SEMLOOP: CLOSING THE LOOP BETWEEN SEMANTIC RELEVANCE AND EEG FOUNDATION MODEL UTILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG foundation models (EEG FMs) learn transferable representations from heterogeneous neural data, yet adapting them to task-specific patterns remains challenging. Language offers a natural interface for incorporating task knowledge, but task-relevant semantics can vary substantially in their utility to the target model. Existing approaches largely rely on predefined semantics, leaving a gap between semantic relevance and model utility. We propose **SemLoop**, a closed-loop semantic adaptation framework that bridges this gap through model feedback. SemLoop comprises three mechanisms: **EEG-Grounded Semantic Alignment (EGSA)** grounds language-derived neurophysiological concepts in EEG representations; **Model-Validated Semantic Refinement (MVSR)** uses LLM-based evaluator and planner agents to iteratively refine semantics with target-model feedback; and **Experience-Internalized Semantic Planning (EISP)** internalizes successful refinements for subsequent planning. Across nine EEG benchmarks and multiple EEG FMs, SemLoop consistently outperforms full fine-tuning and predefined semantic-guided approaches. Further analyses reveal heterogeneous utility among task-relevant semantics and show that model-validated refinement shifts them toward higher model utility. The code is available at `https://anonymous.4open.science/r/SemLoop-75F0`.

1 INTRODUCTION

Large-scale self-supervised pre-training has enabled EEG foundation models (EEG FMs) to learn transferable representations from heterogeneous EEG data, providing a general backbone for diverse downstream tasks (Kuruppu et al., 2026; Lai et al., 2025; Wu et al., 2025). However, these representations may not fully capture the discriminative neural patterns required by specific tasks, making effective downstream adaptation essential (Li et al., 2026b; Wang et al., 2025b; Duan et al., 2023). Language offers a flexible and interpretable interface for introducing explicit task knowledge into this process, allowing task-relevant neurophysiological concepts to directly inform downstream adaptation (Jiang et al., 2025).

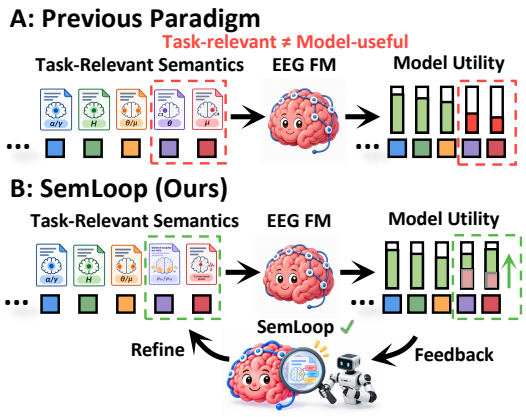

Figure 1: **Motivation of SemLoop.** (A) Existing approaches rely on predefined task-relevant semantics, which are not necessarily model-useful. (B) SemLoop closes this gap by iteratively refining semantics through model feedback.

Recent studies have explored this direction through EEG–text alignment, semantic fusion, and language-guided representation learning (Camaret Ndir et al., 2025; Ye et al., 2026; Cho et al., 2026). However, they primarily focus on incorporating predefined semantic knowledge into EEG models, while overlooking whether such knowledge is actually useful to the target model. In particular, **task-relevant semantics can vary substantially in their model utility**, making semantic relevance alone insufficient to deter-

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.

mine what knowledge should guide adaptation. More importantly, existing methods establish a path from knowledge to models, but lack the reverse path for model behavior to validate and refine that knowledge (Fig.1 (A)). This raises a key question: **Can model feedback identify and refine the semantics that are most useful for adapting EEG foundation models?**

