# Subject-Aware Contrastive Learning for EEG Foundation Models

**Antonis Karantonis** *,1,4,5
akarantonis@di.uoa.gr

**Konstantinos Barmpas** *,2,5
konstantinos.barmpas16@imperial.ac.uk

**Dimitrios A. Adamos** 2,5

**Nikolaos Laskaris** 3,5 **Stefanos Zafeiriou** 2,4,5 **Yannis Panagakis** 1,4,5

[1]National and Kapodistrian University of Athens [2]Imperial College London
[3]Aristotle University of Thessaloniki [4]Archimedes/Athena RC, Greece
[5]Cogitat Ltd., London, U.K.

## Abstract

Foundation models are beginning to reshape EEG representation learning, but existing approaches remain dominated by self-supervised reconstruction objectives. In this work, we introduce the first subject-aware contrastive EEG foundation model, leveraging subject identity as a natural supervisory signal. Building on a patch-based architecture inspired by recent Large Brainwave Foundation Models (LBMs), we pretrain a lightweight transformer encoder using contrastive learning, where positive pairs are drawn from different segments and sessions of the same subject. Unlike contrastive foundation models in other domains, which depend on augmentations to construct positive samples, our method relies on naturally occurring intra-subject variability across EEG sessions. We evaluate the model through both representation metrics (alignment, uniformity and smooth effective rank) and downstream tasks (under linear probing and full fine-tuning). Results show that our model produces well-structured representation spaces, achieving strong representation quality and competitive performance compared to other LBMs.

## 1 Introduction

Brain-Computer Interfaces (BCIs) aim to provide a direct pathway between the human brain and external devices, enabling applications ranging from assistive technologies to cognitive state monitoring. The underlying signals can be recorded through electroencephalography (EEG), a non-invasive technique that captures the brain's oscillatory activity with high temporal resolution, Niedermeyer & da Silva (2004). Early approaches to EEG-based BCIs relied heavily on hand-crafted features designed from neuroscience insights, such as power spectral densities and bandpower ratios Bashashati et al. (2007); Handy (2009); Rao (2013); McFarland et al. (2006). While these methods provided initial progress, they often failed to generalize across subjects due to the strong inter-subject variability inherent in EEG signals, Barmpas et al. (2024), where individual anatomical and physiological differences significantly affect observed neural patterns Jayaram & Barachant (2013); Barmpas et al. (2023b).

With the advent of deep learning, the field shifted towards end-to-end data-driven feature extraction. Models such as convolutional neural networks were shown to learn highly discriminative spatio-temporal representations of EEG signals Lawhern et al. (2018); Santamaría-Vázquez et al. (2020); Song et al. (2023); Barmpas et al. (2023a). This reduced dependence on domain-specific feature engineering enabled significant performance improvements across diverse paradigms, including

Learning from Time Series for Health NeurIPS 2025 Workshop.

motor imagery, event-related potentials and cognitive workload estimation. However, despite these advances, deep learning models in EEG often demand large amounts of labeled data and are typically trained on specific tasks or paradigms. This reliance on supervision restricts their generalizability and makes deployment resource-intensive in new contexts.

In parallel, the rise of foundation models in language, vision, and speech Brown et al. (2020); Touvron et al. (2023); Baevski et al. (2020) has introduced a new paradigm: large-scale self-supervised pretraining on heterogeneous unlabeled data. These models acquire general representations that transfer broadly, reducing the need for extensive bespoke training on every new task. Inspired by this trend, researchers have begun to explore Large Brainwave Models (LBMs) for EEG decoding. Early efforts include BIOT Yang et al. (2023), EEGPT Wang et al. (2024), and CBraMod Wang et al. (2025), which applied transformers and self-supervised training across multiple EEG datasets. Jiang et al. (2024) (and more recently Barmpas et al. (2025)) used a combination of patch-based and codebook-based tokenization along with masked modeling objective, demonstrating the potential of scalable unified EEG foundation models. Most existing LBMs rely on reconstruction-style pretraining, aiming to recover masked or transformed signal patches. Yet contrastive learning, which has proven highly effective in vision and other biosignal domains, has not been systematically explored for EEG.

