Mitigating Subject Dependency in EEG Decoding with Subject-Specific Low-Rank Adapters

Timon Klein* **Piotr Minakowski*** **and** **Sebastian Sager*†**

Abstract

Subject-specific distribution shifts represent an important obstacle to the development of foundation models for EEG decoding. To address this, we propose Subject-Conditioned Layer, an adaptive layer designed as a drop-in replacement for standard linear or convolutional layers in any neural network architecture. Our layer captures subject-specific variability by decomposing its weights into a shared, subject-invariant component and a lightweight, low-rank correction unique to each subject. This explicit separation of general knowledge from personalized adaptation allows existing models to become robust to subject shifts. Empirically, models equipped with our layer outperform both a shared-weight-only model (subject-agnostic model) and the average of individually trained subject-specific models. Consequently, the Subject-Conditioned Layer offers a practical and scalable path towards building effective cross-subject foundation models for EEG.

1 Introduction

Developing a foundation model for decoding complex brain signals like Electroencephalography (EEG) is a central challenge in neuroscience and machine learning [2]. Progress is slowed by the inherent properties of EEG data: low signal-to-noise ratio, non-stationarity, and, most critically, high inter-subject variability [35]. Neural patterns vary so widely across individuals, and even within the same individual over time, that data is effectively fragmented, exacerbating the problem of data scarcity [3, 25]. Consequently, large models trained on pooled data from many subjects often fail to outperform simpler models tailored to individuals, stalling progress towards generalizable solutions.

This situation contrasts sharply with fields like computer vision (CV) and natural language processing (NLP). While these fields have also relied on domain-specific architectures, the Transformer [34] provided a universal paradigm that unlocked predictable scaling laws, enabling massive foundation models to emerge [13, 37]. EEG decoding has yet to benefit from similar scaling. The challenge of cross-subject generalization has kept effective dataset sizes small, preventing the successful application of large-scale, general-purpose architectures. As a result, the field remains dominated by specialized models that have not yet converged to a standard, foundational approach [2, 12].

To bridge this gap, we address the problem of subject-specific distribution shifts, which we identify as the primary obstacle to building foundation models for multi-subject problems such as EEG classification. We propose Subject-Conditioned Layer, an approach that decomposes the parameters of a neural network into two components: a set of base weights shared across all subjects and a subject-specific, low-rank correction that efficiently captures individual variability. The Subject-Conditioned Layer module is designed as a drop-in replacement for standard linear or convolutional layers, making it universally applicable to a wide range of model architectures with applications in EEG decoding and beyond.

*Department of Mathematics; Otto von Guericke University Magdeburg; 39106 Magdeburg; Germany.
†Max Planck Institute for Dynamics of Complex Technical Systems; 39106 Magdeburg; Germany.
Mail correspondence: timon.klein@ovgu.de

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

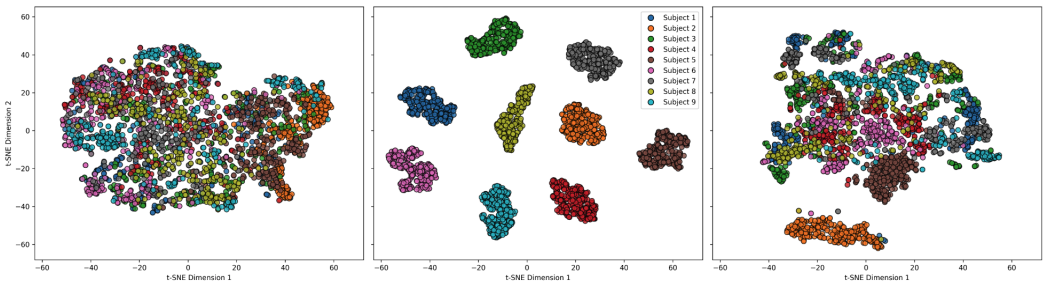

Figure 1: A t-SNE projection of latent representations from different layer components. The panels compare representations derived from: **(Left)** the shared weights (W_{general}), **(Middle)** the subject-specific weights (W_s), and **(Right)** the fused representations from the complete Subject-Conditioned Layer. Representations from the general weights show no subject-specific structure, while those from the subject-specific weights form distinct clusters for each subject. Our full Subject-Conditioned Layer model preserves subject identity while forming meaningful sub-clusters within each subject, which correspond to the different target labels, cf. Figure 3. All representations are from the classification token predicted by the Patched Brain Transformer; a linear classification head was used for visualization clarity.