To address this challenge, we propose **SemLoop**, a **ground–validate–refine–internalize** paradigm for closed-loop semantic adaptation of EEG foundation models. SemLoop operates through three mechanisms: **EEG-Grounded Semantic Alignment (EGSA)** grounds language-derived neurophysiological concepts in EEG representations; **Model-Validated Semantic Refinement (MVSR)** employs LLM-based evaluator and planner agents to validate and refine semantic knowledge through model feedback; and **Experience-Internalized Semantic Planning (EISP)** internalizes successful refinement experiences for subsequent semantic planning. Together, these mechanisms transform predefined, one-way semantic guidance into an adaptive closed-loop process (Fig.1 (B)). Our main contributions are summarized as follows:

- We propose **SemLoop**, a closed-loop semantic adaptation paradigm for EEG foundation models that bridges semantic relevance and model utility, moving beyond predefined, one-way semantic guidance.
- We develop **EEG-Grounded Semantic Alignment (EGSA)** and **Model-Validated Semantic Refinement (MVSR)** to ground neurophysiological semantics in EEG representations and employ LLM-based evaluator and planner agents to iteratively validate and refine them through model feedback.
- We further develop **Experience-Internalized Semantic Planning (EISP)** to internalize successful refinement experiences, enabling validated experiences to improve subsequent semantic planning.

2 Related Work

EEG Foundation Models and Adaptation. Large-scale self-supervised pre-training has given rise to EEG foundation models (FMs) that learn transferable representations from heterogeneous EEG corpora (Xiong et al., 2025). Representative models, including CBraMod (Wang et al., 2025a), LaBraM (Jiang et al., 2024), EEGPT (Wang et al., 2024), and BrainOmni (Xiao et al., 2025), have demonstrated broad generalization across diverse downstream tasks (Abiri et al., 2019; Aggarwal & Chugh, 2022; Lotte et al., 2018; Värbu et al., 2022). Yet, general-purpose representations do not eliminate the need for task specialization, as downstream tasks may emphasize neural patterns not explicitly captured during pre-training (Li et al., 2026b; Wang et al., 2025b; Duan et al., 2023). Existing studies have therefore explored full fine-tuning and parameter-efficient tuning to adapt EEG FMs to downstream tasks (Hu et al., 2021). More recently, language has opened a complementary paradigm for adaptation, providing an explicit interface for introducing task knowledge beyond what can be learned from neural signals alone (Jiang et al., 2025).

Semantic-Guided EEG Representation Learning. Building on this paradigm, existing approaches align EEG signals with textual representations or leverage task-related descriptions and neural concepts to guide representation learning and downstream prediction (Camaret Ndir et al., 2025; Ye et al., 2026; Cho et al., 2026). While these studies demonstrate the promise of semantic guidance, the incorporated semantics are typically predefined and remain disconnected from feedback about their utility to the target model. Consequently, whether task-relevant semantics are also model-useful remains underexplored. SemLoop closes this gap by grounding semantic guidance in target-model behavior, establishing a feedback loop in which the knowledge used for adaptation is iteratively validated and refined according to its model utility.

3 Method

3.1 Problem Formulation and Framework Overview

We consider downstream adaptation of a pre-trained EEG foundation model f_θ on a target dataset $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$. Task-relevant semantics can vary substantially in their utility to the target model, making semantic relevance alone insufficient for determining which semantics should guide