In this work, we introduce the first subject-aware contrastive EEG foundation model, trained with Normalized Temperature-scaled Cross Entropy (NT-Xent) loss Chen et al. (2020) on patch representations. Our approach leverages subject identity as a natural supervisory signal and yields well-structured representation spaces with strong alignment, uniformity and smooth effective rank, while remaining competitive with prior LBM approaches.

# 2 Background

Large-scale self-supervised pretraining has recently been extended beyond language and vision to biosignals. Abbaspourazad et al. (2023) demonstrated that contrastive learning on massive unlabeled ECG and PPG recordings can produce embeddings that encode subject-level physiology and demographic information. Their evaluation emphasized not only downstream accuracy but also representation metrics such as alignment, uniformity and smooth effective rank, highlighting the importance of evaluating latent space structure in biosignal foundation models. Inspired by this work, we investigate whether contrastive learning can also serve as a viable path for LBMs.

To adapt EEG for large-scale pretraining, we adopt a patch-based tokenization scheme similar to that introduced in Jiang et al. (2024) and Barmpas et al. (2025). In this formulation, raw EEG signals are segmented into fixed-length temporal patches across channels, which are then embedded using temporal convolutions and enriched with spatial and temporal embeddings before being passed to a transformer encoder. This design provides standardized, sequence-like inputs that facilitate scalable pretraining across heterogeneous EEG datasets.

Building on these insights, our work combines the contrastive pretraining philosophy of Abbaspourazad et al. (2023) with the patch-based representation strategy of Jiang et al. (2024) and Barmpas et al. (2025), introducing the first Subject-Aware Contrastive Brainwave Foundation Model.

# 3 Model Architecture

## 3.1 Patch-Based Representation

Let $X \in \mathbb{R}^{C \times T}$ denote the input EEG signal, where $T$ is the number of time points and $C$ is the number of electrodes. Similar to Jiang et al. (2024) and Barmpas et al. (2025), the signal is first segmented into temporal patches. To ensure that the models can deal with EEG signals of variable channels and time durations, the following approach is utilized: during model pre-training, each input is represented by $P$ patches of length $w$ (corresponding to a 1-second window), while only the number of channels is allowed to vary. This results in a segmented input sample of $P$ patches (i.e, $\mathbf{x} \in \mathbb{R}^{P \times w}$). These patches undergo embedding via temporal convolutions, enriched with spatial and temporal embeddings, and are subsequently processed by a transformer encoder. This setup provides flexibility to handle heterogeneous EEG recordings but always includes a fixed number of patches, ensuring consistent input length for the encoder.

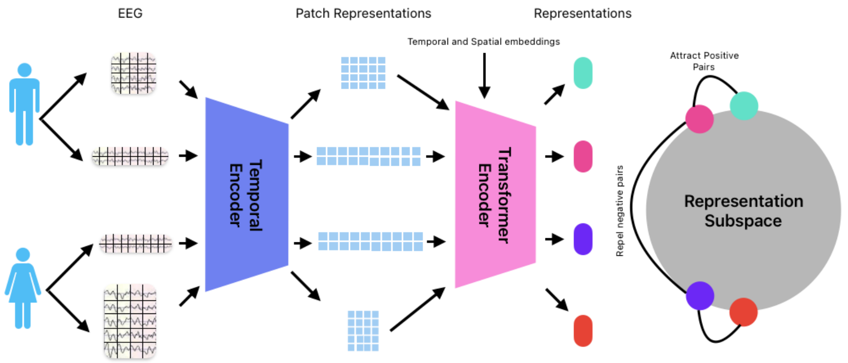

Figure 1: Overview of our subject-aware contrastive EEG foundation model. EEG signals are segmented into patches, embedded via a temporal encoder, enriched with spatial and temporal embeddings, and processed through a transformer encoder. Contrastive learning attracts positive pairs from the same subject and repels negatives from different subjects.

### 3.2 Positive Pairs Selection

A critical design decision concerns the construction of positive pairs for contrastive learning. We draw positives not only from different segments within the same recording, but also from recordings of the same subject across sessions, including potentially different days. This naturally occurring variability functions similarly to augmentations in our setting, allowing the model to learn invariances related to session-level fluctuations and more stable subject characteristics. Studies confirm that intra-subject, inter-session EEG variability (i.e., from recordings of the same person on different days) is substantial and therefore valuable for building robust models Huang et al. (2023). We use samples from different subjects as negatives, structuring the latent space around subject identity. Unlike common contrastive pipelines, we do not employ synthetic augmentations.

