UniNeuro: A Unified EEG Foundation Model for Heterogeneous EEG Task Learning

Haiyang Lu
luhaiyang@tongji.edu.cn
Tongji University
School of Computer Science and Technology
Shanghai, China

Lianghua He✉
helianghua@tongji.edu.cn
Shanghai Eye Disease Prevention and Treatment Center
Shanghai, China
Tongji University
School of Computer Science and Technology
Shanghai, China

Hongzhou Chen
chenhongzhou@tongji.edu.cn
Tongji University
School of Computer Science and Technology
Shanghai, China

Xiao Chen
2432123@tongji.edu.cn
Tongji University
School of Computer Science and Technology
Shanghai, China

Wenqi Zhang
2534183@tongji.edu.cn
Tongji University
School of Computer Science and Technology
Shanghai, China

Abstract

Electroencephalography (EEG), as a non-invasive physiological signal, plays an increasingly important role in human-centered multimedia systems, supporting applications ranging from emotion recognition and cognitive load assessment to motor-imagery-based control. However, due to the substantial differences in temporal scales, neural patterns, and label semantics across EEG tasks, it remains challenging to unify heterogeneous downstream tasks within a single model, while reconciling shared and task-specific EEG features. To this end, we propose **UniNeuro**, a unified EEG foundation framework that seamlessly integrates heterogeneous-task pretraining and downstream adaptation into a single discriminative learning pipeline. Within the backbone, we introduce a **task-conditional Mixture-of-Experts (MoE) with gradient reversal**, enforcing shared experts to capture cross-task features while dynamically routed experts focus on task-specific features, naturally decoupling the two types of representations. At the downstream adaptation stage, we propose **Hierarchical Prototype-Guided Contrastive Learning (HPGCL)** to explicitly structure the latent space at both task and label levels, enhancing task-level clustering and within-task class separability. Pretrained on over 10,000 hours of EEG data and evaluated on six diverse tasks, UniNeuro shows strong unified multi-task performance and narrows the gap with single-task fine-tuning, offering a scalable "brain-modality" encoder for next-generation interactive multimedia systems.

MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836038

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Neural networks**; *Learning latent representations*; • **Human-centered computing** → HCI theory, concepts and models.

Keywords

EEG foundation model, EEG signal, Mixture-of-Experts, Contrastive Learning

ACM Reference Format:
Haiyang Lu, Lianghua He, Hongzhou Chen, Xiao Chen, and Wenqi Zhang. 2026. UniNeuro: A Unified EEG Foundation Model for Heterogeneous EEG Task Learning. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3836038

1 Introduction

Electroencephalography (EEG) has emerged as a pivotal human-centric sensing modality in multimedia systems[26][6]. To support applications such as immersive VR interaction and wellness assessment, these systems must often simultaneously decode a spectrum of user states—ranging from emotional fluctuations and psychiatric indicators to interaction intents [24][34][14]. However, these tasks vary markedly in temporal scales, neural patterns, and label semantics.

To build a shared foundation model supporting such heterogeneous tasks, EEG research has shifted toward large-scale pretraining to learn transferable neural representations[40][16][35]. Despite significant progress, existing foundation models typically rely on task-specific fine-tuning, which incurs prohibitive computational overhead and inhibits real-time deployment. Therefore, a core challenge remains: how to achieve unified learning across heterogeneous downstream tasks within a single model. Recent efforts, such as NeuroLM [15], have attempted to address this by reformulating

heterogeneous EEG classification into a unified autoregressive prediction paradigm through EEG-language alignment and instruction tuning. While this represents a pivotal step toward general-purpose EEG modeling, the generative paradigm often lacks the efficiency and precision required for the closed-set classification benchmarks that dominate the EEG field[42][13]. This raises a critical question: can we directly unify multiple heterogeneous EEG tasks into a single, efficient discriminative framework?

In this end, we propose **UniNeuro**, a unified EEG foundation framework built around two novel components. First, at the backbone level, we introduce a gradient-reversal-enhanced task-conditional mixture-of-experts (MoE) design[7][3][22]. Always-activated shared experts are constrained via gradient reversal to focus on task-shared EEG features, while dynamically routed experts specialize for task-specific features, naturally decoupling shared and task-specific representations. Second, at the downstream level, we organize the learned feature space through Hierarchical Prototype-Guided Contrastive Learning, where supervised contrastive objectives[18][20][13]are applied at both the task level and the label level to encourage coarse-grained task structure and within-task class separability. Together, these two mechanisms support unified discriminative learning over heterogeneous EEG tasks within a single architecture. Our backbone employs an online–target architecture[10][2] to predict high-level latent features, prioritizing representation stability over signal reconstruction to facilitate unified multi-task discrimination.

UniNeuro is pretrained on more than 10,000 hours of EEG data and evaluated under a unified benchmark covering six heterogeneous downstream tasks. The results demonstrate strong unified multi-task performance and substantially narrow the gap between unified learning and conventional single-task fine-tuning.

The main contributions of our work are fourfold:

- **Unified EEG multi-task foundation framework.** We present UniNeuro, a unified EEG foundation framework that brings heterogeneous EEG tasks into a single discriminative paradigm, supporting heterogeneous-task pretraining, downstream adaptation within one learning pipeline.
- **Gradient-reversal task-conditional mixture-of-experts backbone.** We introduce a backbone with shared and routed experts, where shared experts are explicitly constrained via gradient reversal to learn task-shared EEG features while routed experts capture task-specific features, effectively decoupling shared and task-specific representations.
- **Hierarchical prototype-guided contrastive learning.** We explicitly organize the latent feature space at both the task level and the label level through Hierarchical Prototype-Guided Contrastive Learning, improving structured representation learning and multi-task discrimination.
- **Unified downstream multi-task evaluation.** Extensive experiments on six heterogeneous EEG benchmarks demonstrate that UniNeuro achieves strong unified multi-task performance and narrows the gap to task-specific fine-tuning. Further ablation studies and visualization analyses show that the task-conditional Mixture-of-Experts and Hierarchical Prototype-Guided Contrastive Learning jointly shape a structured latent space that supports unifying heterogeneous downstream tasks within a single discriminative framework.

2 Related Work

2.1 EEG Foundation Models

Recent EEG foundation models leverage large-scale, cross-dataset pretraining to learn transferable neural representations. BIOT[40] tokenizes heterogeneous biosignals into unified "biosignal sentences," accommodating variations in channel configurations, sequence lengths, and missing data. LaBraM[16] employs a neural tokenizer and masked neural code prediction for cross-dataset pretraining. EEGPT[35] combines masked dual self-supervision with spatio-temporal representation alignment, while CBraMod[36] uses a criss-cross Transformer to capture spatial and temporal dependencies, highlighting the value of EEG-specific architectural priors.