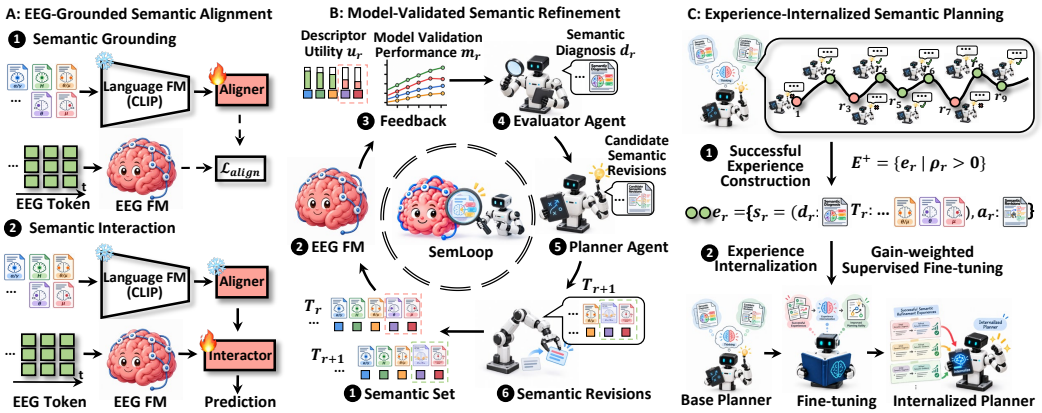

Figure 2: **Overview of SemLoop.** (A) **EEG-Grounded Semantic Alignment (EGSA)** grounds language-derived semantics in EEG representations and integrates them for downstream prediction. (B) **Model-Validated Semantic Refinement (MVSR)** closes the loop between semantic knowledge and the EEG FM, with evaluator and planner agents iteratively diagnosing and revising semantics through model feedback. (C) **Experience-Internalized Semantic Planning (EISP)** internalizes refinement experiences into the planner, enabling their reuse in subsequent semantic planning.

adaptation. We therefore aim to **identify and refine model-useful semantics through target-model feedback**. To this end, we propose **SemLoop**, a **ground–validate–refine–internalize** framework that closes the loop between semantic knowledge and model utility. As illustrated in Fig. 2, SemLoop comprises **EEG-Grounded Semantic Alignment (EGSA)**, **Model-Validated Semantic Refinement (MVSR)**, and **Experience-Internalized Semantic Planning (EISP)**.

3.2 EEG-Grounded Semantic Alignment (EGSA)

Language-derived neurophysiological semantics provide explicit task knowledge, yet are not inherently aligned with EEG representations. EGSA addresses this gap by grounding language semantics in the EEG representation space, enabling their direct interaction with EEG features.

Semantic Grounding. For each EEG sample $\mathbf{X}_i$, the frozen EEG foundation model f_θ produces $\mathbf{Z}_i = f_\theta(\mathbf{X}_i)$. We associate each EEG sample $\mathbf{X}_i$ with a set of EEG-observable neurophysiological descriptions S_i based on its characteristics. Each description $s^j \in \mathcal{S}_i$ is encoded by a frozen CLIP text encoder f_{text} (Radford et al., 2021) and projected into the EEG space through a lightweight Semantic Aligner g_ϕ: $\mathbf{b}^j = g_\phi\big(f_{\text{text}}(s^j)\big)$, where $\mathbf{b}^j$ denotes its EEG-grounded representation. We then optimize g_ϕ through bidirectional multi-positive contrastive alignment. For each EEG sample, all associated semantic descriptions are treated as positives, while the remaining semantics in the batch serve as negatives. Let $\mathcal{P}_i$ denote the indices of semantic descriptions associated with $\mathbf{X}_i$, and define $\ell_{ij} = \alpha\,\text{sim}(\mathbf{Z}_i, \mathbf{b}^j)$, where α is a learnable logit scale. The EEG-to-semantic objective is:

$$\mathcal{L}_{\text{E}\to\text{S}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{P}_i|}\sum_{j\in\mathcal{P}_i}\log\frac{\exp(\ell_{ij})}{\sum_k \exp(\ell_{ik})}. \tag{1}$$

The semantic-to-EEG objective $\mathcal{L}_{\text{S}\to\text{E}}$ is defined symmetrically, yielding

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\left(\mathcal{L}_{\text{E}\to\text{S}} + \mathcal{L}_{\text{S}\to\text{E}}\right). \tag{2}$$

This objective learns the Semantic Aligner that aligns semantics with EEG representations.

