EMPOWERING BRAIN-COMPUTER INTERFACE DECODING WITH LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding brain signals supports a broad range of brain–computer interface (BCI) applications spanning cognitive, motor, and clinical domains. These tasks vary substantially in their targeted brain states, neural patterns, and output spaces. Despite recent advances in brain foundation models, the field still relies on disjointed prediction heads and task-specific adaptation pipelines, limiting unified knowledge sharing and robust generalization across tasks. To overcome these limitations, we introduce UniMind, a unified multi-task brain decoding framework driven by large language models (LLMs). UniMind reformulates heterogeneous task outputs into a unified question–answer paradigm, casting diverse neural signals into a universal semantic manifold optimized via a single shared autoregressive objective. To reconcile task heterogeneity within a single model, we design a task-aware query selector to dynamically extract task-relevant representations. Furthermore, a neuro-language connector is proposed to distill high-dimensional spatiotemporal brain patterns into compact neural-semantic representations directly interpretable by LLMs. Across ten scalp EEG datasets spanning five task domains, UniMind achieves an average improvement of 8.94 percentage points in balanced accuracy. It shows strong cross-subject performance and generalizes to unseen datasets. Beyond scalp EEG, we show that UniMind also benefits intracortical spike decoding in non-human primates, raising average R^2 and producing more separable spike representations. These results establish UniMind as a scalable framework for unified neural decoding across diverse subjects, tasks, and recording modalities.

1 INTRODUCTION

Decoding high-dimensional neural activity into task-relevant representations is a foundational challenge in brain–computer interface (BCI) research. With advances in neural sensing and computational modeling, BCI systems have been increasingly explored across diverse cognitive, motor, and clinical domains. Non-invasive scalp electroencephalography (EEG), owing to its high temporal resolution, portability, and relatively low cost, has supported a broad range of applications including emotion recognition (Koelstra et al., 2012), motor imagery classification (Amin et al., 2019), clinical assessment (Roy et al., 2019; Boonyakitanont et al., 2020), sleep staging (Aboalayon et al., 2016), and stress estimation (Sharma et al., 2022). Meanwhile, invasive intracortical recordings enable high-bandwidth decoding such as speech restoration (Willett et al., 2023; Metzger et al., 2023) and cursor control (Hochberg et al., 2006).

However, different BCI tasks vary substantially in the brain states they target, the neural patterns required for prediction, and their output spaces. Existing high-performing models are therefore predominantly designed and optimized for a specific dataset or task (Dar et al., 2020; Song et al., 2021; Peh et al., 2022; Jing et al., 2023). Recent pretrained EEG foundation models, such as BIOT (Yang et al., 2023a), LaBraM (Jiang et al., 2024), and CBraMod (Wang et al., 2025), have sought to learn robust representations through self-supervised pretraining on large-scale unlabeled EEG corpora. Nevertheless, this progress leaves the downstream decoding paradigm largely unchanged: each task still relies on a task-specific prediction head and a separately optimized adaptation pipeline, resulting in redundant tuning effort and limiting knowledge transfer across related tasks. These limitations highlight the need for a general-purpose neural decoding framework that handles heterogeneous

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

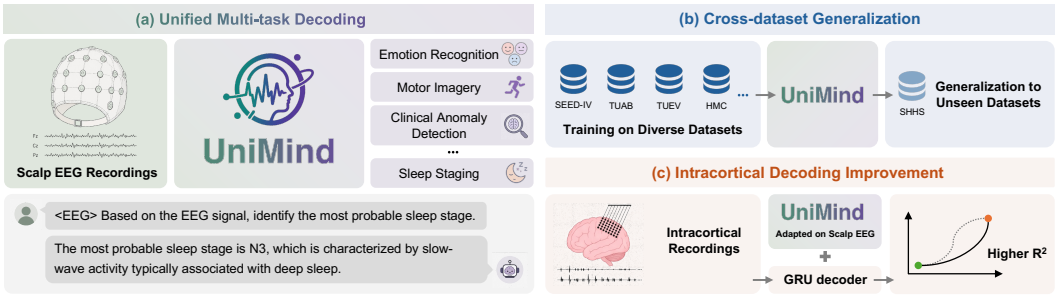


Figure 1: Overview of UniMind and its key capabilities. (a) A single model unifies heterogeneous scalp EEG tasks as question–answer generation. (b) Multi-dataset training advances generalization to unseen datasets. (c) The EEG-adapted LLM input space regularizes an intracortical decoder and improves kinematic decoding.

tasks within a unified model, enabling knowledge sharing across related objectives and potentially stronger generalizability beyond seen datasets.

Achieving such a unified neural decoding framework remains challenging due to two fundamental obstacles. First, unlike visual or linguistic modalities where inputs share a consistent structure (*e.g.*, pixel grids or fixed vocabularies), EEG signals lack naturally organized structures to support generalizable representations. They are inherently non-stationary and suffer from distributional shifts across subjects, recording hardware, and experimental paradigms (Jayaram et al., 2016; Wu et al., 2022; Saha & Baumert, 2020). This variability makes it difficult for conventional decoders to map diverse physiological inputs into a shared representation space without extensive domain-specific tuning. Second, EEG tasks exhibit substantial heterogeneity in both their underlying neural patterns and target label spaces. As different decoding tasks depend on distinct spatiotemporal features (*e.g.,* spatially structured modulations of sensorimotor mu and beta rhythms for motor imagery (McFarland et al., 2000; Kim et al., 2016) vs. stage-specific spectral dynamics and temporal dependencies across epochs for sleep staging (Supratak et al., 2017; Phan et al., 2019)), straightforward parameter sharing across these diverse objectives may lead to negative transfer.

Large language models (LLMs) provide a promising pathway to address these challenges. Their large-scale pretraining establishes a universal semantic manifold that has been shown to be capable of anchoring and aligning heterogeneous representations in visual and robotic domains (Li et al., 2023; Liu et al., 2023; Xu et al., 2025; Kim et al., 2024). The recently proposed NeuroLM (Jiang et al., 2025) takes an initial step in this direction by aligning EEG embeddings with language space through domain-adversarial training. However, such domain-level alignment is insufficient for the LLM to learn and interpret the underlying neural patterns in the signals, leaving a large performance gap relative to task-specific models.

Here we introduce UniMind (Figure 1), a unified multi-task neural decoding framework that reformulates diverse BCI tasks as natural-language question–answer problems, mapping them into a common autoregressive prediction space. UniMind explicitly identifies and translates the task-relevant neural patterns required for the LLM to understand and reason about. Specifically, a Task-aware Query Selector (TQS) maintains spatial and temporal query pools and dynamically selects queries according to the distinct neural dynamics underlying each decoding task. Conditioned on these selected queries, a Neuro-Language Connector (NLC) employs dual-branch cross-attention to aggregate task-relevant spatiotemporal patterns and project them into neural-semantic features interpretable by the LLM. By coupling task-aware feature selection with fine-grained neural–language alignment, UniMind enables heterogeneous EEG tasks to share a single backbone, generative objective, and autoregressive decoder, without relying on task-specific prediction heads.

On ten scalp EEG datasets spanning five task domains, UniMind outperforms NeuroLM by 8.94 percentage points in average balanced accuracy and achieves strong cross-subject and cross-dataset performance. These results indicate that LLM-derived semantic priors facilitate more generalizable EEG representations that transfer to unseen data. The learned neural–language space also benefits intracortical spike decoding in non-human primates, raising average R^2 and yielding more sepa-

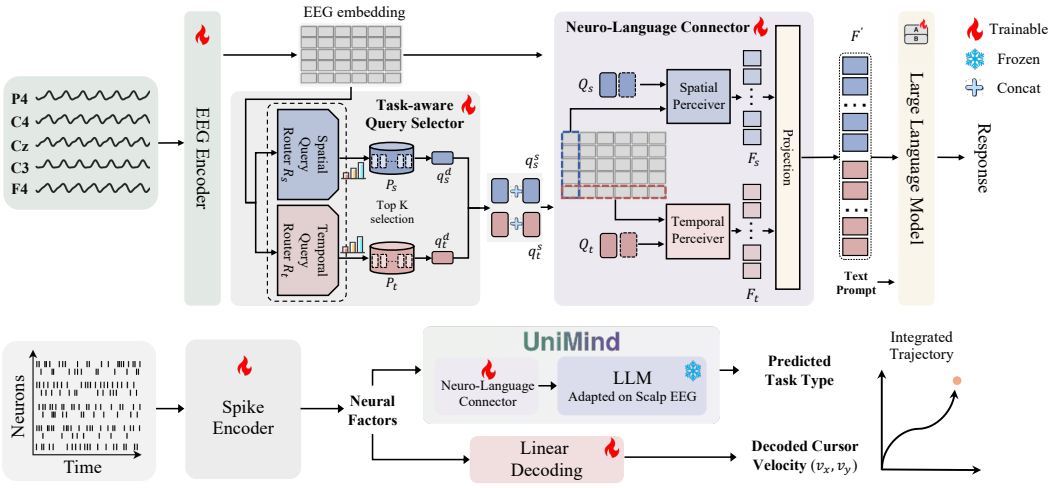


Figure 2: Overview of the UniMind architecture. For scalp EEG decoding, the Task-aware Query Selector dynamically selects spatial and temporal queries, and the Neuro-Language Connector distills task-relevant neural patterns into representations interpreted by the LLM. For intracortical decoding, the EEG-adapted LLM provides an auxiliary semantic objective to regularize latent representations for kinematic regression.

rable spike representations. Together, these findings provide preliminary evidence that UniMind captures latent regularities that generalize across subjects, datasets, and recording modalities. Our contributions are as follows:

- We propose UniMind, a unified multi-task neural decoding framework that formulates heterogeneous BCI tasks as conditional language generation with a single autoregressive objective and no task-specific heads.
- We design a Task-aware Query Selector that dynamically selects task-relevant queries from shared spatial and temporal pools, and a Neuro-Language Connector that distills compact spatiotemporal patterns and maps them into the LLM embedding space.
- Across ten scalp EEG datasets, UniMind outperforms multi-task and task-specific baselines on average and generalizes across unseen datasets and recording modalities.