Despite their strong transferability, these models typically require separate adaptation for each downstream task. They therefore facilitate cross-dataset knowledge transfer but do not directly enable heterogeneous EEG tasks to be jointly learned within a single model.

2.2 Task Unification in EEG Modeling

Research on unifying heterogeneous EEG tasks remains limited. One emerging direction is to express different tasks through a common generative interface. Drawing on sequence-to-sequence models such as T5[28, 30], NeuroLM[15] reformulates heterogeneous EEG tasks as autoregressive prediction through EEG–language alignment, providing a unified interface across task-specific output spaces.

Computer vision and language modeling, meanwhile, provide another possible route: unifying heterogeneous tasks within a shared discriminative label space. Representative studies unify label spaces across datasets for object detection[45], entity–relation extraction[38], and independent multi-label segmentation tasks[42]. These approaches demonstrate that heterogeneous label systems can be jointly modeled without converting classification into sequence generation.

Most current downstream EEG tasks are still closed-set classification problems, and a unified discriminative route is therefore clearly the more practical solution. Based on this observation, our work focuses on a practical question: how to unify heterogeneous EEG tasks within a single discriminative framework while preserving strong task performance.

2.3 Structured Unified Learning

To make such a unified discriminative framework effective for heterogeneous downstream tasks, two issues are particularly important: how to balance feature sharing and task specialization inside a shared backbone, and how to organize the latent space under heterogeneous joint training.

Mixture-of-experts architectures provide a natural mechanism for addressing the first issue. MMoE[22] models task relationships through task-specific gates over a shared expert pool, while PLE[31] separates shared and task-specific experts to mitigate task conflicts and negative transfer. In brain-signal modeling, BrainMoE[39] applies MoE to large-scale fMRI representation learning through

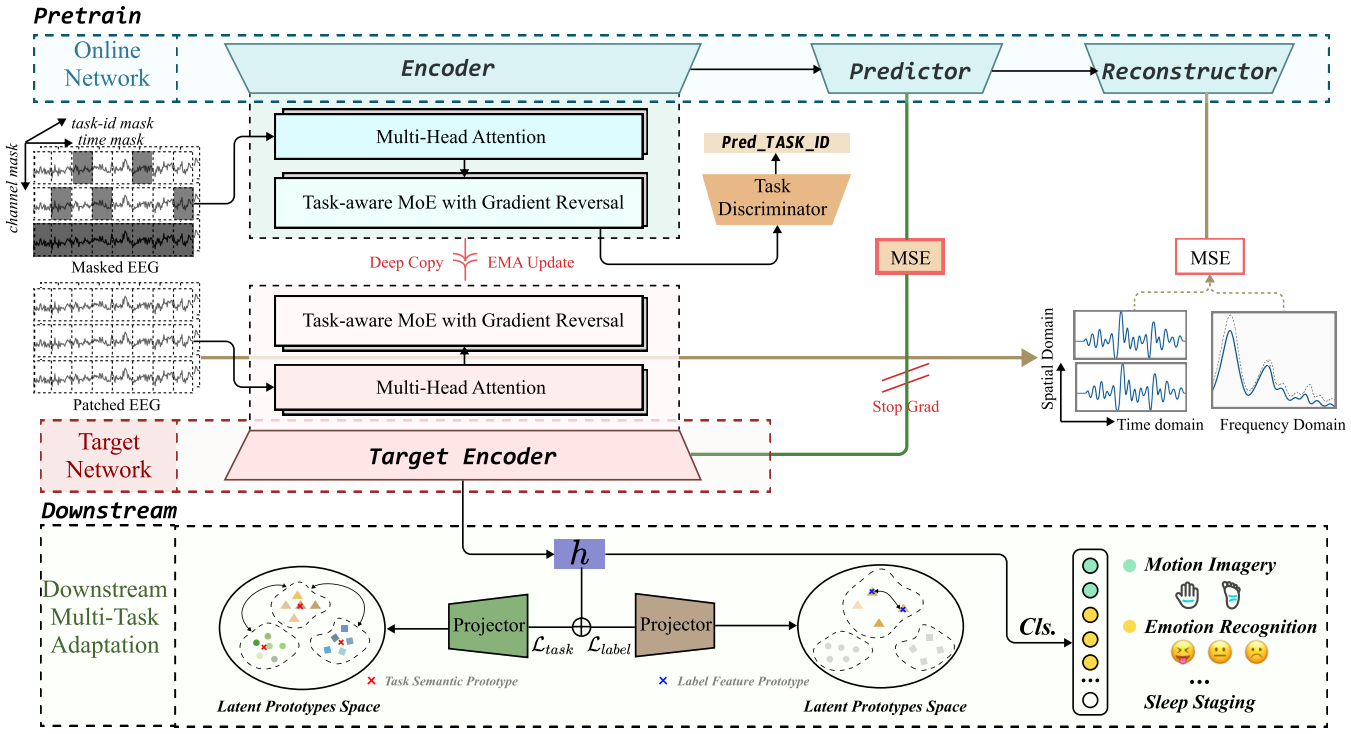

Figure 1: Architecture overview of the UniNeuro framework for unified EEG modeling. The pipeline consists of the following components: (a) Dual parallel networks (online and target) for robust latent feature extraction. (b) Reconstructor for masked EEG recovery. (c) Mixture-of-Experts (MoE) module for shared and task-specific feature specialization. (d) Downstream dual projectors for task-level and label-level representation organization. (e) Unified classifier operating on the structured latent space for dataset-conditioned predictions.

cognition joint embedding, and EEGMamba[11] incorporates task-aware experts into a bidirectional state-space model for multi-task EEG classification.

These studies motivate sparse expert routing for balancing shared and task-specific representations in heterogeneous EEG modeling.

The second issue concerns latent-space organization. Following the BYOL[10] and JEPA[2] paradigms, EEGPT[35] predicts high-level latent targets to reduce sensitivity to fine-grained signal noise[25]. SupCon[18] improves discriminability through supervised instance-level contrast, while ContraWR[41] introduces a global "world representation" as a contrastive reference for noisy EEG data. BENDR[19] further extends contrastive pretraining across EEG tasks and acquisition devices.