Semantic Interaction. We then map a set of task-level neurophysiological descriptors $T = \{t^k\}_{k=1}^K$ through the Semantic Aligner to construct a grounded semantic bank: $\mathbf{B} = \{\mathbf{b}^k\}_{k=1}^K$, where $\mathbf{b}^k = g_\phi\big(f_{\text{text}}(t^k)\big)$. Given EEG representation $\mathbf{Z}_i$, we propose a Semantic Interaction Module that integrates grounded semantics through cross-attention and a gated residual connection:

$$\widetilde{\mathbf{Z}}_i = \mathbf{Z}_i + \sigma(\gamma)\,\text{Attn}(\mathbf{Z}_i, \mathbf{B}), \tag{3}$$

where the cross-attention mechanism selectively attends to relevant semantics in $\mathbf{B}$, and $\sigma(\gamma)$ provides a learnable gate that controls the strength of semantic guidance. The resulting representation $\widetilde{\mathbf{Z}}_i$ is used for downstream prediction.

3.3 Model-Validated Semantic Refinement (MVSR)

Task-relevant semantics can vary substantially in their utility to the target EEG model. MVSR closes the loop between semantic knowledge and model behavior by evaluating descriptor-level utility and using model-grounded feedback to iteratively refine the semantic descriptor set.

Descriptor-Level Utility. At refinement round r, let $T_r = \{t_r^k\}_{k=1}^K$ denote the current semantic descriptor set, with the grounded semantic bank $\mathbf{B}_r = \{\mathbf{b}_r^k\}_{k=1}^K$. We estimate the utility of each descriptor t_r^k by measuring how removing its grounded representation $\mathbf{b}_r^k$ affects model prediction:

$$u_r^k = \frac{1}{N}\sum_{i=1}^{N}\left[\mathcal{L}_i\left(\mathbf{B}_r \setminus \{\mathbf{b}_r^k\}\right) - \mathcal{L}_i(\mathbf{B}_r)\right], \tag{4}$$

where $\mathcal{L}_i(\mathbf{B})$ denotes the prediction loss for $\mathbf{X}_i$ under semantic bank $\mathbf{B}$. A positive u_r^k indicates that removing t_r^k increases the prediction loss, suggesting that the descriptor is useful to the current model, whereas a small or negative value indicates limited or potentially adverse utility.

Agent-Guided Semantic Refinement. The resulting utilities provide model-grounded feedback for semantic refinement. An evaluator agent $\mathcal{A}_{\mathrm{E}}$ analyzes the current descriptors T_r, their utility scores u_r, and current validation performance m_r to produce a semantic diagnosis: $d_r = \mathcal{A}_{\mathrm{E}}(T_r, u_r, m_r)$. The diagnosis identifies which semantics should be preserved or revised and which semantics may be missing. Conditioned on this diagnosis, a planner agent $\mathcal{A}_P$ generates P candidate semantic revisions:

$$\left\{\widetilde{T}_{r+1}^{(p)}\right\}_{p=1}^{P} = \mathcal{A}_P(T_r, d_r). \tag{5}$$

Each candidate is grounded through the learned semantic projector and evaluated by the target EEG model. The best-performing candidate is selected as the next refinement state:

$$T_{r+1} = \arg\max_{\widetilde{T}_{r+1}^{(p)}} M\left(\widetilde{T}_{r+1}^{(p)}\right), \tag{6}$$

where $M(\cdot)$ denotes the validation performance of the target model. MVSR permits non-monotonic transitions to encourage semantic exploration, while retaining the historically best-performing semantic descriptor set for final prediction. We instantiate both agents with DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), and their complete prompt templates are provided in Appendix G.

3.4 Experience-Internalized Semantic Planning (EISP)

MVSR discovers effective semantic refinements through iterative model feedback, but repeated exploration limits efficient reuse of such experience. EISP internalizes successful refinements into the planner, turning feedback-driven exploration into reusable semantic planning capability.

