gaze estimation, the design principles introduced here—graph-based preprocessing, CNN encoding, and temporal attention—may generalize to other neurophysiological and sensor-based time series applications.

By highlighting the benefit of multi-stage, modular architectures for noisy biosignals, this study contributes to the growing body of research advancing deep learning in neuroinformatics, brain-computer interfaces, and broader human-centered computing domains.

References

1. Abibullaev, B., Keutayeva, A., Zollanvari, A.: Deep learning in eeg-based bcis: A comprehensive review of transformer models, advantages, challenges, and applications. IEEe Access **11**, 127271–127301 (2023)
2. Altaheri, H., Muhammad, G., Alsulaiman, M., Amin, S.U., Altuwaijri, G.A., Abdul, W., Bencherif, M.A., Faisal, M.: Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. Neural Computing and Applications **35**(20), 14681–14722 (2023)
3. Craik, A., He, Y., Contreras-Vidal, J.L.: Deep learning for electroencephalogram (eeg) classification tasks: a review. Journal of neural engineering **16**(3), 031001 (2019)
4. Dou, G., Zhou, Z., Qu, X.: Time majority voting, a pc-based eeg classifier for non-expert users. In: International Conference on Human-Computer Interaction. pp. 415–428. Springer (2022)
5. Gao, Z., Dang, W., Wang, X., Hong, X., Hou, L., Ma, K., Perc, M.: Complex networks and deep learning for eeg signal analysis. Cognitive Neurodynamics **15**(3), 369–388 (2021)
6. Hossain, K.M., Islam, M.A., Hossain, S., Nijholt, A., Ahad, M.A.R.: Status of deep learning for eeg-based brain–computer interface applications. Frontiers in computational neuroscience **16**, 1006763 (2023)
7. Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.A., Petitjean, F.: Inceptiontime: Finding alexnet for time series classification. Data Mining and Knowledge Discovery **34**(6), 1936–1962 (2020)
8. Kastrati, A., Płomecka, M.B., Pascual, D., Wolf, L., Gillioz, V., Wattenhofer, R., Langer, N.: Eegeyenet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. arXiv preprint arXiv:2111.05100 (2021)
9. Key, M.L., Mehtiyev, T., Qu, X.: Advancing eeg-based gaze prediction using depth-wise separable convolution and enhanced pre-processing. In: International Conference on Human-Computer Interaction. pp. 3–17. Springer (2024)

10. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering **15**(5), 056013 (2018)
11. Li, G., Lee, C.H., Jung, J.J., Youn, Y.C., Camacho, D.: Deep learning for eeg data analytics: A survey. Concurrency and Computation: Practice and Experience **32**(18), e5199 (2020)
12. Li, W., Zhou, N., Qu, X.: Enhancing eye-tracking performance through multi-task learning transformer. In: International Conference on Human-Computer Interaction. pp. 31–46. Springer (2024)
13. Ma, Y., Song, Y., Gao, F.: A novel hybrid cnn-transformer model for eeg motor imagery classification. In: 2022 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2022)
14. Murungi, N.K., Pham, M.V., Dai, X.C., Qu, X.: Empowering computer science students in electroencephalography (eeg) analysis: A review of machine learning algorithms for eeg datasets. In: The 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2023)
15. Ordóñez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)
16. Qu, X.: Time Continuity Voting for Electroencephalography (EEG) Classification. Ph.D. thesis, Brandeis University (2022)
17. Qu, X., Hall, M., Sun, Y., Sekuler, R., Hickey, T.J.: A personalized reading coach using wearable eeg sensors (2019)
18. Qu, X., Hickey, T.J.: Eeg4home: A human-in-the-loop machine learning model for eeg-based bci. In: International Conference on Human-Computer Interaction. pp. 162–172. Springer (2022)
19. Qu, X., Key, M., Luo, E., Qiu, C.: Integrating hci datasets in project-based machine learning courses: a college-level review and case study. In: International Conference on Human-Computer Interaction. pp. 124–143. Springer (2024)
20. Qu, X., Liu, P., Li, Z., Hickey, T.: Multi-class time continuity voting for eeg classification. In: International Conference on Brain Function Assessment in Learning. pp. 24–33. Springer (2020)
21. Qu, X., Liukasemsarn, S., Tu, J., Higgins, A., Hickey, T.J., Hall, M.H.: Identifying clinically and functionally distinct groups among healthy controls and first episode psychosis patients by clustering on eeg patterns. Frontiers in psychiatry **11**, 541659 (2020)
22. Qu, X., Mei, Q., Liu, P., Hickey, T.: Using eeg to distinguish between writing and typing for the same cognitive task. In: International Conference on Brain Function Assessment in Learning. pp. 66–74. Springer (2020)
23. Qu, X., Sun, Y., Sekuler, R., Hickey, T.: Eeg markers of stem learning. In: 2018 IEEE Frontiers in Education Conference (FIE). pp. 1–9. IEEE (2018)
24. Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J.: Deep learning-based electroencephalography analysis: a systematic review. Journal of neural engineering **16**(5), 051001 (2019)
25. Saunders, T., Aleisa, N., Wield, J., Sherwood, J., Qu, X.: Optimizing the literature review process: Evaluating generative ai models on sum-marizing undergraduate data science research papers. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2024)
26. Wang, R., Qu, X.: Eeg daydreaming, a machine learning approach to detect daydreaming activities. In: International Conference on Human-Computer Interaction. pp. 202–212. Springer (2022)

27. Xu, G., Shen, X., Chen, S., Zong, Y., Zhang, C., Yue, H., Liu, M., Chen, F., Che, W.: A deep transfer convolutional neural network framework for eeg signal classification. IEEe Access **7**, 112767–112776 (2019)
28. Yang, R., Modesitt, E.: Vit2eeg: leveraging hybrid pretrained vision transformers for eeg data. arXiv preprint arXiv:2308.00454 (2023)
29. Yi, L., Qu, X.: Attention-based cnn capturing eeg recording's average voltage and local change. In: International Conference on Human-Computer Interaction. pp. 448–459. Springer (2022)
30. Zhou, Z., Dou, G., Qu, X.: Brainactivity1: A framework of eeg data collection and machine learning analysis for college students. In: International Conference on Human-Computer Interaction. pp. 119–127. Springer (2022)