2 Related Work

2.1 Neural Signal Foundation Models

Large-scale pretraining and representation learning have been actively explored for EEG, with related efforts extending to intracortical and MEG recordings. For EEG, BENDR (Kostas et al., 2021) applies masked contrastive prediction to latent representations extracted from raw waveforms, while BIOT (Yang et al., 2023a) introduces frequency-domain channel tokenization to enable contrastive pretraining across EEG and ECG. Masked prediction has also become a prevalent approach. LaBraM (Jiang et al., 2024) learns a vector-quantized neural tokenizer and predicts discrete neural codes for masked EEG channel patches, whereas CBraMod (Wang et al., 2025) reconstructs masked signal patches using a criss-cross Transformer that separately models temporal and spatial dependencies. Autoregressive objectives provide another means of capturing neural dynamics. Neuro-GPT (Cui et al., 2024) causally reconstructs masked latent EEG segments, while BrainGPT (Yue et al., 2024) performs electrode-wise autoregressive prediction to model temporal dependencies within individual electrodes. Beyond EEG, similar efforts have been developed for intracortical population activity (Ye & Pandarinath, 2021; Azabou et al., 2023) and MEG (Défossez et al., 2023).

Despite differences in recording modality, these models typically follow a common downstream paradigm: a pretrained encoder is paired with a task-specific prediction head and finetuned independently for each task. This independent adaptation keeps heterogeneous tasks isolated, limiting

explicit knowledge sharing across tasks during downstream learning. UniMind instead formulates heterogeneous decoding tasks as natural–language question–answer problems within a shared language space. This formulation makes UniMind *complementary* to neural signal foundation models by connecting transferable neural representations to a common semantic space, enabling joint learning and knowledge sharing across heterogeneous tasks through a shared backbone without task-specific prediction heads.

2.2 LLM-BASED NEURAL DECODING

An emerging line of work leverages pretrained language models for neural decoding. Early studies mainly use pretrained language representations as semantic priors for cross-modal alignment. Wang & Ji (2022) project word-level EEG features into BART for open-vocabulary EEG-to-text generation, while DeWave (Duan et al., 2023) further combines discrete EEG representations with EEG-text contrastive alignment and a pretrained language decoder to translate continuous EEG recordings without word-level event markers. This paradigm has also been extended to other recording modalities. BIT (Zhang et al., 2026) maps intracortical representations into an audio-LLM for speech decoding. Recent work has expanded language models beyond static alignment targets. NeuroLM (Jiang et al., 2025) converts EEG signals into vector-quantized neural tokens and performs multi-task instruction tuning with an LLM. However, existing methods primarily bridge neural signals and language through fixed projections or discrete tokenization, resulting in suboptimal performance and limited generalization. In contrast, UniMind dynamically selects task-relevant spatiotemporal features via shared query pools, thereby enabling knowledge sharing and generalization beyond seen datasets.

3 METHOD

As illustrated in Figure 2, UniMind is a unified neural decoding framework that reformulates heterogeneous EEG tasks as question–answer problems, enabling them to share a single backbone and generative objective without task-specific prediction heads. Specifically, a Neural Encoder first transforms raw multi-channel signals into token sequences, from which a Task-aware Query Selector (TQS) dynamically retrieves task-relevant queries. These queries are then fed into a Neuro-Language Connector (NLC), which distills spatiotemporal patterns via dual-branch cross-attention and projects them into the LLM input space. Finally, an LLM decodes all tasks through a unified autoregressive objective.

3.1 NEURAL ENCODER

Given a neural signal segment $X \in \mathbb{R}^{C\times S}$, where C and S denote the number of recording channels and temporal samples, respectively, the neural encoder transforms segments with varying channel configurations and lengths into structured token sequences. Each channel is partitioned into non-overlapping temporal patches of length t, producing $T = \left\lfloor \frac{S}{t} \right\rfloor$ patches per channel. The resulting patches are encoded into a sequence of patch token embeddings, with a learnable class token prepended, yielding $\boldsymbol{E} \in \mathbb{R}^{(C\times T+1)\times D_{\mathrm{e}}}$, where D_{e} denotes the hidden dimension of the encoder. Any pretrained signal encoder that produces per-patch embeddings can serve this role. In this work, we use the LaBraM encoder for EEG and a GRU for intracortical recordings.

3.2 TASK-AWARE QUERY SELECTOR

Given the encoded EEG representation $\boldsymbol{E}$, the Task-aware Query Selector (TQS) constructs queries for downstream neural feature extraction. A unified decoder must handle heterogeneous EEG tasks with distinct spatiotemporal patterns, where fixed shared queries may cause task interference. To address this, TQS dynamically selects input-adaptive queries while preserving static task-specific information shared across samples.

Specifically, TQS maintains a learnable query pool $\boldsymbol{P} = [P_1, P_2, \cdots, P_{n_q}]^\top \in \mathbb{R}^{n_q\times D_{\mathrm{e}}}$ and a query router $\mathcal{R}$ for query selection. It consists of two parallel branches to capture complementary spatial and temporal neural patterns, with their query pools and routers denoted by $\boldsymbol{P}_s, \mathcal{R}_s$ and $\boldsymbol{P}_t, \mathcal{R}_t$, respectively. Given the input EEG representation $\boldsymbol{E}$, each router $\mathcal{R}_i$ computes relevance

scores for all candidate queries in the corresponding pool $\boldsymbol{P}_i$ and selects the top-scoring subset. The dynamically selected queries are obtained as

$$q_i^d = \text{TopK}(\boldsymbol{P}_i; \mathcal{R}_i(\boldsymbol{E}), n_i) \in \mathbb{R}^{n_i \times D_e}, \quad i \in \{s, t\}, \tag{1}$$

where s and t denote the spatial and temporal branches, and n_s, n_t are the numbers of selected queries in each branch. Besides dynamic queries, each branch incorporates a static query token $q_i^s \in \mathbb{R}^{1 \times D_e}$ to capture stable task characteristics shared across inputs. The final query token sequence can be formulated as $Q_i = [q_i^d; q_i^s] \in \mathbb{R}^{(n_i+1) \times D_e}, i \in \{s, t\}$. The resulting queries Q_s and Q_t are then passed to the Neuro-Language Connector for cross-modal alignment.

3.3 Neuro-Language Connector

Given the task-relevant queries Q_s and Q_t from TQS, the Neuro-Language Connector (NLC) extracts compact spatiotemporal patterns from the encoded representation via dual-branch cross-attention. The spatial branch models cross-channel dependencies, while the temporal branch captures time-varying dynamics. In both branches, the selected queries serve as Queries and the corresponding EEG embeddings serve as Keys and Values. Formally:

$$\boldsymbol{F}_s = \text{CrossAttn}(Q_s, \boldsymbol{E}_s) \in \mathbb{R}^{(n_s+1) \times T \times D_e}, \tag{2}$$

$$\boldsymbol{F}_t = \text{CrossAttn}(Q_t, \boldsymbol{E}_t) \in \mathbb{R}^{(n_t+1) \times C \times D_e}, \tag{3}$$

where $\boldsymbol{E}_s \in \mathbb{R}^{T \times C \times D_e}$ and $\boldsymbol{E}_t \in \mathbb{R}^{C \times T \times D_e}$ denote the spatial and temporal views of the patch-level EEG embeddings, respectively, excluding the class token. The resulting $\boldsymbol{F}_s$ and $\boldsymbol{F}_t$, which encode spatially and temporally aggregated EEG features, are then concatenated with the class-token $\boldsymbol{F}_{\text{cls}} \in \mathbb{R}^{1 \times D_e}$ to form the combined feature sequence:

$$\boldsymbol{F} = \text{Concat}(\boldsymbol{F}_s, \boldsymbol{F}_t, \boldsymbol{F}_{\text{cls}}) \in \mathbb{R}^{((n_s+1) \times T + (n_t+1) \times C + 1) \times D_e}. \tag{4}$$

Finally, a linear projection maps $\boldsymbol{F}$ into the LLM input space:

$$\boldsymbol{F}' = \boldsymbol{F}W \in \mathbb{R}^{((n_s+1) \times T + (n_t+1) \times C + 1) \times D_L}, \tag{5}$$

where D_L denotes the LLM hidden size. By transforming spatiotemporal neural patterns into semantically structured representations, the NLC helps the LLM interpret neural signals based on their underlying neurophysiological characteristics.

3.4 Multi-task Decoding with LLM

The projected features $\boldsymbol{F}'$ are then passed to the LLM for multi-task decoding. We apply UniMind to two settings.

Scalp EEG Decoding. To support unified training across heterogeneous EEG decoding objectives, we construct a multi-task instruction tuning dataset containing 1.3 million samples from five EEG task types. Each original task label is converted into a natural-language answer, and for each task type we design ten semantically equivalent instruction templates that describe the corresponding decoding objective. During training, one template is randomly sampled for each EEG instance and paired with its textual answer. This diversification encourages the model to associate task semantics with neural patterns rather than memorize fixed instruction forms. Further details regarding the instruction templates are provided in Appendix B.

During multi-task training, the neural encoder, TQS, and NLC are jointly trained to adapt to diverse downstream tasks. The LLM is finetuned using LoRA (Hu et al., 2022) to enable parameter-efficient adaptation while largely preserving its pretrained semantic structure. All EEG decoding tasks are formulated as conditional language modeling. Given the neural-semantic features $\boldsymbol{F}'$, a separator token [SEP], and a task instruction P, we construct the input as $X = [\boldsymbol{F}', [\text{SEP}], P]$. The model predicts the target answer $Y = \{y_1, \ldots, y_L\}$ and is optimized using the causal cross-entropy loss:

$$\mathcal{L}_{\text{LM}} = -\sum_{i=1}^{L} \log p(y_i \mid X, y_{<i}). \tag{6}$$

Table 1: Comparison of multi-task models on ten EEG datasets. B.Acc. and W.F1 denote balanced accuracy and weighted F1, respectively. The best result in each column is highlighted in bold.