## 4 Experiments

### 4.1 Training and Evaluation

We trained our model following the details described in Appendix A, with architecture ablations reported in Appendix B, and using the datasets of Appendix C. To assess the quality of the learned representations, we adopt both intrinsic contrastive metrics and extrinsic downstream performance:

**Contrastive metrics:** We report alignment and uniformity Wang & Isola (2020), as well as Smooth Effective Rank (SER) as introduced in Abbaspourazad et al. (2023). These metrics provide insight into whether the representation space is well-structured, offering a good indication of potential downstream task performance even before linear probing or fine-tuning.

**Downstream tasks:** Similar to Lee et al. (2025b), we evaluate on supervised EEG benchmarks under two protocols. In the linear probing setup, a frozen foundation model is paired with a shallow classifier, highlighting the linear separability of the representations. In the fine-tuning setup, the foundation model and classification head are both updated for each task. This dual evaluation reflects both the immediate usability of pretrained representations and their adaptability to specific downstream applications. For downstream performance, the models were evaluated in downstream classification tasks for the following four EEG datasets as described in the benchmark Lee et al. (2025a): Motor paradigm in High Gamma Schirrmeister et al. (2017), a Working Memory dataset Pavlov et al. (2022), Sleep-EDF Kemp et al. (2000) and Eyes Open vs Closed classification on the Physionet Motor dataset Schalk et al. (2004).

### 4.2 Downstream Task Performance

**Linear probing.** We compared our model against an open-source model that uses similar patch representation, namely Jiang et al. (2024). As shown in Table 1, our model achieves the best overall mean performance across tasks, outperforming LaBraM Jiang et al. (2024). These results indicate that the representations learned through subject-aware contrastive pretraining have strong separability, a sign of strong foundation model features. The fact that our lightweight 2-layer encoder surpasses a larger reconstruction-oriented model in this setting highlights the promise of contrastive training for EEG representation learning.

Table 1: Linear probing results: classification accuracy of logistic regression model trained on latent features. Bold and underlined values indicate best and second-best performance respectively (per task or overall).

| Model | Motor | Memory | Sleep | Eyes | Mean |
|---|---|---|---|---|---|
| LaBraM | 0.297 | **0.670** | 0.608 | 0.717 | 0.573 |
| Ours (contrastive) | **0.360** | 0.580 | **0.610** | **0.810** | **0.590** |

**Fine-tuning.** In the full fine-tuning setup, our model delivers competitive performance compared to other open-source state-of-the-art LBMs, namely BIOT Yang et al. (2023)[*], EEGPT Wang et al. (2024) and CBraMod Wang et al. (2025). As shown in Table 2, it achieves the best performance on Eyes and strong results across other tasks. While these numbers demonstrate that the learned representations already transfer reasonably well, there is clear headroom for improvement. We expect that scaling pretraining to larger and more diverse datasets, along with exploring parameter-efficient or task-specific fine-tuning strategies, can help convert the strong representation quality observed in linear probing into even stronger fine-tuned performance.

Table 2: Fine-tuning results (accuracy). Bold and underlined values indicate best and second-best performance respectively (per task or overall).

| Model | Motor | Memory | Sleep | Eyes | Mean |
|---|---|---|---|---|---|
| EEGPT (encoder) | 0.313 | 0.520 | 0.633 | 0.793 | 0.565 |
| CBraMod | **0.614** | **0.574** | **0.635** | 0.839 | **0.666** |
| BIOT | 0.443 | 0.510 | – | 0.763 | 0.572 |
| Ours (contrastive) | 0.450 | 0.560 | 0.610 | **0.840** | 0.615 |

The findings shown in Table 2 highlight that our subject-aware contrastive objective introduces a promising new direction for EEG foundation models. Using a considerably lighter architecture, our model achieves competitive, and in some cases superior, performance compared to existing approaches. As the first contrastive formulation that explicitly incorporates subject identity, it establishes a strong and efficient baseline with clear potential for further advancement through architectural extensions and the exploration of additional contrastive objectives.