However, these objectives do not explicitly model the hierarchy between task identities and task-dependent labels. Because labels from different EEG tasks are not directly comparable, unified learning requires hierarchical semantic anchors that organize representations at both the task and within-task label levels.

3 Methods

3.1 Problem Formulation

Given a collection of heterogeneous EEG tasks, each sample is denoted as (x_i, m_i, d_i, y_i), where $x_i \in \mathbb{R}^{C_{d_i} \times T_{d_i}}$ is an EEG segment from task m_i with C_{d_i} channels and T_{d_i} temporal samples, $m_i \in \mathcal{M} = \{1, 2, \ldots, M\}$ is the task identity, $d_i \in \mathcal{U} = \{1, 2, \ldots, U\}$ is the dataset identity, and $y_i \in \mathcal{Y}_{d_i}$ is the corresponding dataset-specific label. Here, a single task may be supported by samples from multiple datasets.

We use $t_{m_i} = E_{\text{task}}(m_i)$ to denote the learnable embedding of task m_i, and aim to learn a task-conditional EEG backbone f_θ, shared across heterogeneous tasks, that maps the input EEG signal and task condition into a global representation h_i, i.e.,

$$h_i = f_\theta(x_i, \tilde{t}_{m_i}), \qquad h_i \in \mathbb{R}^D. \tag{1}$$

where $\tilde{t}_{m_i}$ denotes an optional task condition. Based on h_i, the model further produces the task-specific prediction

$$\hat{y}_i = \phi_{\text{cls}}(h_i, \tilde{t}_{m_i}). \tag{2}$$

where $\tilde{t}_{m_i}$ is also optional. The goal is to learn a unified model that captures transferable shared EEG structure while preserving task-specific discriminability across heterogeneous tasks.

3.2 Overview of the Framework

We propose a unified EEG foundation framework, **UniNeuro**, which supports multiple heterogeneous downstream tasks,as illustrated in Figure 1. The framework integrates pretraining and downstream adaptation into a single pipeline, enabling the model to capture transferable shared EEG representations while preserving task-specific discriminability.

In the pretraining phase, the backbone learns latent EEG representations through dual parallel networks (online and target) and a task-aware mixture-of-experts (MoE) that balances shared and task-specific feature extraction. Masked EEG components across temporal, channel, and task dimensions are reconstructed, allowing the model to capture high-quality spatio-temporal representations and task-relevant semantics.

During downstream adaptation, the target network features are projected into two latent spaces: a *task-level projector* for coarse-grained task semantics and a *label-level projector* for within-task label separability. Both spaces are optimized using Hierarchical Prototype-Guided Contrastive Learning. A unified classifier produces predictions in the dataset-conditioned label space.

Overall, UniNeuro enables a single model to simultaneously support multiple EEG tasks. Optional task-context recovery allows the model to maintain robust performance even when task identity is unavailable.

3.3 Pretraining

During pretraining, each EEG sample x_i is first transformed into a sequence of latent patches using an EEG patch embedding. A task embedding t_{m_i} is concatenated with each patch to preserve task semantics, and a channel embedding is added to account for different channel configurations across datasets. These task-conditional embeddings form the input to the backbone.

3.3.1 Online-Target Pretraining Architecture. Because low-level signal reconstruction can overfit noise in EEG, we use high-level latent prediction as the primary pretraining objective, complemented by auxiliary reconstruction. The backbone comprises parallel online and target networks[33]. The online network encodes inputs masked along the temporal, channel, and task dimensions, whereas the target network processes unmasked inputs and provides stable latent targets through momentum-updated parameters:

$$\theta_{\text{tg}} \leftarrow \lambda\theta_{\text{tg}} + (1-\lambda)\theta_{\text{on}}, \tag{3}$$

where θ_{tg} and θ_{on} denote the target and online parameters, respectively, and λ is the momentum coefficient. An auxiliary reconstructor further regularizes the online representations by recovering masked temporal, channel, and frequency components. Together, these objectives produce stable task-conditional representations for unified downstream discrimination.

3.3.2 Task-Conditional Mixture-of-Experts with Gradient Reversal. Separating task-invariant and task-specific EEG features is central to heterogeneous-task learning. Although MoE backbones combine shared and routed experts, task-aware routing alone does not explicitly constrain their respective roles. We therefore augment a task-conditional MoE with gradient reversal[8]: an adversarial task objective suppresses task-specific information in shared experts, while task-conditioned routing directs routed experts toward task-relevant representations.

MoE Architecture. The task-conditional MoE consists of E_s always-activated shared experts and E_r dynamically routed experts, with a Top-K sparsity constraint (Fig. 2). Denote the backbone feature as h_i and the task embedding as t_{m_i}. Routing weights are jointly

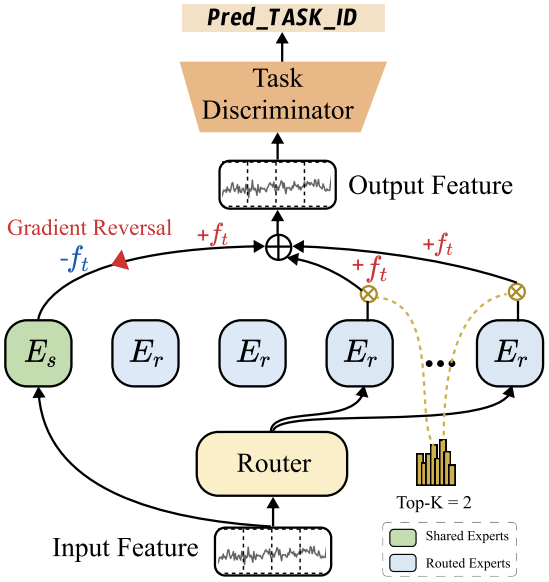

Figure 2: Gradient-reversal task-conditional mixture-of-experts.

computed from the backbone feature and task embedding as:

$$r_i = \text{softmax}(W_r h_i + W_t t_{m_i}), \qquad r_i \in \mathbb{R}^{E_r}, \tag{4}$$

where W_r and W_t are learnable parameters. Only the top-K experts with the highest routing scores are activated for each sample.