Model	Metric	Multi-subject		Cross-subject								Average
		SEEDIV	SEEDV	SHU	Siena	TUAB	TUEV	TUSL	HMC	SleepEDF	Workload	
NeuroLM	B. Acc.	36.35	33.98	54.49	62.50	79.69	46.79	68.45	67.37	67.16	64.45	58.12
	W. F1	43.89	34.93	53.86	96.84	78.93	73.59	68.39	71.26	86.84	66.57	67.51
UniMind-1.7B	B. Acc.	**46.07**	**38.69**	**59.75**	**83.19**	**80.25**	**62.44**	**82.81**	**74.06**	**69.96**	**73.33**	**67.06**
	W. F1	**48.01**	**39.22**	**59.70**	**98.55**	**80.49**	**76.12**	**83.89**	**75.85**	**87.88**	**73.16**	**72.29**

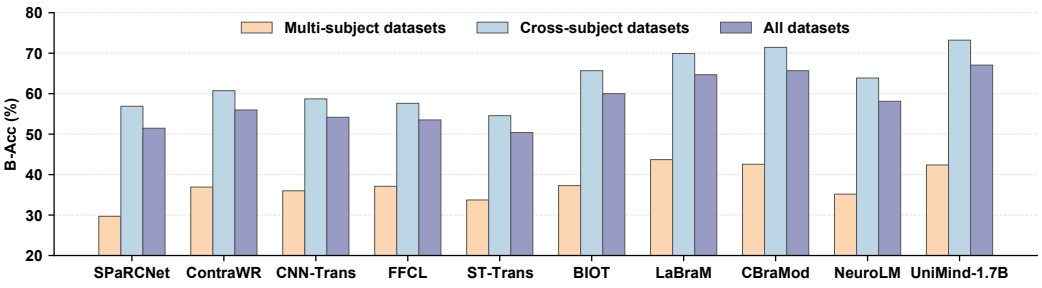


Figure 3: Comparison with task-specific and multi-task models. Results are averaged over datasets under the multi-subject, cross-subject, and combined settings.

Intracortical Spike Decoding. Multi-task EEG training shapes the LLM input space to accommodate neural signal representations. We repurpose this adapted space for intracortical spike decoding. A GRU decoder processes spike firing rates along the temporal dimension, and the NLC projects the GRU output into the LLM input space. The LLM is frozen, and the GRU and NLC are jointly trained on two kinematic tasks (center-out reaching and random-target reaching). Since only two tasks are available from a single subject, TQS is not used. Each trial is paired with a task instruction and a textual label describing the movement category. The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}} + \lambda \mathcal{L}_{\text{LM}}, \tag{7}$$

where $\mathcal{L}_{\text{reg}}$ is the kinematic regression loss (predicting 2D cursor velocity at each time bin) and $\mathcal{L}_{\text{LM}}$ (Eq. 6) backpropagates through the NLC and GRU only. λ controls the weight of the language modeling loss, which regularizes GRU representations toward the LLM input space while preserving the primary regression target. Further details are provided in Appendix F.

4 Experiments

4.1 Experimental Setup

Datasets & Metrics. We evaluate UniMind on ten publicly available EEG datasets across five task domains: emotion recognition (SEEDIV (Zheng et al., 2019), SEEDV (Liu et al., 2022)), motor imagery classification (SHU (Ma et al., 2022)), clinical anomaly detection (Siena (Goldberger et al., 2000; Detti et al., 2020), TUAB (Obeid & Picone, 2016), TUEV (Obeid & Picone, 2016), TUSL (Obeid & Picone, 2016)), sleep stage classification (HMC (Alvarez-Estevez & Rijsman, 2021), SleepEDF (Kemp et al., 2000)), and cognitive workload classification (Workload (Zyma et al., 2019)). For intracortical decoding, we use 96-channel Utah array recordings from one macaque (subject Mihili) performing two motor reaching tasks (Farshchian et al., 2019), with 2D hand velocity as the decoding target (details in Appendix F.1).

Following prior work (Jiang et al., 2024; Wang et al., 2025), the SEED series use multi-subject splits, and the remaining eight datasets use cross-subject splits with non-overlapping subjects. We report balanced accuracy and weighted F1 as primary metrics.

Compared Methods. We compare UniMind with both multi-task and task-specific EEG decoding methods. For multi-task comparison, we consider the B, L, and XL variants of NeuroLM (Jiang

Table 2: Cross-dataset generalization on the unseen SHHS dataset.

Dataset	Model	B. Acc.	Kappa	W. F1
SHHS	LaBraM	58.24	0.5289	65.46
	CBraMod	55.85	0.5264	65.44
	UniMind	**65.59**	**0.6472**	**74.16**

Table 3: Ablation of NLC and TQS components. The baseline directly projects EEG embeddings into the LLM input space using an MLP. Balanced accuracy (%) is reported.

Model	SEEDIV	SEEDV	SHU	Siena	TUAB	TUEV	TUSL	HMC	SleepEDF	Workload	Average
Baseline	45.77	38.84	55.95	77.65	80.14	57.67	78.83	72.83	68.59	67.33	64.36
+NLC	45.99	**39.43**	57.90	79.23	79.32	**64.98**	81.24	73.07	68.87	71.33	66.14
+NLC+TQS	**46.07**	38.69	**59.75**	**83.19**	**80.25**	62.44	**82.81**	**74.06**	**69.96**	**73.33**	**67.06**

et al., 2025) and report the best-performing variant for each dataset. For task-specific comparison, we consider representative EEG decoding models, including SPaRCNet (Jing et al., 2023), ContraWR (Yang et al., 2023b), CNN-Transformer (CNN-Trans) (Peh et al., 2022), FFCL (Li et al., 2022), ST-Transformer (Song et al., 2021), BIOT (Yang et al., 2023a), LaBraM (Jiang et al., 2024), and CBraMod (Wang et al., 2025). Notably, direct comparison inherently favors the task-specific baselines, as they are separately finetuned and optimized for each target dataset, whereas UniMind uses a single unified model across all tasks.

Implementation Details. We use pretrained LaBraM as the neural encoder and Qwen3-1.7B (Yang et al., 2025) as the LLM backbone. Additional architectural details, training configurations, and computational cost analysis are provided in Appendix C.

4.2 Comparison with State-of-the-Art Methods

Comparison with Multi-Task Models. Table 1 compares UniMind with NeuroLM under the same multi-task setting. UniMind consistently outperforms NeuroLM on all ten datasets, with an average improvement of 8.94 percentage points in balanced accuracy. The gains are larger on datasets with more classes or fewer samples. On SEED-IV (4 classes), SEED-V (5 classes), and TUEV (6 classes), balanced accuracy improves by 9.72, 4.71, and 15.65 percentage points, suggesting that the NLC enables the LLM to better exploit spatiotemporal neural patterns for multi-class discrimination. On the small-scale TUSL and Workload datasets (approximately 1k samples), UniMind achieves gains of 14.36 and 8.88 percentage points, as TQS facilitates knowledge sharing from data-rich to data-scarce tasks through shared query pools. Additionally, we evaluate UniMind with a larger Qwen3-4B backbone and observe consistent improvements; full results are reported in Appendix D.

Comparison with Task-Specific Models. We further compare UniMind with task-specific baselines, including both specialized models and pretrained EEG foundation models independently finetuned on each dataset. This per-dataset optimization inherently favors these baselines. Despite this, UniMind achieves the best average balanced accuracy across all datasets with a single unified model. As shown in Figure 3, in the cross-subject setting, UniMind outperforms all baselines, suggesting that joint multi-task training yields more robust cross-subject generalization. In the multi-subject setting, UniMind performs competitively with the best task-specific models.

Benefit of Joint Multi-task Training. To investigate whether joint multi-task training benefits individual tasks, we compare UniMind trained on all ten datasets jointly against training on each dataset separately, with all other settings kept identical. As shown in Figure 4, joint training improves balanced accuracy on nine out of ten datasets. The gains are most pronounced on data-scarce datasets such as TUSL and Workload, as well as datasets with related tasks such as the emotion and clinical groups, where cross-task knowledge transfer effectively compensates for limited training samples and reinforces shared neural patterns. These results confirm that UniMind's unified formulation enables effective knowledge sharing across heterogeneous EEG tasks.

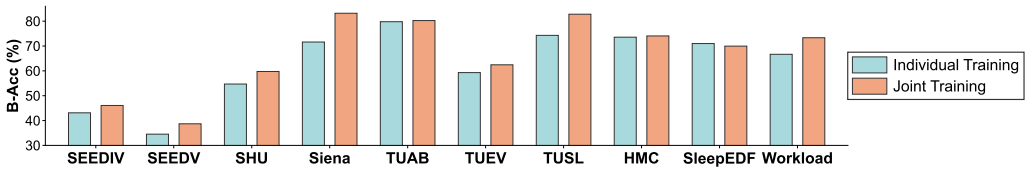


Figure 4: Comparison of individual training and joint multi-task training across ten datasets.

Table 4: Intracortical decoding performance on held-out sessions. R^2 is reported for center-out (CO) and random-target (RT) tasks. LM Loss uses a generic pretrained LLM, whereas UniMind Loss uses the LLM adapted through Scalp EEG training.

Method	**CO R^2**	**RT R^2**
Baseline GRU	0.7505	0.4759
Baseline GRU + LM Loss	NaN†	
Baseline GRU + UniMind Loss	0.7029	0.6607
Δ R^2	−0.0476	+0.1848
Avg. improvement	+0.0686 (**+16.24%**)	

† Training diverged. Model did not converge.

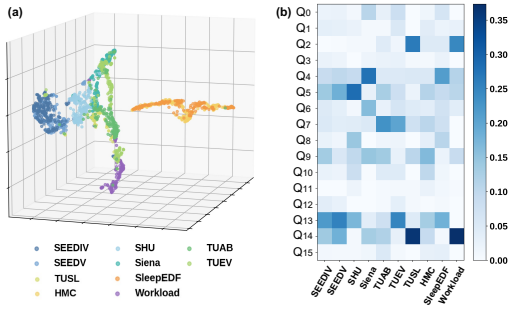


Figure 5: Visualization of EEG representations and task-aware query routing. (a) UMAP projection of pre-LLM EEG representations across datasets. (b) Dataset-wise activation frequencies of task-aware queries.

Cross-Dataset Generalization. We further test UniMind on SHHS (Quan et al., 1997), a sleep staging dataset not included in training. All models are trained on HMC and SleepEDF and directly evaluated on SHHS without target-domain finetuning. For LaBraM and CBraMod, a single shared five-class classification head is jointly trained on the source datasets and directly transferred to SHHS. We disable the static query in TQS to avoid dataset-specific priors. As shown in Table 2, UniMind consistently outperforms both baselines, exceeding LaBraM by 7.35 points in balanced accuracy and 8.70 points in weighted F1. These results suggest that task-aware feature extraction coupled with a shared language decoding space yields more generalizable representations.