## 5 Conclusion

In this work we introduced the Subject-Aware Contrastive Brainwave Foundation Model, combining a patch-based representation with NT-Xent loss to structure the latent space around subject identity. Our experiments show that contrastive pretraining produces well-organized embeddings, with strong SER, rank, and contrastive uniformity/alignment scores. These results highlight the promise of contrastive learning as a complementary paradigm to reconstruction-oriented EEG foundation models.

Our approach could also be extended beyond subject-specific supervision to incorporate task-specific goals, aligning the learned representations more directly with BCI applications. Furthermore, future work will also require scaling to larger and more diverse datasets and exploring multi-modal extensions that combine EEG with other physiological signals to enrich representation quality.

[*] BIOT could not be tested on Sleep since the benchmark electrodes are missing from the pre-trained model

## Acknowledgements

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility.

## References

Abbaspourazad, S., Elachqar, O., Miller, A. C., Emrani, S., Nallasamy, U. & Shapiro, I. (2023) Large-scale training of foundation models for wearable biosignals. doi:10.48550/arXiv.2312.05409. ICLR 2024 camera-ready (v2, 2024-03-06).

Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*.

Barmpas, K., Lee, N., Panagakis, Y., Adamos, D., Laskaris, N. & Zafeiriou, S. (2025) Advancing brainwave modeling with a codebook-based foundation model. *arXiv preprint arXiv:2505.16724*.

Barmpas, K., Panagakis, Y., Adamos, D. A., Laskaris, N. & Zafeiriou, S. (2023a) Brainwave-scattering net: A lightweight network for EEG-based motor imagery recognition. *Journal of Neural Engineering*, 20(5):056014.

Barmpas, K., Panagakis, Y., Bakas, S., Adamos, D. A., Laskaris, N. & Zafeiriou, S. (2023b) Improving generalization of CNN-based motor-imagery EEG decoders via dynamic convolutions. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1997–2005.

Barmpas, K., Panagakis, Y., Zoumpourlis, G., Adamos, D. A., Laskaris, N. & Zafeiriou, S. (2024) A causal perspective on brainwave modeling for brain–computer interfaces. *Journal of Neural Engineering*, 21(3):036001.

Bashashati, A., Fatourechi, M., Ward, R. K. & Birch, G. E. (2007) A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural Engineering*, 4(2):R32–R57. doi:10.1088/1741-2560/4/2/R03.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R. & Curio, G. (2007) The non-invasive berlin brain–computer interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550. doi:10.1016/j.neuroimage.2007.01.051.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020) Language models are few-shot learners.

Buckwalter, G., Chhin, S., Rahman, S., Obeid, I. & Picone, J. (2021) Recent advances in the TUH EEG corpus: Improving the interrater agreement for artifacts and epileptiform events. In *2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–3. doi: 10.1109/SPMB52430.2021.9672302.

Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020) A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*.

Detti, P., Vatti, G. & Zabalo Manrqiue de Lara, G. (2020) EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8:846. doi:10.3390/pr8070846.

Handy, T. C. (2009) *Brain Signal Analysis: Advances in Neuroelectric and Neuromagnetic Methods*. MIT Press. ISBN 9780262013086. doi:10.7551/mitpress/9780262013086.001.0001.

Huang, G., Zhao, Z., Zhang, S., Hu, Z., Fan, J., Fu, M., Chen, J., Xiao, Y., Wang, J. & Dan, G. (2023) Discrepancy between inter- and intra-subject variability in EEG-based motor imagery brain–computer interface: Evidence from multiple perspectives. *Frontiers in Neuroscience*, 17:1122661. doi:10.3389/fnins.2023.1122661.

Jayaram, V. & Barachant, A. (2013) EEG-based brain–computer interfaces and the issue of inter-subject variability: a review. *Frontiers in Human Neuroscience*, 7:00042.

Jiang, W.-B., Zhao, L.-M. & Lu, B.-L. (2024) Large brain model for learning generic representations with tremendous EEG data in BCI.

Kemp, B., Zwinderman, A. H., Tuk, B., Kamphuisen, H. A. C. & Oberye, J. J. L. (2000) Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194. doi:10.1109/10.867928.