The outputs of shared and routed experts are combined to form the final MoE feature:

$$h_i^{\text{MoE}} = \underbrace{\frac{1}{E_s}\sum_{j=1}^{E_s} \text{Expert}_j^{\text{shared}}(h_i)}_{\text{shared experts}} + \underbrace{\sum_{j\in \text{TopK}(r_i)} r_{i,j}\cdot \text{Expert}_j^{\text{routed}}(h_i)}_{\text{routed experts}}. \tag{5}$$

Task-Adversarial Gradient Reversal. To encourage the shared experts to learn task-invariant representations, we introduce gradient reversal on the gradients from the task discrimination loss backpropagated to the shared experts. Let ϕ_{task} denote the task classifier and m_i the task identity. The adversarial loss for shared experts is:

$$\mathcal{L}_{\text{task-shared}} = -\mathcal{L}_{\text{CE}}\left(\phi_{\text{task}}(h_i^{\text{shared}}), m_i\right), \tag{6}$$

which inverts the gradient during backpropagation, preventing shared experts from encoding task-specific information. In contrast, routed experts are trained normally to predict task identity. This design ensures a natural decoupling: shared experts capture task-invariant EEG representations, while routed experts encode task-specific features. Furthermore, the task classifier also allows "optional task-context recovery": when task identity is unavailable, the model can infer task semantics from the input, improving robustness in practical deployment.

3.3.3 Pretraining Losses. The total pretraining objective consists of multiple components:

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{latent}} + \mathcal{L}_{\text{load}} + \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{router}}, \tag{7}$$

where each term is defined as follows:

- **Reconstruction Loss** $\mathcal{L}_{\text{recon}}$: encourages the reconstructor to recover masked EEG components, including temporal, channel, and task identity dimensions, as well as patch embeddings and short-time Fourier transform features.
- **Predictor Alignment Loss** $\mathcal{L}_{\text{latent}}$:

$$\mathcal{L}_{\text{latent}} = \left\| p_i^{\text{pred}} - \text{sg}(h_i^{\text{tg}}) \right\|_2^2, \tag{8}$$

  where p_i^{pred} is the predictor output of the online network, h_i^{tg} is the target network feature, and $\text{sg}(\cdot)$ denotes stop-gradient.
- **Load Balancing Loss** $\mathcal{L}_{\text{load}}$: encourages balanced utilization of routed experts to prevent collapse to a few experts.
- **Task Prediction Loss** $\mathcal{L}_{\text{task}}$: trains the model to predict the task identity from latent representations, supporting task-conditional modeling.
- **Router Consistency Loss** $\mathcal{L}_{\text{router}}$: regularizes the routing outputs to maintain stability and consistency across similar inputs.

3.4 Downstream Adaptation

Although the backbone learned rich latent EEG representations during pretraining, these features are not automatically structured for multiple heterogeneous downstream tasks. In practice, task-level clusters may overlap, and within-task class separability may be insufficient, leading to interference when training a unified classifier. To address these challenges, we design a hierarchical representation learning scheme that explicitly organizes the latent space at both task and label levels.

3.4.1 Hierarchical Prototype-Guided Contrastive Learning. We propose *Hierarchical Prototype-Guided Contrastive Learning (HPGCL)* to disentangle task-level and label-level semantics in the latent space. Let h_i denote the high-level feature from the target network. The feature is projected into two spaces:

Task-level latent space: captures coarse-grained task semantics. For each task m, we maintain a momentum-updated prototype c_m. Task-level supervised contrastive loss encourages features of the same task to cluster around their prototype while pushing features of different tasks apart:

$$\mathcal{L}_{\text{task}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(z_i^{\text{task}}, c_{m_i})/\tau_t\right)}{\sum_{m=1}^{M} \exp\left(\text{sim}(z_i^{\text{task}}, c_m)/\tau_t\right)}, \tag{9}$$

where τ is a temperature hyperparameter.

Label-level latent space: captures fine-grained within-task label separability. For each class y under task m, we maintain a task-conditioned label prototype $l_{m,y}$. Label-level contrastive loss encourages features to cluster around the correct label prototype:

$$\mathcal{L}_{\text{label}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(z_i^{\text{label}}, l_{m_i,y_i})/\tau_l\right)}{\sum_{y \in \mathcal{Y}_{m_i}} \exp\left(\text{sim}(z_i^{\text{label}}, l_{m_i,y})/\tau_l\right)}, \tag{10}$$

The combined HPGCL objective is:

$$\mathcal{L}_{\text{HPGCL}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{label}}. \tag{11}$$

Table 1: Summary of the datasets used in this work.

Dataset	Task Type	#Subj.	Hours	Cls.
Pretraining Datasets				
TUEG[27] (subset)	Clinical EEG	4000	≈ 6000	SSL
SEED-IV[43]	Emotion Recognition	15	45	4
SEED-V[21]	Emotion Recognition	20	60	5
SleepEDFx[17]	Sleep Staging	197	3849	5
KaggleERN[23]	ERP Detection	26	30	2
BCI2B[29]	Motor Imagery	9	26.3	2
TSUBenchmark[37]	SSVEP Classification	35	14	40
Downstream Benchmark Datasets				
TUAB[4]	Abnormal Detection	2383	1137	2
TUEV[12]	Event Classification	370	148	6
SEED[44][5]	Emotion Recognition	15	45	3
HMC[1]	Sleep Staging	151	582	5
PhysioP300[9]	ERP Detection	10	2.3	2
BCICIV2a[32]	Motor Imagery	9	13.4	4

3.4.2 Unified Discriminative Classification. On top of the structured latent space, a unified classifier ϕ_{cls} produces predictions in the dataset-conditioned label space. Let y_i be the ground-truth label; the cross-entropy loss is:

$$\mathcal{L}_{\text{CE}} = -\sum_i y_i \log \phi_{\text{cls}}(h_i). \tag{12}$$

The final downstream adaptation objective combines hierarchical contrastive learning with supervised classification:

$$\mathcal{L}_{\text{downstream}} = \lambda_{\text{HPGCL}}\mathcal{L}_{\text{HPGCL}} + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}}, \tag{13}$$

This design ensures that a single model can simultaneously support multiple heterogeneous EEG tasks while preserving both task-level and label-level feature structure. Optional task-context recovery allows inference even when task identity is unavailable, enhancing robustness and deployability.

4 Experiments

4.1 Datasets

We use a heterogeneous collection of EEG datasets for both pretraining and downstream evaluation, covering several representative paradigms, including clinical EEG analysis, emotion recognition, sleep staging, ERP decoding, motor imagery (MI), and SSVEP-based BCI, as summarized in Table 1.