Extension to Intracortical Decoding. We explore the benefit of the UniMind LLM input space for intracortical spike decoding. We freeze the UniMind LLM and introduce $\mathcal{L}_{\text{LM}}$ as an auxiliary objective to regularize a baseline GRU decoder for non-human primate kinematic decoding on two tasks: center-out reaching (CO) and random-target reaching (RT). No finetuning is performed on the test sessions, which are recorded on different days from all training sessions. As shown in Table 4, aligning with a pretrained LLM that has not been adapted through EEG training leads to training divergence. In contrast, aligning with the UniMind LLM yields a mean relative R^2 improvement of 16.24% across the two tasks, with large gains on the more challenging RT task. The CO task shows a slight decrease, likely reflecting a trade-off between in-distribution fit and generalization to diverse movement patterns. These results suggest that the neural-semantic space shaped by UniMind generalizes beyond human scalp EEG to invasive neural recordings.

4.3 Ablation Study

Analysis of Key Components. We conduct an incremental ablation study to evaluate the contributions of key components. The baseline directly projects the EEG embeddings into the LLM embedding space using an MLP. Unless otherwise noted, all ablation experiments are conducted with the UniMind-1.7B configuration hereafter. As shown in Table 3, introducing the NLC improves the average accuracy from 64.36% to 66.14%, with particularly large gains on TUEV (+7.31 points) and Workload (+4.00 points). Incorporating the TQS further increases the overall performance.

Analysis of Query Types and Configuration. We further analyze query design choices in TQS. Combining static and dynamic queries yields the best performance across all five evaluated datasets,

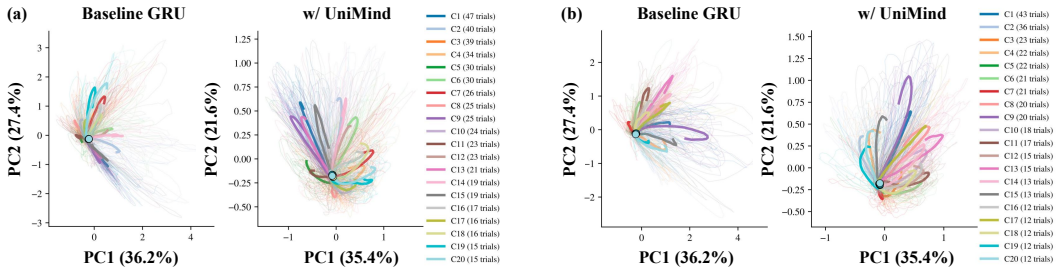


Figure 6: Two-dimensional PCA projections of GRU latent factors for the RT task. (a) and (b) correspond to two held-out test sessions. Left panels show the baseline GRU. Right panels show the GRU with UniMind Loss. Trials are grouped by velocity-based behavioral clustering; colors denote cluster identity; bold lines are trial-averaged trajectories.

indicating that the two query mechanisms provide complementary information. For TQS hyperparameters, average performance peaks at a pool size of n_q=16 with n_s=2 spatial and n_t=1 temporal selected queries; both too few and too many queries degrade accuracy. Detailed results are provided in Appendix E.

Comparison with Contrastive Alignment Paradigm. A central design choice of UniMind is to perform EEG decoding within the LLM's generative space, rather than merely aligning EEG with text embeddings. To validate this, we construct a contrastive EEG-text baseline following EEG-CLIP (Camaret Ndir et al., 2025). Specifically, the LaBraM encoder is finetuned end-to-end to align EEG features with pre-extracted text embeddings using the SoftCLIP loss (Gao et al., 2024). UniMind consistently outperforms both LaBraM and the contrastive baseline (Table 5), highlighting advantages of question-conditioned decoding over static text-embedding alignment. By integrating neural embeddings with task questions under a shared autoregressive objective, UniMind enables task semantics to guide the interpretation of task-relevant neural patterns in the LLM generative space, leading to more effective decoding.

Table 5: Contrastive alignment comparison. Balanced accuracy (%) is reported.

Method	HMC	TUSL
LaBraM	72.86	76.25
Contrastive	72.16	68.11
UniMind	**74.06**	**82.81**

4.4 Qualitative Analysis

Scalp EEG Representation Visualization. As shown in Fig. 5(a), the UMAP projection forms distinct clusters corresponding to each dataset, with datasets from the same domain positioned close to one another, such as emotion, sleep, and TUH clinical datasets. Fig. 5(b) shows a consistent pattern in query routing, where related tasks tend to activate overlapping queries. For example, the emotion datasets commonly select Q_5, Q_{13}, and Q_{14}, the sleep datasets share Q_{13}, and TUAB and TUEV both emphasize Q_7. These results indicate that TQS captures both domain-specific neural patterns and shared knowledge across related tasks.

Intracortical Neural Manifold Visualization. Figure 6 visualizes PCA projections of the GRU latent factors for the RT task on two held-out test sets, where trials are grouped by velocity-based behavioral clustering (details in Appendix F.6). The baseline decoder produces overlapping trajectories, while the decoder trained with UniMind Loss yields more separable trajectory clusters. This is consistent with the R^2 improvements in Table 4.

5 Conclusion

We introduce UniMind, a unified framework for multi-task brain decoding. Extensive experiments demonstrate that UniMind consistently outperforms both multi-task and task-specific baselines across ten EEG benchmarks, exhibits strong zero-shot generalization to an unseen dataset, and yields more separable neural representations that benefit intracortical spike decoding. Promising directions for future work include joint training across non-invasive and invasive recording modalities and scaling to larger, more diverse subject populations.

AI USE STATEMENT

AI-assisted tools were employed solely for proofreading and refining the language of this manuscript. All suggested revisions were manually reviewed and approved by the authors. The authors bear full responsibility for the final content, including all scientific claims, experimental results, and presented artifacts.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide detailed descriptions of all datasets, preprocessing steps, evaluation metrics, and data splits in Appendix A. Architectural details, training configurations, and hyperparameter settings are documented in Appendix C. All datasets used in this work are publicly available, and we will release the source code upon acceptance.

REFERENCES

Khald Ali I Aboalayon, Miad Faezipour, Wafaa S Almuhammadi, and Saeid Moslehpour. Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation. *Entropy*, 18(9):272, 2016.

Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLOS ONE*, 16(8):e0256111, 2021.

Syed Umar Amin, Mansour Alsulaiman, Ghulam Muhammad, Mohamed Amine Mekhtiche, and M Shamim Hossain. Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion. *Future Generation Computer Systems*, 101:542–554, 2019.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44937–44956, 2023.

Poomipat Boonyakitanont, Apiwat Lek-Uthai, Krisnachai Chomtho, and Jitkomut Songsiri. A review of feature extraction and performance evaluation in epileptic seizure detection using EEG. *Biomedical Signal Processing and Control*, 57:101702, 2020.

Tidiane Camaret Ndir, Robin T Schirrmeister, and Tonio Ball. EEG-CLIP: Learning EEG representations from natural language descriptions. *Frontiers in Robotics and AI*, 12:1625731, 2025.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-GPT: Towards a foundation model for EEG. In *2024 IEEE International Symposium on Biomedical Imaging*, pp. 1–5. IEEE, 2024.

Muhammad Najam Dar, Muhammad Usman Akram, Sajid Gul Khawaja, and Amit N Pujari. CNN and LSTM-based emotion charting using physiological signals. *Sensors*, 20(16):4551, 2020.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, 2023.

Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8(7):846, 2020. ISSN 2227-9717. doi: 10.3390/pr8070846. URL https://www.mdpi.com/2227-9717/8/7/846.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. DeWave: Discrete encoding of EEG waves for EEG to text translation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9907–9918, 2023.

A Farshchian, JA Gallego, JP Cohen, Y Bengio, LE Miller, and SA Solla. Adversarial domain adaptation for stable brain-machine interfaces. In *International Conference on Learning Representations*, 2019.

Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. SoftCLIP: Softer cross-modal alignment makes CLIP stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1860–1868, 2024.

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. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Leigh R Hochberg, Mijail D Serruya, Gerhard M Friehs, Jon A Mukand, Maryam Saleh, Abraham H Caplan, Almut Branner, David Chen, Richard D Penn, and John P Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, 2006.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Vinay Jayaram, Morteza Alamgir, Yasemin Altun, Bernhard Scholkopf, and Moritz Grosse-Wentrup. Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *International Conference on Learning Representations*, 2025.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100 (17):e1750–e1762, 2023.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Youngjoo Kim, Jiwoo Ryu, Ko Keun Kim, Clive C Took, Danilo P Mandic, and Cheolsoo Park. Motor imagery classification using mu and beta rhythms of EEG with strong uncorrelating transform based complex common spatial patterns. *Computational Intelligence and Neuroscience*, 2016(1): 1489692, 2016.

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

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

Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. *Biomedical Signal Processing and Control*, 72:103342, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2022.

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, 2022.

Dennis J McFarland, Laurie A Miner, Theresa M Vaughan, and Jonathan R Wolpaw. Mu and beta rhythm topographies during motor imagery and actual movements. *Brain Topography*, 12(3): 177–186, 2000.

Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023.

Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp EEG. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3599–3602. IEEE, 2022.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019.

Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'Connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. ChronoNet: A deep recurrent neural network for abnormal EEG identification. In *Conference on Artificial Intelligence in Medicine in Europe*, pp. 47–56. Springer, 2019.

Simanto Saha and Mathias Baumert. Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review. *Frontiers in Computational Neuroscience*, 13:87, 2020.

Lakhan Dev Sharma, Vijay Kumar Bohat, Maria Habib, Ala'M Al-Zoubi, Hossam Faris, and Ibrahim Aljarah. Evolutionary inspired approach for mental stress detection using EEG signal. *Expert Systems with Applications*, 197:116634, 2022.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for EEG decoding. *arXiv preprint arXiv:2106.11170*, 2021.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5350–5358, 2022.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

Dongrui Wu, Yifan Xu, and Bao-Liang Lu. Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19, 2022.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chaoqi Yang, M Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260, 2023a.

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

Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with Neural Data Transformers. *Neurons, Behavior, Data Analysis, and Theory*, 5(3):1–18, 2021.