Our contributions are threefold: We introduce Subject-Conditioned Layer, a novel adaptive layer that explicitly separates shared knowledge from subject-specific signatures and is designed as a drop-in replacement for standard linear and convolutional layers in any neural network architecture. We provide qualitative evidence through embedding visualization that our layer successfully disentangles subject-invariant neural patterns from individual subject signatures, confirming the mechanism behind our method's success. Through a comprehensive evaluation on the BCI Competition IV datasets, we demonstrate that both CNN and ViT like architectures equipped with Subject-Conditioned Layer outperform strong baselines: (i) a subject-agnostic model, (ii) the average of subject-specific models, and (iii) the average of subject-specific LoRA models. The code is publicly available [3].

2 On the Difficulty of Multi-Subject Training

We consider a classification problem where data is collected from multiple subjects. Since the signal contains subject-specific characteristics, our goal is to learn a general model that captures class-discriminative structures across subjects. Let $\mathcal{X}_s = \{x_{sj}\}_{j=1}^{M_s}$ and $\mathcal{Y}_s = \{y_{sj}\}_{j=1}^{M_s}$ contain subject-specific features and labels, respectively, for subjects $s \in \{1, \ldots, N\}$. In short, we write $\mathcal{D} = (\mathcal{X}, \mathcal{Y}) = \bigcup_{s=1}^{N} \mathcal{D}_s = \bigcup_{s=1}^{N} (\mathcal{X}_s, \mathcal{Y}_s)$.

A single linear layer trained for all subjects implicitly estimates $p(\mathcal{Y}|\mathcal{X})$, but cannot properly capture subject-specific features because it lacks a mechanism to adapt its weights per subject. It can only represent the shared average structure. We address this with a probabilistic framework that models subject-specific effects. Our main goal is to learn the subject-independent classifier $p(\mathcal{Y}|\mathcal{X})$ that for a given $\mathcal{D}$ marginalizes over s,

$$p(\mathcal{Y} \mid \mathcal{X}) = \textstyle\sum_{s=1}^{N} p(\mathcal{Y} \mid \mathcal{X}, s). \tag{1}$$

Further, assume that we have latent random variables $\mathcal{Z} = \xi(\mathcal{X}, s)$ as intermediates on the map from $\mathcal{X}$ to $\mathcal{Y}$, where we define ξ as a transformation to latent representation. Then we can write

$$p(\mathcal{Y} \mid \mathcal{X}, s) = p(\mathcal{Y} \mid \mathcal{Z}, s)\; p(\mathcal{Z} \mid \mathcal{X}, s). \tag{2}$$

Concerning the first term of Equation (2) we assume that

$$p(\mathcal{Y} \mid \mathcal{Z}, s) = p(\mathcal{Y} \mid \mathcal{Z}),$$

which states that once the latent representation $\mathcal{Z}$ is known, the participant index s and the datapoint x_{sj} provide no additional information about the label y_{sj}. This conditional independence assumption

[3] https://github.com/timonkl/SubjectConditionedLayer

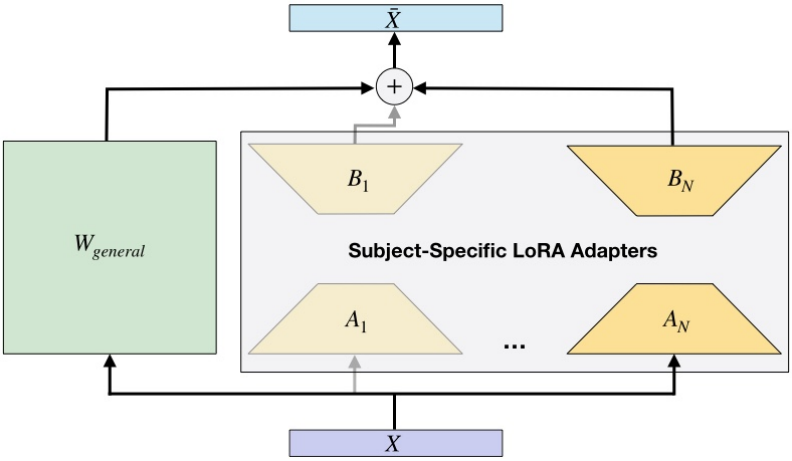

Figure 2: **Subject-Conditioned Layer**. The input X is processed by a shared general weight matrix W_{general}. For each subject, a subject-specific low-rank adapter, parameterized by matrices A_s and B_s of rank r, computes a correction that is added to the output of the general weight matrix, yielding the layer output $\bar{X}$.

implies that $\mathcal{Z}$ serves as a sufficient statistic of x_{sj} to predict y_{sj}. A key step is to verify whether this property holds for the proposed method.