Dataset Splits: For pre-training, we follow the same setting as EEGPT[35] and randomly sample 10% of the training data as the validation set. For downstream multi-task evaluation, we adopt a unified dataset split protocol to ensure fair comparison across tasks. Specifically, for TUAB[4] and TUEV[12], we follow the official train/test split and further split the training patients into training and validation sets with an 80/20 ratio. For SEED[44], we divided all 15 trials into training, validation, and test sets in a 9:3:3 ratio and merged the sessions. For BCICIV2a[32], and PhysioP300[9], we use leave-one-subject-out (LOSO) evaluation. For HMC[1], we use the first 100 subjects for training, the next 25 for validation, and the remaining 26 for testing.

Preprocessing: All EEG signals are preprocessed using a uniform pipeline. Specifically, all recordings are resampled to 256 Hz

Table 2: Unified downstream multi-task evaluation on six representative EEG benchmarks. The *Multi-task* column indicates whether a model is trained under the unified multi-task setting. Within the single-task specialist group (BIOT[40], LaBraM[16], EEGPT[35]), the best result is underlined. Within the unified multi-task group (NeuroLM[15], Ours-Tiny, Ours-Base, Ours-Large), the best result is highlighted in bold.

TUAB

Model	Multi-task	Balanced Acc.	AUC-PR	AUROC
BIOT	✗	0.7959±0.0057	0.8792±0.0023	0.8815±0.0043
LaBraM	✗	0.8140±0.0019	0.8965±0.0016	0.9022±0.0009
EEGPT	✗	0.7983±0.0030	0.8827±0.0106	0.8718±0.0050
NeuroLM	✓	0.7826±0.0065	0.6975±0.0081	0.7816±0.0079
Ours-Tiny	✓	0.7723±0.0075	0.8484±0.0135	0.8627±0.0062
Ours-Base	✓	0.7912±0.0035	0.8501±0.0164	0.8624±0.0026
Ours-Large	✓	**0.7964±0.0067**	**0.8536±0.0025**	**0.8653±0.0011**

TUEV

Model	Multi-task	Balanced Acc.	Cohen's Kappa	Weighted F1
BIOT	✗	0.5281±0.0225	0.5273±0.0249	0.7492±0.0082
LaBraM	✗	0.6409±0.0065	0.6637±0.0093	0.8312±0.0052
EEGPT	✗	0.6232±0.0114	0.6351±0.0134	0.8187±0.0063
NeuroLM	✓	0.4560±0.0048	0.4285±0.0048	0.7153±0.0028
Ours-Tiny	✓	0.5584±0.0096	0.5638±0.0033	0.7320±0.0064
Ours-Base	✓	0.5739±0.0025	0.5827±0.0092	0.7462±0.0051
Ours-Large	✓	**0.5813±0.0061**	**0.5884±0.0034**	**0.7588±0.0048**

SEED

Model	Multi-task	Balanced Acc.	Cohen's Kappa	Weighted F1
BIOT	✗	0.7097±0.0024	0.5682±0.0051	0.7134±0.0027
LaBraM	✗	0.7318±0.0019	0.5994±0.0031	0.7354±0.0021
EEGPT	✗	0.7228±0.0068	0.5915±0.0075	0.7043±0.0071
NeuroLM	✓	0.5554±0.0075	0.3393±0.0117	0.5599±0.0068
Ours-Tiny	✓	0.6626±0.0016	0.5312±0.0042	0.6318±0.0071
Ours-Base	✓	0.6835±0.0047	0.5313±0.0064	0.6324±0.0053
Ours-Large	✓	**0.6860±0.0013**	**0.5566±0.0029**	**0.6534±0.0011**

HMC

Model	Multi-task	Balanced Acc.	Cohen's Kappa	Weighted F1
BIOT	✗	0.6862±0.0041	0.6295±0.0113	0.7091±0.0147
LaBraM	✗	0.7286±0.0101	0.6812±0.0073	0.7554±0.0024
EEGPT	✗	0.7344±0.0120	0.6821±0.0125	0.8126±0.0093
NeuroLM	✓	0.6737±0.0050	0.6188±0.0057	**0.7126±0.0034**
Ours-Tiny	✓	0.6652±0.0120	0.6320±0.0051	0.6946±0.0103
Ours-Base	✓	0.6673±0.0138	0.6482±0.0098	0.6953±0.0110
Ours-Large	✓	**0.6750±0.0112**	**0.6578±0.0086**	0.7032±0.0085

PhysioP300

Model	Multi-task	Balanced Acc.	Cohen's Kappa	AUROC
BIOT	✗	0.5485±0.0325	0.0968±0.0647	0.5308±0.0333
LaBraM	✗	0.6477±0.0110	0.2935±0.0277	0.7068±0.0134
EEGPT	✗	0.6502±0.0063	0.2999±0.0139	0.7168±0.0051
NeuroLM	✓	0.5292±0.0143	0.2219±0.0156	0.5234±0.0129
Ours-Tiny	✓	0.5517±0.0090	0.2401±0.0149	0.6371±0.0111
Ours-Base	✓	0.5620±0.0067	0.2433±0.0104	0.6412±0.0096
Ours-Large	✓	**0.5746±0.0013**	**0.2531±0.0102**	**0.6412±0.0089**

BCICIV2a

Model	Multi-task	Balanced Acc.	Cohen's Kappa	Weighted F1
BIOT	✗	0.4590±0.0196	0.2787±0.0261	0.4282±0.0289
LaBraM	✗	0.5613±0.0058	0.4151±0.0069	0.5520±0.0052
EEGPT	✗	0.5846±0.0070	0.4462±0.0094	0.5715±0.0051
NeuroLM	✓	0.4314±0.0133	0.3145±0.0114	0.4467±0.0003
Ours-Tiny	✓	0.5249±0.0015	0.3794±0.0027	0.5164±0.0052
Ours-Base	✓	0.5415±0.0022	**0.3931±0.0050**	**0.5349±0.0016**
Ours-Large	✓	**0.5461±0.0067**	0.3927±0.0074	0.5315±0.0019

to unify temporal resolution. To remove physiological artifacts and line noise, a band-pass filter from 0.1 Hz to 75 Hz and a 50 Hz notch filter are applied. The signals are then segmented into fixed-length 4 s windows, and the amplitudes are standardized to microvolts (μV) with consistent channel alignment across datasets. These preprocessed segments are subsequently used for both pre-training and downstream evaluation.

4.2 Experimental Settings

Pretraining. We pretrained three scales of the UniNeuro model: Tiny (11M), Base (27M), and Large (86M), where the total parameter count includes all MoE components.