Tongtian Yue, Xuange Gao, Shuning Xue, Yepeng Tang, Longteng Guo, Jie Jiang, and Jing Liu. BrainGPT: Unleashing the potential of EEG generalist foundation model by autoregressive pre-training. *arXiv preprint arXiv:2410.19779*, 2024.

Yizi Zhang, Linyang He, Chaofei Fan, Tingkai Liu, Han Yu, Trung Le, Jingyuan Li, Scott Linderman, Lea Duncker, Francis R Willett, et al. A cross-species neural foundation model for end-to-end speech decoding. In *International Conference on Learning Representations*, 2026.

Wei-Long Zheng and Bao-Liang Lu. A multimodal approach to estimating vigilance using EEG and forehead EOG. *Journal of Neural Engineering*, 14(2):026017, 2017. URL `http://stacks.iop.org/1741-2552/14/i=2/a=026017`.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. EmotionMeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49 (3):1110–1122, 2019.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

APPENDIX CONTENTS

A Dataset Details and Evaluation Metrics 15

A.1 Dataset Overview 15
A.2 Data Splits 15
A.3 Data Preprocessing 16
A.4 Evaluation Metrics 16

B Instruction Templates 17

C Implementation Details 17

C.1 Architectural Details 17
C.2 Training Configurations 19
C.3 Computational Cost 19

D Complete Quantitative Results 19

D.1 Baseline Models 19
D.2 Detailed Results 21
D.3 Regression Results 21

E Additional Ablation Results 22

E.1 Analysis of Query Types 22
E.2 Effect of Query Pool Size 22
E.3 Effect of Selected Query Numbers 23

F Intracortical Decoding Details 23

F.1 Dataset and Experimental Setup 23
F.2 Instruction-Tuning Format 25
F.3 Evaluation Metric 25
F.4 Baseline Architecture 25
F.5 Training Details 25
F.6 Behavioral Clustering and Manifold Analysis 26

A Dataset Details and Evaluation Metrics

A.1 Dataset Overview

As shown in Table S1, to comprehensively evaluate UniMind, we utilize ten publicly available EEG datasets spanning five representative task domains. (1) For **emotion recognition**, we adopt SEEDIV (Zheng et al., 2019) and SEEDV (Liu et al., 2022), both containing multi-session EEG recordings collected while subjects viewed emotion-eliciting video clips. SEEDIV includes four emotion categories (happy, sad, fear, and neutral), while SEEDV extends to five (happy, sad, fear, disgust, and neutral). (2) For **motor imagery classification**, we incorporate SHU (Ma et al., 2022), a large-scale dataset comprising EEG signals recorded during imagined left- and right-hand movements, enabling binary decoding of motor intentions for brain–computer interface applications. (3) For **clinical anomaly detection**, we use Siena (Goldberger et al., 2000; Detti et al., 2020), TUAB (Obeid & Picone, 2016), TUEV (Obeid & Picone, 2016), and TUSL (Obeid & Picone, 2016). Siena targets binary seizure detection by distinguishing ictal from non-ictal EEG segments; TUAB focuses on identifying abnormal EEG recordings; TUEV classifies six types of clinical events, such as epileptiform discharges and eye movements; and TUSL categorizes seizure, slowing, and complex background activity. (4) For **sleep stage classification**, we employ HMC (Alvarez-Estevez & Rijsman, 2021) and SleepEDF (Kemp et al., 2000). These datasets provide long-term polysomnographic EEG recordings with expert annotations, where each segment is labeled as one of five standard sleep stages: Wake, N1, N2, N3, and REM. (5) For **cognitive workload classification**, we utilize the Workload dataset (Zyma et al., 2019), which contains EEG recordings acquired during mental arithmetic tasks designed to elicit varying levels of cognitive load.

Table S1: Overview of downstream EEG datasets and tasks used for evaluation.

Task	Dataset	Rate (Hz)	#Channels	Duration	#Subjects	Label
Emotion Recognition	SEEDIV	1000	62	1 s	15	4-class
	SEEDV	1000	62	1 s	16	5-class
Motor Imagery	SHU	250	32	4 s	25	Binary
Clinical Abnormality Detection	Siena	512	29	10 s	14	Binary
	TUAB	256	23	10 s	2,383	Binary
	TUEV	250	23	5 s	370	6-class
	TUSL	250	23	10 s	28	3-class
Sleep Classification	HMC	256	4	30 s	151	5-class
	SleepEDF	100	2	30 s	78	5-class
Workload Classification	Workload	500	19	4 s	36	Binary

A.2 Data Splits

Each dataset is partitioned into training, validation, and test sets following task-specific strategies. (1) **SEEDIV and SEEDV**: Trials within each session are split chronologically. SEEDIV contains 24 trials per session, divided into 16 for training, 4 for validation, and 4 for testing; SEEDV contains 15 trials per session, equally divided into three consecutive parts (5:5:5) for training, validation, and testing. For both datasets, all sessions are subsequently merged within each partition. (2) **SHU**: A subject-independent protocol is adopted, with subjects 1-15 allocated for training, 16-20 for validation, and 21-25 for testing. (3) **Siena**: Raw recordings are segmented into 10-second non-overlapping samples, yielding 51,307 samples in total. A subject-independent split is employed: data from the last two subjects (PN16 and PN17) are reserved as the test set, while the remaining 12 subjects' data are partitioned into training and validation sets at an 8:2 ratio. (4) **TUAB and TUEV**: The official training/test partition is preserved, with the training subjects further divided into 80% for training and 20% for validation. (5) **TUSL**: The dataset is randomly partitioned into 60% for training, 20% for validation, and 20% for testing. (6) **HMC**: The first 100 subjects constitute the training set, the middle 25 the validation set, and the remaining 26 the test set. (7) **SleepEDF**: Subjects are randomly assigned to training, validation, and test sets at an 8:1:1 ratio. (8) **Workload**: Subjects 0-25 are designated for training, 26-30 for validation, and 31-35 for testing.

A.3 Data Preprocessing

We develop a standardized preprocessing pipeline comprising band-pass filtering, notch filtering, resampling, and normalization, designed to be universally applicable across diverse BCI tasks and decoding models. Following established practice, we apply a 0.1-75 Hz band-pass filter to retain task-relevant frequency components while suppressing low-frequency drift and high-frequency noise. To remove power-line interference, we first perform a Fast Fourier Transform (FFT) on the raw EEG signals to identify the dominant interference frequency, and then apply a notch filter at either 50 Hz or 60 Hz accordingly. All EEG signals are subsequently resampled to a uniform rate of 200 Hz. To mitigate the effects of low-amplitude signals and enhance numerical stability during model optimization, we apply z-score normalization to most datasets. For the SHU dataset, we instead adopt 95th-percentile normalization to better accommodate its signal distribution characteristics.

A.4 Evaluation Metrics

Due to the inherent class imbalance in EEG datasets, we evaluate model performance using three primary metrics.

(1) Balanced Accuracy is defined as the average recall obtained on each class. By equally weighting the recall of all classes, it effectively mitigates the bias caused by imbalanced class distributions, ensuring that minority classes are not overlooked during evaluation:

$$\text{Balanced Accuracy} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}, \tag{S1}$$

where C denotes the number of classes, and TP_i and FN_i denote the numbers of true positives and false negatives for class i, respectively.

(2) Cohen's Kappa measures the level of agreement between the predicted labels and the true labels, while correcting for the agreement that could occur by random chance. This statistic provides a more robust assessment of classification performance than simple accuracy, especially in datasets where class distribution is uneven:

$$\text{Cohen's Kappa} = \frac{p_o - p_e}{1 - p_e}, \tag{S2}$$

where p_o denotes the observed agreement between the predicted and true labels, and p_e denotes the agreement expected by chance.

(3) Weighted F1 Score is the harmonic mean of precision and recall calculated for each class individually and then averaged by weighting each class's score according to its support (i.e., the number of true instances). This metric balances false positives and false negatives while accounting for class imbalance, providing a comprehensive measure of overall model effectiveness:

$$\text{Weighted F1} = \sum_{i=1}^{C} w_i \cdot \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}, \tag{S3}$$

where w_i denotes the normalized support of class i and is defined as

$$w_i = \frac{n_i}{\sum_{j=1}^{C} n_j}, \tag{S4}$$

with n_i denoting the number of true instances belonging to class i. The class-wise precision and recall are defined as

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \quad \text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \tag{S5}$$

where FP_i denotes the number of false positives for class i.

Since our model produces discrete class labels directly from LLM outputs without associated probability scores, metrics that require confidence values such as **AUROC** (Area Under the Receiver Operating Characteristic Curve) and **AUC-PR** (Area Under the Precision-Recall Curve) are not applicable. Therefore, **Weighted F1** and **Cohen's Kappa** are prioritized for both binary and multi-class classification tasks to ensure reliable evaluation under these constraints.

B Instruction Templates

UniMind is trained with instruction tuning. To enhance instruction diversity, we design ten distinct prompt templates per dataset. During training, each sample is randomly paired with one of the corresponding task-specific instructions.

Instruction Templates for the SEEDIV Dataset

1. Which emotion category is represented by this EEG segment? [Neutral, Sad, Fear, Happy]
2. What emotional state does this EEG segment reflect? [. . .]
3. Based on this EEG segment, which emotion label is most likely? [. . .]
4. . . . (similar instructions)

Instruction Templates for the SEEDV Dataset

1. Given this EEG signal, which emotion does it reflect? [Disgust, Fear, Sad, Neutral, Happy]
2. Which emotion category is represented by this EEG segment? [. . .]
3. How should the emotion represented by this EEG segment be classified? [. . .]
4. . . . (similar instructions)

Instruction Templates for the SHU Dataset

1. Which hand motor imagery class is represented by this EEG segment? [Left hand, Right hand]
2. What imagined hand movement does this EEG segment reflect? [. . .]
3. Based on this EEG segment, which hand motor imagery task is most likely? [. . .]
4. How should the hand motor imagery in this EEG segment be classified? [. . .]
5. . . . (similar instructions)

C Implementation Details

C.1 Architectural Details

We use pretrained LaBraM (Jiang et al., 2024) as the neural encoder and Qwen3-1.7B and Qwen3-4B (Yang et al., 2025) as the LLM backbones. For the Task-aware Query Selector (TQS), the spatial and temporal branches each maintain a pool of 16 learnable queries. The spatial branch selects two dynamic queries and the temporal branch selects one, with an additional static query appended in each branch. Each query router is implemented as a two-layer MLP. For the Neuro-Language Connector (NLC), each branch comprises an eight-head cross-attention layer followed by a two-layer MLP. A linear projection then maps the concatenated spatial, temporal, and class-token representations to the LLM embedding dimension. During multi-task training, the neural encoder, TQS, and NLC are jointly optimized, while the LLM is adapted via LoRA (Hu et al., 2022).