The second term of Equation (2), $p(\mathcal{Z} \mid \mathcal{X}, s)$, describes how the latent representation is learned. Specifically,

$$\xi(\mathcal{X}, s) \sim p(\mathcal{Z} \mid \mathcal{X}, s), \tag{3}$$

i.e., ξ is treated as a stochastic latent variable in the representation space $\mathcal{Z}$, sampled conditionally on the input $\mathcal{X}$ and the participant index s. The subject-specific map $\xi(\mathcal{X}, s)$ is parameterized by a set of general weights and a subject-specific low-rank correction.

3 Method

3.1 Subject-Conditioned Layer

We propose a novel subject-aware layer for multi-subject machine learning tasks as

$$\bar{X} = \sigma\left(XW_{\text{general}}^{\top} + \textstyle\sum_{s=1}^{N}(M_s \cdot X)W_s^{\top}\right), \tag{4}$$

where $W_{\text{general}} \in \mathbb{R}^{m\times n}$ and $W_s \in \mathbb{R}^{m\times n}$ are shared and subject-specific weight matrices, respectively. The mask matrices M_s with binary entries extract subject-s-specific samples from the batch data X. Note that this approach may lead to a reduction in batch sizes for subject-dependent weights W_s, if the number of subjects N increases. As a remedy, adapting the total batch size was proposed in [31].

3.2 Low Rank approximation for Subject-Dependent Weights

Low-rank adaptation (LoRA) [14] is a well-established method for fine-tuning models in few-shot settings. Compared to full-rank training, LoRA shows superior performance on limited data [27], while also substantially reducing the number of training parameters and the associated memory consumption. We define subject-specific weights as low-rank matrix decompositions

$$W_s^{\top} := A_s B_s, \tag{5}$$

with $A_s \in \mathbb{R}^{n\times r}$, $B_s \in \mathbb{R}^{r\times m}$ and rank r. Similarly to [14], we initialize $A_s = \mathcal{N}(0, \sigma^2)$, and $B_s = 0$ for all subjects s.

3.3 Regularization of Subject-Dependent Weights

A central challenge is to ensure that the general weight matrix W_{general} learns shared patterns while the subject-specific weights W_s only capture individual deviations. To prevent W_s from learning general representations, we employ a regularization strategy. First, we constrain the complexity of the subject-specific corrections by enforcing a relatively small rank r, as defined in Equation 5. See Tables 2 and 3 for exact values.

Second, we introduce an optimization-based regularization method by scaling W_s with $\frac{\alpha}{r}$, where α is a hyperparameter. This scaling adjusts the effective learning rate of the adapter matrices relative to the general weights. For $\alpha < r$, the adapters are updated conservatively, encouraging reliance on W_{general}. In contrast to $\alpha > r$, where the adapters are emphasized.

The weights W_s could also be regularized by adding a weight regularization term (e.g., an L_2 penalty) to the loss function. We did not pursue this here, opting instead to demonstrate the capabilities of the LoRA only approach.

4 Data

We conduct our experiment on two Brain Computer Interface datasets from BCI Competition IV [32]. The data are publicly available[4] with the MOABB-Package [7, 1]. The BCI Competition IV 2a dataset [4] consists of 22-channel EEG recordings from nine subjects. For each subject, data were collected in two sessions (training and evaluation), with 288 trials per session. The labels are balanced across four motor imagery classes: left hand, right hand, foot, and tongue. The BCI Competition IV 2b dataset [20, 19] consists of 3-channel EEG recordings from nine subjects. For each subject, data were collected across five sessions, three for training and two for evaluation, totaling between 680 and 760 trials. The labels are balanced between the classes: left hand and right hand.