Korczowski, L., Cederhout, M., Andreev, A., Cattan, G., Rodrigues, P., Gautheret, V. & Congedo, M. (2019) Brain invaders calibration-less P300-based BCI with modulation of flash duration dataset (bi2015a). doi:10.5281/zenodo.3266930.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P. & Lance, B. J. (2018) EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013. doi:10.1088/1741-2552/aace8c.

Lee, N., Bakas, S., Barmpas, K., Panagakis, Y., Adamos, D., Laskaris, N. & Zafeiriou, S. (2025a) Assessing the capabilities of large brainwave foundation models. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*.

Lee, N., Barmpas, K., Panagakis, Y., Adamos, D., Laskaris, N. & Zafeiriou, S. (2025b) Are large brainwave foundation models capable yet? insights from fine-tuning.

Luciw, M., Jarocka, E. & Edin, B. (2014) Multi-channel EEG recordings during 3,936 grasp and lift trials with varying weight and friction. *Scientific Data*, 1:140047. doi:10.1038/sdata.2014.47.

Margaux, P., Maby, E., Daligault, S., Bertrand, O. & Mattout, J. (2012) Objective and subjective evaluation of online error correction during P300-based spelling. *Advances in Human-Computer Interaction*, 2012. doi:10.1155/2012/578295.

McFarland, D. J., Anderson, C. W., Muller, K.-R., Schlogl, A. & Krusienski, D. J. (2006) BCI meeting 2005–workshop on BCI signal processing: Feature extraction and translation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):135–138. doi:10.1109/TNSRE.2006.875637.

Niedermeyer, E. & da Silva, F. L. (2004) *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins, 5th edn.

Pavlov, Y. G., Kasanov, D., Kosachenko, A. I. & Kotyusov, A. I. (2022) EEG, pupillometry, ECG and photoplethysmography, and behavioral data in the digit span task and rest. doi:10.18112/openneuro.ds003838.v1.0.2.

Rao, R. P. N. (2013) *Brain-computer interfacing: An introduction*. USA: Cambridge University Press. ISBN 0521769418.

Santamaría-Vázquez, E., Martínez-Cagigal, V., Vaquerizo-Villar, F. & Hornero, R. (2020) EEG-inception: A novel deep convolutional neural network for assistive ERP-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):2773–2782. doi:10.1109/TNSRE.2020.3048106.

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N. & Wolpaw, J. R. (2004) BCI2000: a general-purpose brain–computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W. & Ball, T. (2017) Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420.

Shah, V., von Weltin, E., Lopez, S., McHugh, J. R., Veloso, L., Golmohammadi, M., Obeid, I. & Picone, J. (2018) The temple university hospital seizure detection corpus.

Song, Y., Zheng, Q., Liu, B. & Gao, X. (2023) EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719. doi:10.1109/TNSRE.2022.3230250.

Torkamani-Azar, M., Kanik, S. D., Aydin, S. & Cetin, M. (2019) Prediction of reaction time and vigilance variability from spatiospectral features of resting-state EEG in a long sustained attention task.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. & Lample, G. (2023) LLaMA: Open and efficient foundation language models.

Trujillo, L. (2020) Raw EEG data. doi:10.18738/T8/SS2NHB.

Trujillo, L. T., Stanfield, C. T. & Vela, R. D. (2017) The effect of electroencephalogram (EEG) reference choice on information-theoretic measures of the complexity and integration of EEG signals. *Frontiers in Neuroscience*, 11. doi:10.3389/fnins.2017.00425.

Veloso, L., McHugh, J., von Weltin, E., Lopez, S., Obeid, I. & Picone, J. (2017) Big data resources for EEGs: Enabling deep learning research. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–3. doi:10.1109/SPMB.2017.8257044.

Wang, G., Liu, W., He, Y., Xu, C., Ma, L. & Li, H. (2024) EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak & C. Zhang, eds., *Advances in Neural Information Processing Systems*, vol. 37, pp. 39249–39280. Curran Associates, Inc.

Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T. & Pan, G. (2025) CBramod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*.

Wang, T. & Isola, P. (2020) Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning*.

Yang, C., Westover, M. & Sun, J. (2023) BIOT: Biosignal transformer for cross-data learning in the wild. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine, eds., *Advances in Neural Information Processing Systems*, vol. 36, pp. 78240–78260. Curran Associates, Inc.