Downstream Multi-task Training. We jointly fine-tuned UniNeuro across all six downstream tasks. Each mini-batch was constructed using task-wise block sampling with a fixed task ratio, ensuring that every task contributed to each optimization step. The pretrained target encoder, hierarchical projectors, and unified classifier were jointly optimized using the same optimization setup as in pretraining. Each experiment was repeated with three random seeds. We report task-appropriate metrics, including balanced accuracy (BA), area under the precision–recall curve (AUC-PR), area under the receiver operating characteristic curve (AUROC), weighted F1 score, and Cohen's kappa. For each task, the checkpoint with the best corresponding validation performance was used for test evaluation.

Implementation and Hardware. All models were implemented in PyTorch 2.7.0 and trained on eight NVIDIA A6000 GPUs using mixed-precision (FP16) with CUDA 11.7. Distributed data-parallel training ensured efficient multi-GPU scaling, and fixed random seeds guaranteed reproducibility. Training logs and validation metrics were recorded throughout.

Baseline Methods. For multi-task evaluation, NeuroLM[15] was used as the primary baseline. Additionally, three single-task strong baselines—BIOT[40], LaBraM[16], and EEGPT[35]—were included. Single-task baselines were trained under identical budgets and hyperparameters and fine-tuned independently for each task.

4.3 Unified Downstream Multi-task Evaluation

We evaluate UniNeuro on six representative EEG benchmarks, covering abnormality detection (TUAB[4]), event classification (TUEV[12]), emotion recognition (SEED[44]), sleep staging (HMC[1]), ERP decoding (PhysioP300[9]), and motor imagery classification (BCICIV2a[32]). We compare against three single-task specialist baselines (BIOT[40], LaBraM[16], and EEGPT[35]) and one unified multi-task baseline (NeuroLM[15]). The single-task models serve as reference upper bounds, while NeuroLM provides the primary baseline under the same unified multi-task setting.

As shown in Table 2, UniNeuro consistently outperforms NeuroLM on most benchmarks under the unified setting. Using the

Base model as reference, the Balanced Accuracy gain reaches 11.8% on TUEV and 12.8% on SEED, and remains notable on more challenging low-resource tasks such as PhysioP300 (+3.3%) and BCICIV2a (+11.0%). On TUAB and HMC, although the Balanced Accuracy improvement is smaller, UniNeuro remains competitive and achieves slightly better Cohen's Kappa, indicating more stable class-consistent predictions.

Compared with single-task specialists, the gap varies across tasks. On TUAB and SEED, UniNeuro narrows the gap to 2.3% and 4.8%, respectively, showing that unified multi-task learning can approach task-specific fine-tuning when sufficient shared structure exists across tasks. In contrast, larger deficits remain on PhysioP300 and BCICIV2a, where limited data and finer transient discriminative patterns make joint optimization more sensitive to data imbalance. Overall, these results suggest that UniNeuro is a strong unified discriminative framework for heterogeneous EEG task learning: it delivers clear gains over the prior unified baseline while substantially reducing, though not fully eliminating, the gap to single-task modeling.

4.4 Task-aware MoE Analysis

Under severe dataset imbalance, we observe that expert utilization tends to be dominated by high-resource tasks despite task-wise batch partitioning. To isolate the contribution of task-aware expert routing, we conduct the main Task-MoE ablation on four tasks with relatively balanced training-set sizes: TUEV[12], SEED[44], HMC[1], and BCICIV2a[32]. All experiments follow the same data-split protocol as in Sec. 4.1. This controlled setting reduces the confounding effects of dominant large-scale datasets, allowing performance differences to be attributed more directly to the routing mechanism.

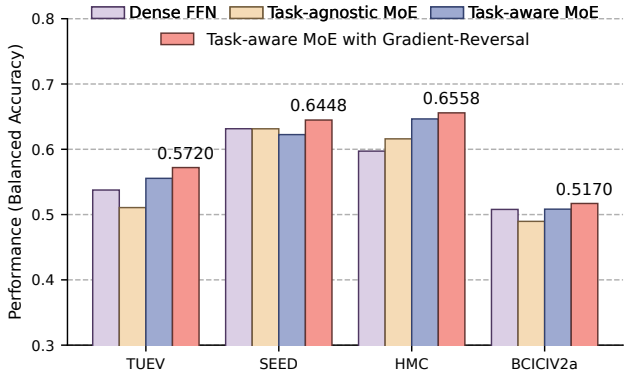

Figure 3: Controlled ablation of expert routing mechanisms on four downstream tasks with relatively balanced dataset scales.

Quantitative Ablation. We compare four parameter-matched variants under the same downstream training setting with task IDs provided: (1) Dense FFN, where MoE blocks are replaced by standard feed-forward layers; (2) Task-agnostic MoE, retaining multiple experts but without task-conditioned routing; (3) Task-aware MoE; and (4) Task-aware MoE with Gradient-Reversal. As shown in Fig. 3, Task-aware MoE improves average balanced accuracy from 0.569 (Dense FFN) and 0.562 (Task-agnostic MoE) to 0.583 across the four tasks, corresponding to relative gains of 2.6% and 3.8%, respectively. Adding Gradient-Reversal further boosts the average balanced accuracy to 0.597, achieving an additional 2.4% improvement. These results indicate that expert capacity alone is insufficient, and task-conditioned routing is essential for reliable multi-task adaptation. Moreover, the combination with Gradient-Reversal enhances adaptation by further promoting task-specific feature separation.

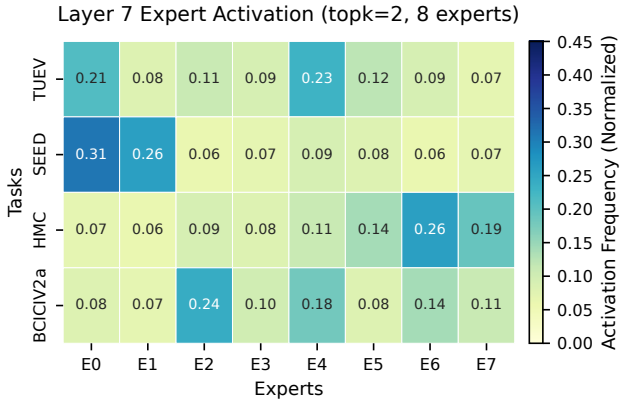

Figure 4: Expert activation heatmaps

Expert Routing Visualization. Figure 4 shows that Task-aware MoE develops distinct task-dependent routing patterns in deeper FFN blocks while maintaining balanced expert utilization overall. In contrast, Task-agnostic MoE produces nearly uniform activation patterns across tasks under the load-balancing objective, suggesting limited expert specialization. This contrast indicates that task conditioning, rather than expert competition alone, is responsible for forming task-specific computation paths.