5 Results

We evaluate our proposed method, Subject-Conditioned Layer, by integrating it into two distinct EEG decoding architectures: EEGNeX [6], a compact Convolutional Neural Network (CNN), and the Patched Brain Transformer (PBT) [17], a Vision Transformer (ViT) based architecture [8] that reflect recent trends in the field [18, 16]. These models demonstrate the versatility of Subject-Conditioned Layer across diverse architecture types. All experiments were conducted on the BCI Competition IV-2a and IV-2b datasets. Further details on the experimental setup are in Appendix C.1.

5.1 Qualitative Analysis

To confirm our hypothesis that our Subject-Conditioned Layer learns to decouple shared from subject-specific information, we visualize the PBT's latent representations using t-SNE [33], see Figure 1. The visualizations reveal a clear separation of the learned representations. As shown in the figure, embeddings from the shared weights (W_{general}) are unstructured by subject (left), whereas those from the subject-specific weights (W_s) form distinct, individual-specific clusters (middle). The full model (right) integrates both aspects: it preserves the subject-level clustering while also forming meaningful sub-clusters corresponding to target labels, Figure 3, demonstrating that it retains task-relevant information. This demonstrates that the low-rank corrections W_s have effectively isolated the idiosyncratic signatures of each subject. Together, these visualizations provide compelling evidence for our method's proposed mechanism.

5.2 Numerical Results

To validate our approach, we compare Subject-Conditioned Layer against several modeling strategies, with results summarized in Table 1. All numerical results are reported as the mean and standard deviation of three reproducible runs with random seeds $\{1, 2, 3\}$.

First, to establish the necessity of subject-specific adaptation, we compare against a **Subject-Agnostic Model**. This baseline is a standard architecture trained on data collected from all subjects, without

[4] http://bnci-horizon-2020.eu/database/data-sets

Table 1: Numerical Results. Performance comparison across model variants: a subject-agnostic model (no personalization), subject-specific models (trained independently per subject), subject-specific LoRA models (entirely low-rank per subject), and Subject-Conditioned Layer.

Method	Total # Params	Active # Params	BCI Comp. IV2a (4 Classes)	BCI Comp. IV2b (2 Classes)
EEGNeX Subject-Agnostic [6]	55 972	55 972	55.00% ± 0.66%	74.28% ± 0.39%
EEGNeX Subject-Specific	55 972	55 972	59.36% ± 14.42%	75.92% ± 14.32%
EEGNex Subject-Specific LoRA	81 124	10 980	53.60% ± 2.91%	69.64% ± 0.27%
EEGNeX Subject-Conditioned Layer	134 884	64 740	**64.72% ± 0.56%**	**76.48% ± 0.25%**
PBT Subject-Agnostic [17]	867 460	867 460	50.82% ± 0.85%	75.49% ± 1.10%
PBT Subject-Specific	867 460	867 460	44.16% ± 12.19%	75.63% ± 14.23%
PBT Subject-Specific LoRA	680 736	135 712	25.81% ± 0.52%	53.70% ± 0.80%
PBT Subject-Conditioned Layer	1 480 612	935 588	**54.51% ± 0.30%**	**76.36% ± 0.23%**

personalization. Our method surpasses this baseline, confirming that a one-size-fits-all approach is insufficient for high performance in EEG decoding, Table 1.

Next, we demonstrate the benefit of knowledge sharing by comparing our method against the common practice of training fully independent **Subject-Specific Models**. This baseline trains a separate model from scratch for each individual. In Table 1, we report the mean performance across all subjects (the complete results are provided in Table 4). Subject-Conditioned Layer consistently outperforms the average of those independently trained models. This finding highlights that the shared backbone (W_{general}) enables effective knowledge transfer across subjects, yielding better generalization than isolated training on limited single-subject datasets.

Finally, to validate our specific architectural design, we introduce a crucial parameter control: **Subject-Specific LoRA Models**. For each subject, a separate low-rank model is trained from scratch. This baseline allows us to test if the performance gain is simply an artifact of having a certain number of trainable parameters per subject. Subject-Conditioned Layer's superiority in this comparison is a critical finding. It proves that our hybrid architecture is more effective than training isolated, parameter-efficient models. This validates that the structure of our approach is important for its performance.

6 Limitations & Outlook

Our work focuses on introducing Subject-Conditioned Layer and empirically demonstrating its effectiveness on the BCI Competition IV 2a and 2b datasets. However, our study has some limitations, which suggest directions for future research.