Instruction Templates for the Siena Dataset

1. Does this EEG segment contain seizure activity? [Non-seizure, Seizure]
2. Is a seizure present in this EEG segment? [. . .]
3. Based on this EEG segment, is seizure activity present or absent? [. . .]
4. How should this EEG segment be classified with respect to seizure activity? [. . .]
5. . . . (similar instructions)

Instruction Templates for the TUAB Dataset

1. Does this segment come from a normal or abnormal EEG recording? [Normal, Abnormal]
2. Is the EEG recording from which this segment was taken normal or abnormal? [. . .]
3. Based on this EEG segment, does the source recording appear normal or abnormal? [. . .]
4. How should the source EEG recording be classified using this segment? [. . .]
5. . . . (similar instructions)

Instruction Templates for the TUEV Dataset

1. What does this EEG signal suggest about the current epileptic state of the patient? [Spike and sharp wave, Generalized periodic epileptiform discharges, Periodic lateralized epileptiform discharges, Eye movement, Artifact, Background]
2. Which epileptic state category best describes this EEG signal? [. . .]
3. Based on this EEG signal, which epileptic state label is most likely? [. . .]
4. How should the epileptic activity in this EEG signal be classified? [. . .]
5. . . . (similar instructions)

Instruction Templates for the TUSL Dataset

1. Which state category best describes this EEG signal? [Seizure, Slowing, Complex background]
2. Which clinical EEG pattern category is represented by this segment? [. . .]
3. Based on this EEG segment, which pattern label is most likely? [. . .]
4. How should this EEG segment be categorized based on its dominant clinical pattern? [. . .]
5. . . . (similar instructions)

Instruction Templates for the HMC and SleepEDF Datasets

1. Which sleep stage is represented by this EEG epoch? [W, N1, N2, N3, REM]
2. What sleep stage does this EEG segment reflect? [. . .]
3. Based on this EEG epoch, which sleep stage is most likely? [. . .]
4. How should the sleep stage of this EEG epoch be classified? [. . .]
5. . . . (similar instructions)

Instruction Templates for the Workload Dataset

1. Which mental workload level is represented by this EEG segment? [High, Low]
2. What level of mental workload does this EEG signal reflect? [. . .]
3. Based on this EEG segment, which workload label is most likely? [. . .]
4. How should the mental workload in this EEG segment be classified? [. . .]
5. . . . (similar instructions)

C.2 Training Configurations

Experiments are conducted on a single machine with 8 NVIDIA GPUs, using Python 3.11.14, PyTorch 2.6.0, and CUDA 12.4. Distributed training is performed via DeepSpeed ZeRO Stage 2. The model is trained for 10 epochs with the AdamW optimizer and a cosine learning rate schedule. Detailed configurations are provided in Table S2.

C.3 Computational Cost

The trainable ratio denotes the proportion of trainable parameters relative to the total parameter count of the complete model. Trainable components include the EEG encoder, TQS, NLC, and LoRA adapters; the language model backbone remains frozen throughout training. Training time refers to the wall-clock duration for 10 epochs on eight NVIDIA GPUs. As shown in Table S3, the parameter-efficient adaptation strategy keeps the trainable ratio below 13% for both configurations, with the Qwen3-1.7B variant completing training in under 7 hours.

D Complete Quantitative Results

D.1 Baseline Models

- **SPaRCNet** (Jing et al., 2023) is a 1D CNN with dense residual connections, incorporating adaptive depth and channel selection mechanisms for EEG decoding.
- **ContraWR** (Yang et al., 2023b) converts EEG signals into multi-channel spectrograms and classifies them using a ResNet-style 2D CNN.
- **CNN-Transformer** (CNN-Trans) (Peh et al., 2022) augments a CNN backbone with transformer layers and employs belief matching (BM) loss for channel-wise EEG artifact detection.
- **FFCL** (Li et al., 2022) integrates parallel CNN and LSTM branches to extract spatial and temporal features, respectively, which are fused via a fully connected layer for classification.
- **ST-Transformer** (ST-Trans) (Song et al., 2021) applies attention mechanisms along both feature-channel and segmented temporal dimensions to capture global spatiotemporal dependencies from EEG signals.

Table S2: Key training and model configuration parameters.

Parameter	Value
Training Epochs	10
Batch Size	256
Learning Rate	4×10^{-5}
Weight Decay	0.01
Warmup Ratio	0.05
LR Scheduler	Cosine decay
Dataloader Workers	8
BFloat16 Precision	True
Gradient Checkpointing	True
DeepSpeed Strategy	ZeRO Stage 2
EEG Encoder	LaBraM-Base
EEG Patch Size	200
EEG Hidden Size	200
Number of EEG Encoder Layers	12
Number of EEG Attention Heads	10
Dynamic Query Pool Size	16
Spatial / Temporal Top-K	2 / 1
Static Spatial / Temporal Queries	1 / 1
Language Model	Qwen3-1.7B
LLM Hidden Size	2048
Number of LLM Layers	28
Number of LLM Attention Heads	16
Number of LLM KV Heads	8
Language Model	Qwen3-4B
LLM Hidden Size	2560
Number of LLM Layers	36
Number of LLM Attention Heads	32
Number of LLM KV Heads	8

Table S3: Computational cost of different language model backbones.

Backbone	Trainable Ratio (%)	Time (hours)
Qwen3-1.7B	12.29	6.56
Qwen3-4B	7.26	8.70

- **BIOT** (Yang et al., 2023a) is a linear transformer for universal biosignal representation learning via hybrid supervised-unsupervised pretraining. Since BIOT accepts a maximum of 18 EEG channels, a 1×1 convolution is applied to project higher-dimensional inputs into an 18-channel format.
- **LaBraM** (Jiang et al., 2024) is an EEG foundation model that learns generic neural representations through masked token prediction with a full-attention transformer.
- **CBraMod** (Wang et al., 2025) is an EEG foundation model that addresses the heterogeneous spatiotemporal structure of EEG signals via a criss-cross transformer backbone.

D.2 Detailed Results

Tables S5–S9 present the complete results for Balanced Accuracy, Cohen's Kappa, and Weighted F1 (reported in %) across all ten datasets. Each experiment is repeated with different random seeds, and we report the mean and standard deviation for each metric. The performance trends across these metrics are consistent with those discussed in the main text.

D.3 Regression Results

To extend UniMind to continuous-valued prediction, we introduce a regression readout token `<reg>` paired with a lightweight regression head. Given an EEG signal $X \in \mathbb{R}^{C \times S}$, the neural encoder first extracts spatiotemporal representations, which are subsequently processed by the TQS and NLC and projected into the language model hidden space. The input sequence for a regression sample is structured as:

```
User:  <eeg> + regression instruction
Assistant:  <reg>
```

The model performs a single forward pass and extracts the last-layer hidden state at the `<reg>` position as the regression representation $\boldsymbol{h}_{\text{reg}} \in \mathbb{R}^d$. Because `<reg>` is positioned after both the EEG tokens and the task instruction, causal self-attention allows $\boldsymbol{h}_{\text{reg}}$ to incorporate information from the neural signal and the task semantics. A two-layer MLP then maps this representation to a scalar prediction $\hat{z}$, optimized using the mean squared error:

$$\mathcal{L}_{\text{reg}} = \frac{1}{B} \sum_{i=1}^{B} (\hat{z}_i - z_i)^2 , \tag{S6}$$

where z_i denotes the ground-truth regression target and B is the batch size.

For regression samples, all token-level language modeling labels are set to the ignore index, such that the loss is computed entirely from the regression head. This design directly predicts continuous values from contextualized hidden representations, avoiding the uncertainty associated with numerical text generation. During training, the presence of `<reg>` also serves as a routing signal: samples containing this token are processed by the regression branch, whereas classification samples follow the standard autoregressive language modeling branch. This formulation enables classification and regression tasks to be jointly optimized within a single model while sharing the neural encoder, TQS, NLC, and language model backbone.

We evaluate this regression formulation on the SEEDVIG dataset (Zheng & Lu, 2017) for continuous vigilance estimation. SEEDVIG simulates a real-world driving environment in which participants gradually transition from an alert state to a fatigued state while operating a simulated driving system. Continuous vigilance levels are quantified using the PERCLOS indicator obtained from SMI eye-tracking glasses. The dataset contains EEG recordings from 23 subjects, acquired using 17 channels at a sampling rate of 200 Hz. The recordings are segmented into 20,355 non-overlapping 8-second samples. We adopt a subject-independent evaluation protocol, using subjects 1-15 for training, subjects 16-19 for validation, and subjects 20-23 for testing.

Regression performance is evaluated using the Pearson correlation coefficient (PCC), R^2, and root mean squared error (RMSE), as reported in Table S4.

Table S4: Performance comparison on the regression task. **Bold** and underlined values denote the best and second-best results, respectively.

Method	PCC↑	R^2 ↑	RMSE↓
SPaRCNet	0.5709	0.2185	0.2806
ContraWR	0.5235	0.0727	0.3057
CNN-Trans.	0.5829	0.1796	0.2877
FFCL	0.4923	0.1740	0.2885
ST-Trans.	0.6020	0.1138	0.2983
BIOT	0.6114	0.1232	0.2971
LaBraM	**0.6347**	0.1808	0.2871
CBraMod	0.5502	0.0737	0.3057
UniMind-1.7B	0.6176	**0.2990**	**0.2659**
UniMind-4B	0.6242	0.2258	0.2795

Table S5: Performance comparison on SEEDIV and SEEDV. "MT" denotes multi-task training. Results are reported as mean ± standard deviation. **Bold** values denote the best overall performance, and underlined the best among multi-task models.