# A Model Configuration and Hyperparameter Settings

Our foundation model was trained using the NT-Xent loss Chen et al. (2020), a widely used contrastive objective. Positives are defined as segments from the same subject (either within the same recording or across different recordings), while negatives are drawn from other subjects. This loss $\mathcal{L}$ encourages embeddings of the same subject to be closer in the latent space, while pushing apart embeddings from different subjects, thereby structuring the representation space around subject identity. Mathematically, $\mathcal{L}$ is defined as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \tag{1}$$

where $\mathbf{z}_i$ is the patch representation for subject $i$.

Training was conducted using mini-batches sampled across subjects to ensure a balanced distribution of positive and negative pairs. Unlike typical contrastive learning pipelines, we did not employ synthetic data augmentations; instead, natural session-level variability in EEG recordings across different days and conditions served as implicit augmentations Huang et al. (2023).

All model configurations are described in detail in the following tables:

Table 3: Configuration of temporal encoder module.

| | Layer | Shape | Kernel | Stride | Padding | Norm(N, C) | Activation |
|---|---|---|---|---|---|---|---|
| Patch Embedding | Conv2d | (1, 8) | (1, 15) | (1, 8) | (0, 7) | GroupNorm(4, 8) | GELU |
| | Conv2d | (8, 8) | (1, 3) | (1, 1) | (0, 1) | GroupNorm(4, 8) | GELU |
| | Conv2d | (8, 8) | (1, 3) | (1, 1) | (0, 1) | GroupNorm(4, 8) | GELU |

Table 4: Hyperparameters for pre-training core foundation model and finetuning on downstream tasks.

| Hyperparameter | Pre-training FM | Finetuning |
|---|---|---|
| Batch size | 256 | 32 |
| Learning rate scheduler | Cosine | Linear |
| Base learning rate | 5e-4 | 5e-4 |
| Min learning rate | 1e-5 | - |
| Total epochs | 40 | 20 |
| Warmup epochs | 4 | 4 |
| Optimizer | AdamW | AdamW |
| Weight decay | 1e-4 | 0.01 |
| Adam $\beta$ | (0.9, 0.999) | (0.9, 0.999) |
| Layer lr decay | - | 0.975 |
| Layer scale init | 0.001 | - |
| Encoder depth | 2 | 2 |
| Hidden dimension | 200 | 200 |
| No. Attention heads | 10 | 10 |
| MLP hidden dimension | 256 | 256 |

## B Architecture Ablations

As described in Section 3, our network consists of a transformer-based encoder. To find the optimal number of layers, we resorted to evaluating contrastive metrics and experimenting with transformer layers of multiple depths. While deeper models tended to overfit, resulting in less stable representation metrics and weaker downstream performance, the shallowest models lacked sufficient capacity. A **2-layer transformer** struck the optimal balance, yielding 1.) consistent and stable contrastive training dynamics and 2.) strong downstream representation quality in a compact efficient architecture.

Table 5: Representation-level metrics across transformer depths and average of linear probing results.

| Layers | Alignment | Uniformity | SER | Average |
|---|---|---|---|---|
| 1 | 0.60 | $-3.48$ | 90.23 | 0.590 |
| 2 | 0.51 | $-3.52$ | **95.03** | **0.594** |
| 8 | 0.48 | $-3.51$ | 80.24 | 0.592 |

## C Datasets

Table 6: Datasets used during the contrastive foundation model's pre-training

| Dataset Names |
|---|
| BCI Competition IV-1 Blankertz et al. (2007) |
| Grasp and Lift Luciw et al. (2014) |
| Inria BCI Challenge Margaux et al. (2012) |
| Physionet MI Schalk et al. (2004) |
| Trujillo 2020 Trujillo (2020) |
| Trujillo 2017 Trujillo et al. (2017) |
| Siena Scalp Detti et al. (2020) |
| SPIS Resting Torkamani-Azar et al. (2019) |
| bi2015a Korczowski et al. (2019) |
| TUAR Buckwalter et al. (2021) |
| TUEP Veloso et al. (2017) |
| TUSZ Shah et al. (2018) |