A limitation is the framework's dependence on subject-specific data to train the personalized adapters, W_s. This means that for previously unseen subjects, predictions can only be made using the shared weights, W_{general}. Although we have shown that such personalization enhances performance, further analysis is required to fully assess the model's generalization capabilities and its viability for practical deployment. Subject-Conditioned Layer is well-suited for a pre-training and fine-tuning paradigm. One could first train the shared weights W_{general} across a large population. Subsequently, for a new subject, one could freeze the powerful shared model and efficiently fine-tune only a lightweight subject-specific adapter W_s. This iterative application, coupled with scaling, could lead to a Foundation model for EEG Decoding. To tackle the challenge of unseen subjects, future research could investigate the geometry of the learned adapter space. By comparing the adapters of different subjects (e.g., via their similarity), we could potentially discover subject clusters or archetypes. This could allow us to initialize a new subject adapter from a pre-computed group adapter, enabling faster personalization or even zero-shot prediction, thereby greatly enhancing the practical utility of our method.

Funding Acknowledgements

This project has received funding from the European Regional Development Fund (grants timingMatters and IntelAlgen) under the European Union's Horizon Europe Research and Innovation Program, from the German Research Foundation DFG within GRK 2297 'Mathematical Complexity Reduction', and from the German Federal Joint Committee (Grant 01VSF23017), as well as German Research Foundation DFG (Grant 537063406), which we gratefully acknowledge.

References

[1] B. Aristimunha, I. Carrara, P. Guetschel, S. Sedlar, P. Rodrigues, J. Sosulski, D. Narayanan, E. Bjareholt, Q. Barthelemy, R. T. Schirrmeister, R. Kobler, E. Kalunga, L. Darmet, C. Gregoire, A. Abdul Hussain, R. Gatti, V. Goncharenko, J. Thielen, T. Moreau, Y. Roy, V. Jayaram, A. Barachant, and S. Chevallier. Mother of all BCI Benchmarks, 2025.

[2] B. Aristimunha, D. Truong, P. Guetschel, S. Y. Shirazi, I. Guyon, A. R. Franco, M. P. Milham, A. Dotan, S. Makeig, A. Gramfort, J.-R. King, M.-C. Corsi, P. A. Valdés-Sosa, A. Majumdar, A. Evans, T. J. Sejnowski, O. Shriki, S. Chevallier, and A. Delorme. EEG Foundation Challenge: From Cross-Task to Cross-Subject EEG Decoding, 2025.

[3] P. Bomatter and H. Gouk. Is Limited Participant Diversity Impeding EEG-based Machine Learning? 2025.

[4] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller. BCI Competition 2008–Graz data set A. Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology, 16:1–6, 2008.

[5] R. Caruana. Multitask Learning. Machine Learning, 28(1):41–75, July 1997.

[6] X. Chen, X. Teng, H. Chen, Y. Pan, and P. Geyer. Toward reliable signals decoding for electroencephalogram: A benchmark study to EEGNeX. Biomedical Signal Processing and Control, 87:105475, Jan. 2024.

[7] S. Chevallier, I. Carrara, B. Aristimunha, P. Guetschel, S. Sedlar, B. Junqueira Lopes, S. Velut, S. Khazem, and T. Moreau. The largest EEG-based BCI reproducibility study for open science: the MOABB benchmark. working paper or preprint, Apr. 2024.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021.

[9] T. Evgeniou and M. Pontil. Regularized multi–task learning. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD04, page 109–117. ACM, Aug. 2004.

[10] W. Fedus, B. Zoph, and N. Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, 2022.

[11] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen. MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience, 7:267, 2013.

[12] P. Guetschel, S. Ahmadi, and M. Tangermann. Review of deep representation learning techniques for brain–computer interfaces. Journal of Neural Engineering, 21(6):061002, Nov. 2024.

[13] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training Compute-Optimal Large Language Models, 2022.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021.

[15] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive Mixtures of Local Experts. Neural Computation, 3(1):79–87, Feb. 1991.

[16] W.-B. Jiang, L.-M. Zhao, and B.-L. Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI, 2024.

[17] T. Klein, P. Minakowski, and S. Sager. Flexible Patched Brain Transformer model for EEG decoding. Scientific Reports, 15(1), Mar. 2025.

[18] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz. BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data. Frontiers in Human Neuroscience, 15, June 2021.

[19] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller. BCI Competition 2008–Graz data set B. Graz University of Technology, Austria, 16:1–6, 2008.