Model	MT	SEEDIV			SEEDV		
		B. Acc	Kappa	W. F1	B. Acc	Kappa	W. F1
SPaRCNet	×	29.88 (± 1.37)	6.45 (± 0.92)	32.05 (± 1.48)	29.49 (± 0.78)	11.21 (± 1.39)	29.79 (± 0.83)
ContraWR	×	38.38 (± 1.61)	14.83 (± 1.26)	40.21 (± 1.43)	35.46 (± 1.05)	19.05 (± 1.88)	35.44 (± 1.21)
CNN-Trans.	×	35.21 (± 1.42)	10.92 (± 1.17)	36.57 (± 1.39)	36.78 (± 0.78)	20.72 (± 1.83)	36.42 (± 0.88)
FFCL	×	37.81 (± 1.55)	15.35 (± 1.44)	39.76 (± 1.61)	36.41 (± 0.92)	20.78 (± 2.01)	36.45 (± 1.32)
ST-Trans.	×	36.93 (± 1.46)	15.72 (± 1.30)	36.95 (± 1.50)	30.52 (± 0.72)	10.83 (± 1.21)	28.33 (± 1.05)
BIOT	×	36.19 (± 1.38)	23.21 (± 1.47)	42.76 (± 1.52)	38.37 (± 1.87)	22.61 (± 2.62)	38.56 (± 2.03)
LaBraM	×	**47.63** (± 0.73)	**28.77** (± 0.60)	**49.14** (± 0.82)	39.76 (± 1.38)	23.86 (± 2.09)	39.74 (± 1.11)
CBraMod	×	44.20 (± 0.19)	24.17 (± 0.20)	45.58 (± 0.22)	**40.91** (± 0.97)	25.69 (± 1.43)	41.01 (± 1.08)
NeuroLM	✓	36.35 (± 0.14)	18.07 (± 1.46)	43.89 (± 2.98)	33.98 (± 0.96)	17.65 (± 1.22)	34.93 (± 1.36)
UniMind-1.7B	✓	46.07 (± 0.35)	26.22 (± 0.39)	48.01 (± 0.28)	38.69 (± 0.27)	23.22 (± 0.36)	39.22 (± 0.29)
UniMind-4B	✓	46.14 (± 0.23)	25.80 (± 0.22)	46.91 (± 0.17)	40.82 (± 0.03)	**26.28** (± 0.02)	**41.52** (± 0.02)

E Additional Ablation Results

This section presents additional ablation experiments on the Task-aware Query Selector (TQS). Unless otherwise noted, all experiments use the UniMind-1.7B configuration.

E.1 Analysis of Query Types

We compare static queries, dynamic queries, and their combination in the TQS. Static queries encode dataset-specific priors, whereas dynamic queries adapt query selection to each EEG input. As shown in Table S10, combining both query types achieves the best performance across all five datasets. These results indicate that the two query mechanisms provide complementary information for task-relevant feature extraction.

E.2 Effect of Query Pool Size

We analyze the effect of the query pool size n_q on model performance. The spatial and temporal query pools share the same size. As shown in Table S11, average performance peaks at $n_q = 16$. A small pool forces different tasks to select similar queries, while an overly large pool reduces beneficial sharing among related tasks.

Table S6: Performance comparison on SHU and Siena. "MT" denotes multi-task training. Results are reported as mean ± standard deviation. **Bold** values denote the best overall performance, and underlined the best among multi-task models.

Model	MT	SHU			Siena		
		B. Acc	Kappa	W. F1	B. Acc	Kappa	W. F1
SPaRCNet	×	58.70 (± 1.37)	17.39 (± 2.73)	58.03 (± 0.84)	71.15 (± 2.62)	56.26 (± 2.85)	97.58 (± 0.11)
ContraWR	×	58.85 (± 0.42)	17.67 (± 0.85)	56.60 (± 2.22)	65.10 (± 7.20)	37.24 (± 14.32)	96.69 (± 1.37)
CNN-Trans.	×	57.74 (± 1.60)	15.46 (± 2.19)	56.90 (± 3.63)	77.27 (± 4.63)	56.12 (± 3.49)	97.27 (± 0.17)
FFCL	×	58.63 (± 1.38)	17.25 (± 2.79)	56.41 (± 4.99)	68.72 (± 1.84)	53.03 (± 5.80)	97.46 (± 0.58)
ST-Trans.	×	60.35 (± 0.84)	20.69 (± 1.68)	59.98 (± 0.99)	80.98 (± 0.62)	69.21 (± 9.00)	98.16 (± 0.93)
BIOT	×	59.49 (± 0.65)	18.99 (± 1.30)	59.06 (± 0.79)	76.75 (± 1.95)	61.47 (± 4.75)	97.71 (± 0.40)
LaBraM	×	61.66 (± 1.57)	22.70 (± 2.12)	60.73 (± 1.51)	70.82 (± 2.86)	58.02 (± 7.75)	97.71 (± 0.83)
CBraMod	×	**63.70** (± 0.74)	**25.62** (± 1.47)	**61.66** (± 0.63)	73.17 (± 0.44)	60.37 (± 8.12)	97.77 (± 0.66)
NeuroLM	✓	54.49 (± 0.12)	8.99 (± 0.24)	53.86 (± 1.85)	62.50 (± 1.90)	39.21 (± 1.34)	96.84 (± 0.06)
UniMind-1.7B	✓	59.75 (± 0.67)	19.49 (± 1.34)	59.70 (± 0.68)	83.19 (± 0.63)	75.51 (± 0.99)	98.55 (± 0.06)
UniMind-4B	✓	55.76 (± 0.30)	11.51 (± 0.59)	55.48 (± 0.33)	**86.03** (± 0.27)	**77.34** (± 1.26)	**98.62** (± 0.09)

Table S7: Performance comparison on TUAB and TUEV. "MT" denotes multi-task training. Results are reported as mean ± standard deviation. **Bold** values denote the best overall performance, and underlined the best among multi-task models.

Model	MT	TUAB			TUEV		
		B. Acc	Kappa	W. F1	B. Acc	Kappa	W. F1
SPaRCNet	×	78.69 (± 0.47)	50.67 (± 0.29)	75.13 (± 0.80)	41.61 (± 2.62)	42.33 (± 1.81)	70.24 (± 1.04)
ContraWR	×	80.17 (± 0.58)	61.02 (± 0.38)	80.65 (± 0.42)	43.84 (± 3.49)	39.12 (± 2.37)	68.93 (± 1.36)
CNN-Trans.	×	79.53 (± 0.94)	57.33 (± 0.63)	78.76 (± 0.31)	40.87 (± 1.61)	38.15 (± 1.34)	68.54 (± 2.93)
FFCL	×	78.19 (± 0.22)	55.81 (± 0.84)	77.83 (± 0.75)	39.79 (± 1.04)	37.32 (± 1.88)	67.83 (± 1.20)
ST-Trans.	×	79.66 (± 0.31)	61.69 (± 0.45)	80.90 (± 0.93)	39.84 (± 2.28)	37.65 (± 3.06)	68.23 (± 1.90)
BIOT	×	79.59 (± 0.89)	59.42 (± 0.13)	78.82 (± 0.90)	52.81 (± 2.25)	52.73 (± 2.49)	74.92 (± 0.82)
LaBraM	×	**81.40** (± 0.20)	62.52 (± 0.12)	81.47 (± 0.57)	64.09 (± 0.65)	66.37 (± 0.93)	83.12 (± 0.52)
CBraMod	×	79.79 (± 0.22)	**63.02** (± 0.28)	**85.29** (± 0.26)	**66.71** (± 1.07)	**67.72** (± 0.96)	**83.42** (± 0.64)
NeuroLM	✓	79.69 (± 0.53)	54.61 (± 0.98)	78.93 (± 0.91)	46.79 (± 3.56)	45.70 (± 4.98)	73.59 (± 2.19)
UniMind-1.7B	✓	80.25 (± 0.02)	60.71 (± 0.04)	80.49 (± 0.02)	62.44 (± 0.08)	54.10 (± 0.08)	76.12 (± 0.05)
UniMind-4B	✓	80.50 (± 0.02)	61.26 (± 0.05)	80.77 (± 0.02)	63.60 (± 0.17)	56.28 (± 0.08)	77.20 (± 0.04)

E.3 Effect of Selected Query Numbers

We analyze the effect of the numbers of selected spatial and temporal queries, n_s and n_t. As shown in Table S12, average performance peaks at $n_s = 2$ and $n_t = 1$. Too few queries restrict the diversity of captured EEG patterns while too many introduce irrelevant information.

F Intracortical Decoding Details

F.1 Dataset and Experimental Setup

We use intracortical spike recordings from one non-human primate (macaque, subject Mihili) performing two motor reaching tasks (Farshchian et al., 2019). Neural activity was recorded with a 96-electrode Utah array implanted in the primary motor cortex (M1). After quality filtering, 95 neurons are retained. Spike counts are binned at 20 ms intervals to form firing rate matrices of shape (neurons × time bins) for each trial. The decoding target is the 2D hand velocity at each time bin.

The two tasks differ in their movement patterns.

- **Center-out reaching (CO).** The subject reaches from a fixed central position to one of eight peripheral targets, producing stereotyped straight trajectories. Each trial contains one discrete reach.

Table S8: Performance comparison on TUSL and Workload. "MT" denotes multi-task training. Results are reported as mean ± standard deviation. **Bold** values denote the best overall performance, and underlined the best among multi-task models.