Overall, this controlled ablation confirms that task-aware expert routing is key to translating expert modularity into consistent downstream improvements, and that combining it with Gradient-Reversal further enhances multi-task adaptation across heterogeneous EEG tasks.

4.5 Ablation of Hierarchical Prototype-Guided Contrastive Learning (HPG-Contrastive)

We conduct a controlled ablation on the unified six-task downstream benchmark to quantify the individual and joint contributions of the task- and label-level HPG-Contrastive objectives. All variants share the same backbone initialization and training protocol.

Quantitative Ablation. As shown in Fig. 5 and Table 3, the task- and label-level variants improve average BA from 46.98% (CE-only) to 53.48% (+6.50 points) and 60.69% (+13.71 points), respectively. Combining both objectives yields the highest average BA of 64.02%, a 17.04-point improvement over CE-only. In contrast, the flat global variant with a single projector reaches only 47.96% (+0.98 points),

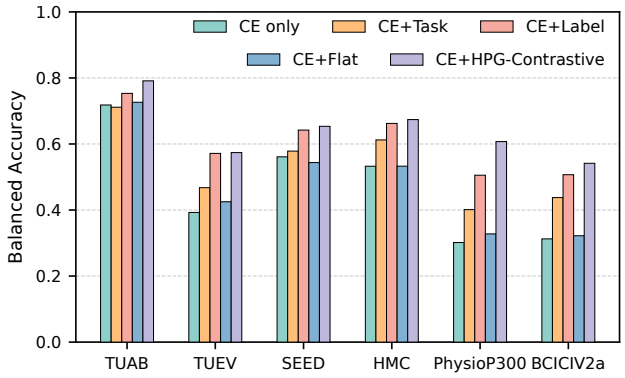

Figure 5: Ablation results of HPG-Contrastive on the unified six-task downstream benchmark. CE-only is the classification-only baseline.

Table 3: Average Balanced Accuracy (%) of contrastive variants; ↑ shows gain over CE-only.

Method	Avg. BA (%)	↑
CE only (cross-entropy)	46.98	–
CE + Task-level HPG-Contrastive	53.48	+6.50
CE + Label-level HPG-Contrastive	60.69	+13.71
CE + Flat Global HPG-Contrastive	47.96	+0.98
CE + HPG-Contrastive (full)	**64.02**	**+17.04**

suggesting that a single contrastive space inadequately models task-dependent label semantics.

Mechanistic Interpretation. The task-level objective establishes coarse inter-task organization, while the label-level objective refines class structure within each task. Their complementary gains, together with the near-baseline performance of the flat global variant, indicate that HPG-Contrastive benefits from explicitly modeling the task–label hierarchy rather than applying contrastive regularization in an undifferentiated global space.

4.6 Representation Space Analysis

To examine how HPG-Contrastive structures the learned representation space, we visualize the task-level projector outputs z^{task} across all six tasks and the label-level outputs z^{label} on SEED using UMAP. Embeddings are colored by task identity and class label, respectively. All comparisons use balanced sampling, identical sample identities, and the same PCA+UMAP configuration, with checkpoints selected a priori for each analysis. Silhouette scores are computed in the original high-dimensional spaces rather than the 2D projections.

As shown in Fig. 6(a)–(b), the CE-only baseline already exhibits visible task-level separation, indicating that the pretrained task-aware backbone captures coarse task-dependent EEG structure. Nevertheless, several task clusters remain partially overlapping.

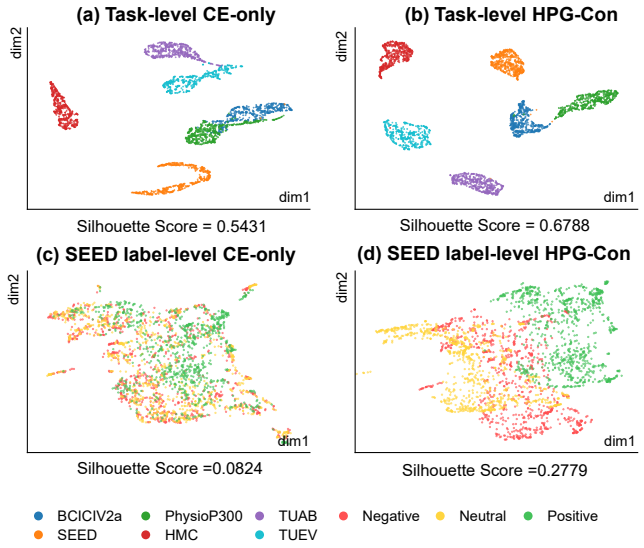

Figure 6: UMAP visualization of learned EEG representations. (a)–(b) Task-level embeddings for CE-only and HPG-Contrastive. (c)–(d) Label-level embeddings on the SEED dataset for CE-only and HPG-Contrastive.

Task-level HPG-Contrastive produces more compact intra-task clusters and clearer inter-task separation, increasing the silhouette score from 0.5431 to 0.6788.

Figure 6(c)–(d) shows the label-level representations on SEED. The three emotion classes overlap substantially under CE-only, indicating limited within-task separability. Label-level HPG-Contrastive improves within-class compactness and between-class separation, increasing the silhouette score from 0.0824 to 0.2779, alongside an absolute BA gain of 0.081. The relatively low final silhouette score nevertheless indicates that affective EEG states remain difficult to separate.

Together, these results show that HPG-Contrastive improves both inter-task organization and within-task class structure, consistent with its downstream performance gains.

5 Conclusions

In this work, we proposed UniNeuro, a unified EEG foundation model that enables joint learning of heterogeneous EEG tasks within a single architecture. By integrating a task-conditional Mixture-of-Experts (MoE) backbone with Hierarchical Prototype-Guided Contrastive Learning (HPGCL), UniNeuro effectively decouples shared and task-specific neural representations and organizes the latent space at both task and label levels. Pretrained on over 10,000 hours of EEG data and evaluated across six diverse downstream tasks, UniNeuro demonstrates strong unified multi-task discriminative performance, and achieves performance competitive with single-task fine-tuning on several tasks. The framework provides a generalizable foundation for EEG-based applications, including abnormality detection, emotion recognition, sleep staging, ERP decoding, and motor imagery, and offers a versatile infrastructure for future multi-task and multimodal EEG research.