[20] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller. Brain–Computer Communication: Motivation, Aim, and Impact of Exploring a Virtual Apartment. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 15(4):473–482, 2007.

[21] M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. Biometrics, pages 673–687, 1990.

[22] J. C. Pinheiro and D. M. Bates. Mixed-effects models in S and S-PLUS. Springer, 2000.

[23] S. Ruder. An Overview of Multi-Task Learning in Deep Neural Networks. CoRR, abs/1706.05098, 2017.

[24] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping, aug 2017.

[25] P. Schmid, T. Klein, P. Minakowski, S. Sager, C. Reichert, R. T. Knight, and S. Dürschmid. Temporal kinetics of brain state effects on visual perception, Aug. 2024.

[26] S. Schotthöfer, E. Zangrando, J. Kusch, G. Ceruti, and F. Tudisco. Low-rank lottery tickets: finding efficient low-rank neural networks via matrix differential equations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 20051–20063. Curran Associates, Inc., 2022.

[27] S. Schotthöfer, T. Klein, and J. Kusch. A geometric framework for momentum-based optimizers for low-rank training, 2025.

[28] S. Schotthöfer and M. P. Laiu. Federated Dynamical Low-Rank Training with Global Loss Convergence Guarantees, 2024.

[29] S. Schotthöfer, H. L. Yang, and S. Schnake. Dynamical Low-Rank Compression of Neural Networks with Robustness under Adversarial Attacks, 2025.

[30] S. Schotthöfer, E. Zangrando, G. Ceruti, F. Tudisco, and J. Kusch. GeoLoRA: Geometric integration for parameter efficient fine-tuning, 2024.

[31] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer, 2017.

[32] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz. Review of the BCI competition IV. Frontiers in Neuroscience, 6, 2012.

[33] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(86):2579–2605, 2008.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need, 2023.

[35] X. Wei, A. A. Faisal, M. Grosse-Wentrup, A. Gramfort, S. Chevallier, V. Jayaram, C. Jeunet, S. Bakas, S. Ludwig, K. Barmpas, M. Bahri, Y. Panagakis, N. Laskaris, D. A. Adamos, S. Zafeiriou, W. C. Duong, S. M. Gordon, V. J. Lawhern, M. Śliwowski, V. Rouanne, and P. Tempczyk. 2021 BEETL Competition: Advancing Transfer Learning for Subject Independence & Heterogenous EEG Data Sets, 2022.

[36] E. Zangrando, S. Schotthöfer, G. Ceruti, J. Kusch, and F. Tudisco. Geometry-aware training of factorized layers in tensor Tucker format. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 129743–129773. Curran Associates, Inc., 2024.

[37] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling Vision Transformers, 2022.

[38] Y. Zhang and Q. Yang. A Survey on Multi-Task Learning, 2021.

A Related Work

Our proposed method addresses the challenge of inter-subject variability in EEG decoding by learning both shared and subject-specific representations. This approach is closely related to established paradigms nonlinear mixed-effects (NLME) modeling and in machine learning, primarily Multi-Task Learning (MTL) and Mixture of Experts (MoE).

Nonlinear Mixed-Effects Modeling. Our approach can be viewed through the lens of nonlinear mixed-effects (NLME) models [21, 22]. In this framing, the shared network W_{general} learns the fixed, population-level effects, while the subject-specific low-rank adapters W_s capture the random effects, parsimonious deviations for each individual. The low-rank constraint on the adapters is analogous to the low-dimensional random effects assumption in classical NLME, which encourages partial pooling of information and borrows statistical strength across subjects. This perspective provides a principled justification for using LoRA [14] as a parameter-efficient method to model subject-specific variations, which is crucial when per-subject data is limited.

Multi-Task Learning. In MTL, a model is trained on several related tasks simultaneously, with the goal of improving performance by leveraging shared information across tasks [5, 23, 9]. Viewing the decoding for each subject as a distinct task, our model fits directly into the MTL framework. It employs a form of hard parameter sharing, where the general weight matrix, W_{general}, learns a common feature space shared across all subjects. The subject-specific weights, W_s, act as task-specific layers that adapt this shared representation to the unique neural patterns of each individual. This architecture allows the model to learn robust, generalizable features from the entire population while still retaining the flexibility to specialize for each subject [38].