Successful Experience Construction. We formulate each refinement transition as an experience $e_r = (\mathbf{s}_r, \mathbf{a}_r, \rho_r)$. The state $\mathbf{s}_r$ summarizes the current semantic descriptor set T_r and its model-grounded feedback, including the semantic diagnosis d_r, while the action $\mathbf{a}_r$ denotes the corresponding semantic revision. We quantify the effectiveness of each experience by its improvement in validation performance: $\rho_r = M(T_{r+1}) - M(T_r)$. We retain model-beneficial transitions with $\rho_r > 0$, yielding: $\mathcal{E}^+ = \{e_r \mid \rho_r > 0\}$.

Experience Internalization. We assign each successful experience a utility weight w_r proportional to its validation improvement ρ_r, emphasizing refinement behaviors that yield larger performance gains. The selected experiences are then internalized into the planner $\mathcal{A}_{\mathrm{P}}$ through gain-weighted supervised fine-tuning:

$$\mathcal{L}_{\mathrm{EISP}} = -\frac{1}{|\mathcal{E}^+|}\sum_{e_r \in \mathcal{E}^+} w_r \log \pi_\psi(\mathbf{a}_r \mid \mathbf{s}_r), \tag{7}$$

where $\pi_\psi(\mathbf{a} \mid \mathbf{s})$ denotes the planner's conditional distribution over semantic revisions, parameterized by trainable parameters ψ. This objective preferentially internalizes refinement behaviors that have been validated as beneficial by the target EEG model. Together, EGSA, MVSR, and EISP form a closed semantic adaptation cycle: semantics are grounded in EEG representations, validated and refined through model feedback, and internalized into the planner for subsequent adaptation. SemLoop thereby transforms one-way semantic guidance into an adaptive process jointly shaped by semantic knowledge and model utility.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate SemLoop on nine public EEG benchmarks spanning motor imagery classification (BCIC-2A (Brunner et al., 2008), BCIC-2B (Steyrl et al., 2016), SHU-MI (Ma et al., 2022)), emotion recognition (FACED (Chen et al., 2023), SEED-V (Liu et al., 2021)), mental disorder diagnosis (Mumtaz2016 (Wajid, 2016)), mental stress detection (MentalArithmetic (Zyma et al., 2019)), imagined speech classification (BCIC2020-3 (Jeong et al., 2022)), and event type classification (TUEV (Obeid & Picone, 2016)). These benchmarks cover diverse task semantics, subject populations, and recording conditions. (see Appendix A and Table 6 for details)

Baselines. We consider two groups of baselines. For standard EEG foundation model adaptation, we evaluate CBraMod (Wang et al., 2025a), LaBraM (Jiang et al., 2024), EEGPT (Wang et al., 2024), and BrainOmni (Xiao et al., 2025) with full fine-tuning. For semantic-guided adaptation, we compare against EEG-CLIP (Camaret Ndir et al., 2025), USEA (Ye et al., 2026), and BLPM (Cho et al., 2026) as representative approaches based on predefined semantic guidance, using the same backbone and evaluation protocol as SemLoop. (see Appendix B for details)

Backbone Models. We evaluate SemLoop with three representative EEG foundation models: CBraMod, LaBraM, and EEGPT. Their diverse architectures and pre-training objectives allow us to assess the generality of SemLoop across EEG foundation models. CBraMod serves as the default backbone unless otherwise specified. (see Appendix C for details)

Evaluation Metrics and Implementation Details. We use Balanced Accuracy (Bal.Acc.), AUROC, and AUC-PR for binary tasks, and Bal.Acc., Cohen's Kappa (Kappa), and weighted F1-score (F1-w) for multi-class tasks. AUROC and Kappa are used for model selection, respectively. We report mean $\pm$ standard deviation over five random seeds. (see Appendix D-F for details)

4.2 MAIN RESULTS

Setup. For a fair comparison, SemLoop and all predefined semantic baselines use the same CBraMod backbone. **Results.** Table 1 presents the overall performance across nine EEG benchmarks. SemLoop achieves superior overall performance over standard EEG foundation model adaptation and existing semantic-guided approaches. The improvements are statistically significant under the two-sided Wilcoxon signed-rank test ($p < 0.001$), conducted on paired results across datasets (Pratt, 1959; Liu et al., 2022; Li et al., 2026a). **Analysis.** Two key observations emerge from the results. (i) The improvements over standard EEG foundation model adaptation demonstrate the effectiveness of introducing semantic guidance into downstream adaptation. (ii) More importantly, SemLoop outperforms approaches based on predefined semantic guidance under the same backbone, suggesting that task relevance alone may not identify the semantics most useful to the target model. These findings highlight the importance of validating and refining semantic knowledge through target-model feedback rather than relying solely on predefined semantics.