Acknowledgments

This work was supported in part by the Yeqisun Joint Funds of the National Natural Science Foundation of China under Grant No. U2441252; the National Natural Science Foundation of China under Grant Nos. 62331001 and 62271155; the Shanghai Municipal Science and Technology Major Project, administered under Project No. 2021SHZDZX0100; the Changjiang Scholars Program of China; the Fundamental Research Funds for the Central Universities; and the "Computational Biology" Project of the Shanghai Key Technology Research and Development Plan under Project No. 25JS2840100.

References

[1] Diego Alvarez-Estevez and Roselyne Rijsman. 2022. Haaglanden Medisch Centrum sleep staging database. *PhysioNet* (March 2022). Version 1.1. doi:10.13026/t79q-fr32

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15619–15629.

[3] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1280–1297.

[4] Silvia López de Diego. 2017. *Automated interpretation of abnormal adult electroencephalograms*. Temple University.

[5] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. 2013. Differential entropy feature for EEG-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 81–84.

[6] Kübra Erat, Elif Bilge Sahin, Furkan Dogan, Nur Merdanoglu, Ahmet Akcakaya, and Pinar Onay Durdu. 2024. Emotion recognition with EEG-based brain-computer interfaces: a systematic literature review. *Multim. Tools Appl.* 83, 33 (2024), 79647–79694. doi:10.1007/S11042-024-18259-Z

[7] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

[8] Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings)*, Francis R. Bach and David M. Blei (Eds.). JMLR.org, 1180–1189. http://proceedings.mlr.press/v37/ganin15.html

[9] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation* 101, 23 (2000), e215–e220.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[11] Yiyu Gui, MingZhi Chen, Yuqi Su, Guibo Luo, and Yuchao Yang. 2024. EEGMamba: Bidirectional state space model with mixture of experts for EEG multi-task classification. *arXiv preprint arXiv:2407.20254* (2024).

[12] Amir Harati, Meysam Golmohammadi, Silvia Lopez, Iyad Obeid, and Joseph Picone. 2015. Improved EEG event classification using differential energy. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE, 1–4.

[13] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2020), 2.

[14] Ravichander Janapati, Vishwas Dalal, and Rakesh Sengupta. 2023. Advances in modern EEG-BCI signal processing: A review. *Materials Today: Proceedings* 80 (2023), 2563–2566.

[15] Weibang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. 2025. NeuroLM: A Universal Multi-task Foundation Model for Bridging the Gap between Language and EEG Signals. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net. https://openreview.net/forum?id=Io9yFt7XH7

[16] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. 2024. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=QzTpTRVtrP

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

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.

[19] Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. 2021. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience* 15 (2021), 653659.

[20] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. 2020. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966* (2020).

[21] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. 2021. Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition. *IEEE Transactions on Cognitive and Developmental Systems* (2021).

[22] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Yike Guo and Faisal Farooq (Eds.). ACM, 1930–1939. doi:10.1145/3219819.3220007

[23] Perrin Margaux, Maby Emmanuel, Daligault Sébastien, Bertrand Olivier, and Mattout Jérémie. 2012. Objective and Subjective Evaluation of Online Error Correction during P300-Based Spelling. *Advances in Human-Computer Interaction* 2012, 1 (2012), 578295.

[24] Dennis J McFarland and Jonathan R Wolpaw. 2011. Brain-computer interfaces for communication and control. *Commun. ACM* 54, 5 (2011), 60–66.

[25] Lorenzo Mur-Labadia, Matthew J. Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. 2026. V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning. *CoRR* abs/2603.14482 (2026). arXiv:2603.14482 doi:10.48550/ARXIV.2603.14482

[26] Ernst Niedermeyer and FH Lopes da Silva. 2005. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins.

[27] Iyad Obeid and Joseph Picone. 2016. The temple university hospital EEG data corpus. *Frontiers in neuroscience* 10 (2016), 196.

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. https://jmlr.org/papers/v21/20-074.html

[29] David Steyrl, Reinhold Scherer, Josef Faller, and Gernot R. Müller-Putz. 2016. Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: a practical and convenient non-linear classifier. *Biomedical Engineering / Biomedizinische Technik* 61 (2016), 77 – 86. https://api.semanticscholar.org/CorpusID:7782626

[30] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 3104–3112. https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

[31] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. In *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, Rodrygo L. T. Santos, Leandro Balby Marinho, Elizabeth M. Daly, Li Chen, Kim Falk, Noam Koenigstein, and Edleno Silva de Moura (Eds.). ACM, 269–278. doi:10.1145/3383313.3412236

[32] Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. 2012. Review of the BCI Competition IV. *Frontiers in Neuroscience* 6 (2012). https://api.semanticscholar.org/CorpusID:790253

[33] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=ry8u21rtl

[34] Swati Vaid, Preeti Singh, and Chamandeep Kaur. 2015. EEG signal analysis for BCI interface: A review. In *2015 fifth international conference on advanced computing & communication technologies*. IEEE, 143–147.

[35] Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. 2024. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS*

2024, Vancouver, BC, Canada, December 10 - 15, 2024, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html

[36] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. 2025. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net. https://openreview.net/forum?id=NPNUHgHF2w

[37] Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. 2016. A benchmark dataset for SSVEP-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 10 (2016), 1746–1752.

[38] Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021. UniRE: A Unified Label Space for Entity Relation Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 220–231. doi:10.18653/V1/2021.ACL-LONG.19

[39] Ziquan Wei, Tingting Dan, Tianlong Chen, and Guorong Wu. 2025. BrainMoE: Cognition Joint Embedding via Mixture-of-Expert Towards Robust Brain Foundation Model. *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)* (2025).

[40] Chaoqi Yang, M Westover, and Jimeng Sun. 2023. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems* 36 (2023), 78240–78260.

[41] Chaoqi Yang, Cao Xiao, M Brandon Westover, and Jimeng Sun. 2023. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI* 2, 1 (2023), e46769.

[42] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. 2020. Object Detection with a Unified Label Space from Multiple Datasets. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV (Lecture Notes in Computer Science)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 178–193. doi:10.1007/978-3-030-58568-6_11

[43] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki. 2018. EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. *IEEE Transactions on Cybernetics* (2018), 1–13. doi:10.1109/TCYB.2018.2797176

[44] Wei-Long Zheng and Bao-Liang Lu. 2015. Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development* 7, 3 (2015), 162–175. doi:10.1109/TAMD.2015.2431497

[45] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2022. Simple Multi-dataset Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 7561–7570. doi:10.1109/CVPR52688.2022.00742