Mixture of Experts. MoE models utilize multiple expert networks and a gating mechanism that determines which expert processes a given input [15]. Sparsely-gated MoE models, in particular, activate a subset of experts for each input, enabling the scaling of model capacity without a proportional increase in computation [31, 10]. Our subject-aware layer can be interpreted as a specialized, non-learned MoE. Here, the subject identity serves as the gating signal, deterministically routing the input to its corresponding expert, the pathway defined by W_s. Unlike traditional MoE, our gating is not a learned function but is provided by the input via the mask M. This simplifies the model by removing the need to learn a routing policy and directly addresses the known source of variation (i.e., subject identity).

Dynamical low-rank. Since our method explicitly models subject-specific variability as lightweight low-rank corrections, it naturally connects to the broader line of dynamical low-rank research [26, 28, 29]. An exciting direction for future work is to explore whether subject adaptation in EEG can benefit from recent advances in geometry-aware training [36, 30], and robust compression strategies [29], thereby unifying personalization with efficient large-scale deployment.

B Extension to Convolutional Layers

While the formulation in Equation 4 is presented for linear transformations, it can be seamlessly extended to convolutional layers. A convolution operation is itself a linear transformation, where the weight matrix W is replaced by a convolutional kernel W. The Subject-Conditioned convolutional layer can thus be expressed as:

$$\bar{X} = \sigma\left(X * \mathcal{W}_{\text{general}} + \textstyle\sum_{s=1}^{N} M_s \cdot (X * \mathcal{W}_s)\right), \tag{6}$$

where $*$ denotes the convolution operation.

To implement the low-rank adaptation (Section 3.2) for the subject-specific kernels, we decompose the operation $X * W_s$ into two sequential convolutions, effectively factorizing the kernel itself. This is analogous to the matrix decomposition in Equation 5. Specifically, the subject-specific update is computed by:

$$\mathcal{W}_s := A_s B_s, \tag{7}$$

where A_s and B_s are the low-rank convolutional kernels. The first kernel, A_s, reduces the channel dimension from C_{in} to the rank r while preserving the spatial kernel size. The second kernel, B_s, is a 1×1 convolution that projects the features from the rank r space back to the C_{out} channel dimension.

Initialization is equal to the linear case, see Section 3.1. This convolutional decomposition allows for parameter-efficient adaptation for each subject while maintaining the inductive biases of convolutional layers.

B.1 Visualization of Latent Space Correspondence

In Section 5.1, we claim that our full model learns to form meaningful sub-clusters that correspond to the downstream task's target labels. Figure 3 provides visual evidence supporting this claim by extending Figure 1.