Model	MT	TUSL			Workload		
		B. Acc	Kappa	W. F1	B. Acc	Kappa	W. F1
SPaRCNet	×	41.85 (± 4.52)	13.99 (± 7.99)	35.00 (± 9.68)	59.77 (± 3.67)	10.67 (± 2.42)	54.60 (± 4.11)
ContraWR	×	58.57 (± 6.62)	35.67 (± 9.68)	54.58 (± 7.98)	69.66 (± 4.33)	31.08 (± 3.81)	69.33 (± 3.78)
CNN-Trans.	×	35.75 (± 3.51)	3.06 (± 4.79)	22.35 (± 2.51)	57.93 (± 5.14)	8.37 (± 4.35)	58.63 (± 4.62)
FFCL	×	39.19 (± 6.88)	6.28 (± 8.88)	21.20 (± 7.86)	70.69 (± 2.46)	41.36 (± 3.12)	72.54 (± 2.93)
ST-Trans.	×	40.00 (± 3.29)	8.60 (± 4.49)	37.93 (± 4.59)	61.03 (± 1.87)	12.48 (± 2.33)	60.67 (± 1.52)
BIOT	×	57.58 (± 3.03)	20.12 (± 2.12)	23.94 (± 0.40)	66.55 (± 2.74)	30.67 (± 3.30)	51.66 (± 2.61)
LaBraM	×	76.25 (± 2.31)	64.07 (± 3.04)	76.14 (± 2.12)	66.09 (± 1.94)	29.81 (± 1.35)	64.72 (± 2.43)
CBraMod	×	73.88 (± 3.20)	61.49 (± 5.50)	74.53 (± 3.60)	77.50 (± 0.22)	**64.67** (± 0.28)	77.32 (± 0.25)
NeuroLM	✓	68.45 (± 3.04)	51.94 (± 4.61)	68.39 (± 2.97)	64.45 (± 2.25)	23.19 (± 1.74)	66.57 (± 1.92)
UniMind-1.7B	✓	**82.81** (± 1.78)	**74.44** (± 2.37)	**83.89** (± 1.52)	73.33 (± 1.02)	46.67 (± 2.04)	73.16 (± 1.01)
UniMind-4B	✓	82.51 (± 3.52)	72.35 (± 4.62)	82.50 (± 2.85)	**79.33** (± 0.77)	58.67 (± 1.54)	**79.26** (± 0.77)

Table S9: Performance comparison on HMC and SleepEDF. "MT" denotes multi-task training. Results are reported as mean ± standard deviation. **Bold** values denote the best overall performance, and underlined the best among multi-task models.

Model	MT	HMC			SleepEDF		
		B. Acc	Kappa	W. F1	B. Acc	Kappa	W. F1
SPaRCNet	×	47.56 (± 11.09)	31.47 (± 13.15)	41.08 (± 13.10)	60.16 (± 1.21)	65.31 (± 1.36)	58.61 (± 1.44)
ContraWR	×	42.42 (± 5.41)	23.40 (± 5.54)	29.87 (± 2.88)	67.05 (± 1.67)	70.46 (± 1.52)	66.92 (± 1.58)
CNN-Trans.	×	65.73 (± 1.41)	59.61 (± 1.05)	68.96 (± 0.65)	60.29 (± 1.35)	66.12 (± 1.21)	58.96 (± 1.30)
FFCL	×	44.27 (± 7.02)	25.42 (± 6.54)	29.02 (± 4.85)	65.66 (± 1.33)	70.13 (± 1.27)	64.79 (± 1.45)
ST-Trans.	×	25.59 (± 1.41)	5.03 (± 1.83)	14.28 (± 1.22)	55.17 (± 1.49)	67.32 (± 1.38)	53.18 (± 1.42)
BIOT	×	68.62 (± 0.41)	62.95 (± 1.13)	70.91 (± 1.47)	64.95 (± 0.26)	71.32 (± 0.32)	60.91 (± 0.40)
LaBraM	×	72.86 (± 1.01)	68.12 (± 0.73)	75.54 (± 0.24)	68.96 (± 0.75)	75.49 (± 0.68)	87.30 (± 0.71)
CBraMod	×	72.69 (± 0.41)	66.85 (± 1.04)	73.95 (± 0.89)	69.47 (± 0.20)	75.13 (± 0.26)	87.40 (± 0.24)
NeuroLM	✓	67.37 (± 0.50)	61.88 (± 0.57)	71.26 (± 0.34)	67.16 (± 0.60)	74.50 (± 0.59)	86.84 (± 0.63)
UniMind-1.7B	✓	**74.06** (± 0.17)	68.68 (± 0.21)	75.85 (± 0.17)	69.96 (± 0.19)	75.49 (± 0.22)	**87.88** (± 0.12)
UniMind-4B	✓	73.99 (± 0.09)	**68.83** (± 0.05)	**76.09** (± 0.03)	**71.37** (± 0.24)	**75.53** (± 0.07)	87.87 (± 0.04)

- **Random-target reaching (RT).** The subject reaches to randomly placed targets in sequence, producing diverse and unpredictable movement trajectories. Each trial contains one reach segment between consecutive targets.

Data from both tasks are used to jointly train a single GRU model. Session-level splits ensure no temporal overlap between training and test sets, and test sessions are recorded on different days from all training sessions. The detailed session allocation is shown in Table S13.

Trial lengths vary across sessions (CO mean ≈61 bins, RT mean ≈40 bins). Trials shorter than 10 bins are discarded. All trials are zero-padded to a fixed length of 80 bins before being fed to the model. During evaluation, only the time bins corresponding to the original (unpadded) trial length contribute to the R^2 computation.

Table S10: Ablation of static and dynamic queries in the Task-aware Query Selector. Balanced accuracy (%) is reported on five representative datasets.

Queries	SEEDIV	SHU	TUEV	HMC	Workload
Static	44.59	56.23	59.43	73.36	70.67
Dynamic	44.99	59.07	60.09	73.43	68.00
Combined	**46.07**	**59.75**	**62.44**	**74.06**	**73.33**

Table S11: Effect of the query pool size n_q on balanced accuracy (%). The spatial and temporal query pools have the same size.

n_q	SEEDIV	SHU	TUEV	HMC	AVG
8	43.72	56.30	58.45	73.26	57.93
16	**46.07**	**59.75**	**62.44**	**74.06**	**60.58**
32	43.51	55.34	59.92	73.61	58.10

Table S12: Effect of the numbers of selected spatial and temporal queries, n_s and n_t, on balanced accuracy (%).

n_s	n_t	SEEDIV	SHU	TUEV	HMC	AVG
1	1	42.68	54.63	**63.54**	73.02	58.47
1	2	44.13	58.92	61.02	**76.14**	60.05
2	1	**46.07**	**59.75**	62.44	74.06	**60.58**
2	2	45.63	55.98	61.43	75.79	59.71
4	4	43.34	57.36	61.70	75.97	59.59

F.2 Instruction-Tuning Format

Each trial is paired with a task instruction and a textual answer, following the same paradigm as the EEG experiments. The instruction template is:

A monkey is performing a motor reaching task. Given the neural population firing rates of subject {subject} with {num_neurons} neurons recorded, which task type is it? Pick one: [CO, RT].

The textual answer is the task label, i.e., center-out (CO) or random-target (RT). The frozen LLM receives the NLC-projected spike representations concatenated with the tokenized instruction and predicts the answer via causal cross-entropy loss $\mathcal{L}_{\text{LM}}$ (Eq. 6). This loss backpropagates through the NLC and GRU encoder, providing an additional training signal that regularizes the learned representations toward the adapted LLM input space.

F.3 Evaluation Metric

We report the coefficient of determination (R^2) between predicted and ground-truth 2D velocity as the decoding metric:

$$R^2 = 1 - \frac{\sum_{t=1}^{T}\sum_{d=1}^{2}(y_{t,d} - \hat{y}_{t,d})^2}{\sum_{t=1}^{T}\sum_{d=1}^{2}(y_{t,d} - \bar{y}_d)^2}, \tag{S7}$$

where $y_{t,d}$ and $\hat{y}_{t,d}$ are the ground-truth and predicted velocity at time bin t for dimension d, and $\bar{y}_d$ is the mean ground-truth velocity for dimension d across all time bins in the test set. $R^2 = 1$ indicates perfect prediction; $R^2 = 0$ corresponds to predicting the mean velocity.

F.4 Baseline Architecture

The baseline GRU decoder is a 2-layer bidirectional GRU with hidden dimension 128. It takes the spike firing rate matrix (95 neurons $\times$ 80 time bins) as input and processes it along the temporal dimension. At each time bin, the GRU hidden state is linearly projected to predict the 2D cursor velocity. This architecture follows standard practice for neural population decoding (Farshchian et al., 2019).

F.5 Training Details

A single GRU model is jointly trained on both CO and RT tasks. The total training loss combines a kinematic regression loss and the UniMind Loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}} + \lambda \cdot \mathcal{L}_{\text{LM}}, \tag{S8}$$

Table S13: **Cross-session experimental design for subject Mihili.** A single model is jointly trained on both reaching tasks, with test sessions held out at the session level.

Task	Split	Sessions	Trials
Center-Out (CO)	Training	M_20140203, M_20140211, M_20140217, M_20140218, M_20140225	1,066
	Test	**M_20140306, M_20140307**	**441**
Random-Target (RT)	Training	M_20131207, M_20131208, M_20140114, M_20140115, M_20140116, M_20140128, M_20140129, M_20140212, M_20140214	6,158
	Test	**M_20140222, M_20140224**	**1,296**

where $\mathcal{L}_{\text{reg}}$ is the mean squared error between predicted and ground-truth 2D velocity at each time bin, $\mathcal{L}_{\text{LM}}$ is the causal cross-entropy loss (Eq. 6) from the frozen LLM, and λ is a task-specific weight set to 1.0 for CO and 0.2 for RT, to compensate for the imbalanced trial counts between the two tasks.

For the "Baseline GRU" row in Table 4, only $\mathcal{L}_{\text{reg}}$ is used ($\lambda = 0$). For "Baseline GRU + UniMind Loss", the NLC projects GRU hidden states into the UniMind LLM input space and the full $\mathcal{L}_{\text{total}}$ is optimized. For "Baseline GRU + LM Loss", the alignment target is a generic pretrained LLM (without EEG finetuning); this variant diverges to NaN during training, suggesting that a pretrained LLM without neural signal adaptation does not provide useful regularization for this task.

Data augmentation is applied during training: channel dropout (rate 0.1), temporal jitter (width 2 bins), each applied with probability 0.3. Only two tasks are available from a single subject, so TQS is not used; the NLC operates with fixed queries.

F.6 Behavioral Clustering and Manifold Analysis

For the neural manifold visualization (Figure 6), we apply PCA to the GRU latent factors (128-dimensional hidden states) and project them to 2D. PCA is fit on all test trials per model variant, then applied to individual trials for plotting.

Since the random-target task has no predefined movement categories, we group trials into 20 behavioral clusters based on ground-truth velocity similarity. Specifically, we compute pairwise distances using temporal correlation and dynamic time warping, then apply agglomerative clustering with Ward linkage. Clusters with fewer than 3 trials are discarded. In the figures, individual trials are plotted with low opacity and the cluster-averaged trajectory is overlaid in bold.