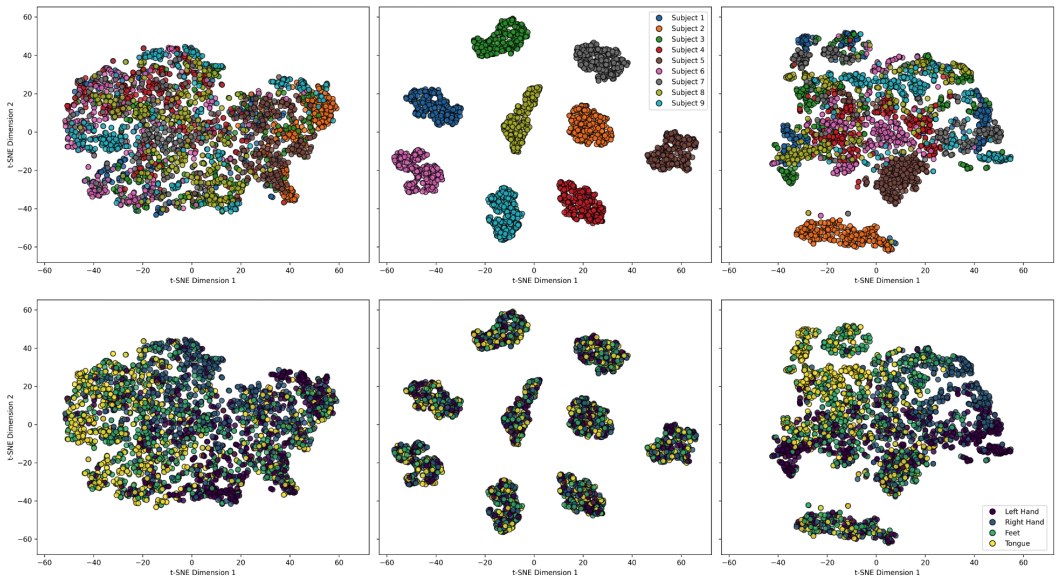

Figure 3: Latent Representations Colored by Subject ID vs. Target Labels. This figure is an extension of Figure 1. The top row repeats the t-SNE projections of representations from the shared weights W_{general}, subject-specific weights W_s, and the full model, with all points colored by subject ID. The bottom row displays the exact same data points, but colored by their ground-truth target labels. This comparison explicitly shows that the sub-clusters formed by our full model (top right) directly align with the target classes (bottom right), confirming that the model learns discriminative, task-relevant features within each subject's representation space.

C Numerical Experiments

C.1 Experiment Design

EEGNeX

We used the EEGNeX implementation from the Braindecode package [11, 24]. In this model, we replaced all standard convolutional layers with the convolutional version of our method, Subject-Conditioned Layer (see Appendix B). We made an exception for the depthwise convolutional layer (DepthwiseConv2D) in Block 3 due to implementation details.

As the original hyperparameters for EEGNeX were not specified, we tuned the baseline learning rate by sweeping over the values $\{10^{-3},\ 10^{-4}\}$. We want to remark that the difference between the numbers reported in Table 1 and those in the original EEGNeX paper stems from different evaluation methods. Here, we report the numbers for cross-session analysis from a 512-timepoint long interval.

The final hyperparameters are reported in Table 2:

Table 2: Hyperparameters EEGNeX.

Hyperparameter	Value
Band-pass filter	[8, 45]
Optimizer	AdamW
Learning rate	1×10^{-3}
Weight decay	0.01
AdamW betas	(0.9, 0.999)
Batch size	64
Training epochs	100
Dropout rate	0.0
Activation function	ELU
Rank adapters	4
LoRA α	1

Patched Brain Transformer

For the PBT, we used the official GitHub repository accompanying the publication and replaced all linear layers with our proposed method, Subject-Conditioned Layer. The only exception was for the visualization in Figure 1 and Figure 3, where we kept the final linear layer to improve clarity.

We used the same hyperparameters as reported in the original PBT paper for all runs (Table 3). For our method's specific hyperparameters, we set the rank $r = 8$ and $\alpha = 1$.

Table 3: Hyperparameters PBT.

Hyperparameter	Value
Band-pass filter	[8, 45]
Optimizer	AdamW
Learning rate	3×10^{-4}
Weight decay	0.01
AdamW betas	(0.9, 0.95)
Batch size	128
Training epochs	1200
Warmup epochs	150
Dropout rate	0.1
Activation function	GELU
Rank adapters	8
LoRA α	1
Data augmentation	Time shifts

D Detailed Numerical Results for Subject-Specific Modeling

Table 4: Subject-Specific Numerical results for EEGNeX and PBT on individual subjects.

Subject	**EEGNeX**		**PBT**	
	BCI Comp. IV2a	BCI Comp. IV2b	BCI Comp. IV2a	BCI Comp. IV2b
1	67.82% ± 6.25%	72.50% ± 1.84%	56.60 % ± 1.85%	71.35% ± 1.28%
2	46.53% ± 1.70%	52.86% ± 0.77%	25.23 % ± 1.31 %	53.10% ± 1.02%
3	77.08% ± 3.27%	51.56% ± 1.99%	55.90 % ± 3.07 %	50.73% ± 0.15%
4	53.01% ± 1.80%	94.27% ± 0.53%	39.00 % ± 1.15 %	95.31% ± 1.11%
5	36.69% ± 3.38%	86.25% ± 1.35%	26.97 % ± 0.59 %	86.98% ± 1.62%
6	40.97% ± 1.58%	81.77% ± 2.32%	34.72 % ± 2.79 %	77.81% ± 1.02%
7	67.94% ± 2.31%	70.94% ± 2.04%	47.80 % ± 7.45 %	78.52% ± 0.72%
8	68.29% ± 0.71%	86.77% ± 1.49%	55.79 % ± 3.38 %	86.46% ± 0.64%
9	75.93% ± 0.43%	83.85% ± 1.98%	55.44 % ± 0.59 %	83.02% ± 1.64%

E Compute Resources

The shown experiments are performed on a single NVIDIA Tesla V100 GPU using PyTorch. Code is publicly available at github.com/timonkl/SubjectConditionedLayer.