4.3 GENERALIZATION ACROSS EEG FOUNDATION MODELS

Setup. We further evaluate SemLoop on LaBraM and EEGPT, comparing it with full fine-tuning and predefined semantic baselines under the same adaptation protocol. **Results.** As shown in Table 2, SemLoop achieves consistent improvements across different EEG foundation model backbones, yielding better overall performance than competing methods on both LaBraM and EEGPT. **Analysis.** The consistent gains across different backbones demonstrate that SemLoop is not depen-

Table 1: Performance comparison with baselines under full fine-tuning. All semantic-guided methods use CBraMod as the backbone.

Methods	BCIC-2A, 4-class			BCIC-2B, 2-class			SHU-MI, 2-class		
	Bal.Acc.	Kappa	F1-w	Bal.Acc.	AUC-PR	AUROC	Bal.Acc.	AUC-PR	AUROC
EEGPT	36.11 ± 1.24	14.81 ± 1.65	31.17 ± 2.70	62.19 ± 5.85	67.15 ± 8.82	68.76 ± 7.83	52.65 ± 1.42	54.58 ± 0.97	54.51 ± 1.38
BrainOmni	38.84 ± 1.45	18.45 ± 1.93	37.12 ± 1.75	61.32 ± 0.89	64.84 ± 1.04	65.79 ± 0.85	60.43 ± 2.32	66.71 ± 1.14	67.27 ± 0.72
LaBraM	44.88 ± 4.29	26.50 ± 5.72	43.76 ± 4.61	63.27 ± 2.62	68.18 ± 2.05	70.21 ± 2.14	60.80 ± 1.61	67.36 ± 2.16	65.88 ± 2.80
CBraMod	46.77 ± 1.35	29.03 ± 1.80	44.83 ± 1.95	65.12 ± 1.41	68.64 ± 1.62	70.72 ± 1.97	60.54 ± 1.58	66.83 ± 1.81	66.53 ± 2.63
USEA	47.08 ± 0.73	29.44 ± 0.97	45.36 ± 1.26	64.26 ± 2.22	68.98 ± 1.13	70.77 ± 1.23	60.31 ± 1.03	65.98 ± 1.04	65.94 ± 0.98
BLPM	46.82 ± 1.82	29.10 ± 2.43	45.07 ± 2.26	64.99 ± 1.34	68.72 ± 1.51	70.77 ± 1.86	60.91 ± 1.95	66.88 ± 1.85	66.58 ± 2.67
EEG-CLIP	44.36 ± 1.73	25.81 ± 2.31	40.30 ± 3.58	64.58 ± 0.69	69.11 ± 0.57	71.02 ± 0.68	60.50 ± 1.37	65.97 ± 2.53	65.52 ± 2.75
Ours	**49.08** ± 1.17	**32.11** ± 1.56	**47.34** ± 1.52	**65.73** ± 0.76	**69.24** ± 0.69	**71.11** ± 1.18	**61.52** ± 1.99	**67.55** ± 1.24	**67.46** ± 2.19
ΔIMP ↑	**+4.25%**	**+9.07%**	**+4.37%**	**+0.94%**	**+0.19%**	**+0.13%**	**+1.00%**	**+0.28%**	**+0.28%**
Methods	FACED, 9-class			SEED-V, 5-class			Mumtaz2016, 2-class		
	Bal.Acc.	Kappa	F1-w	Bal.Acc.	Kappa	F1-w	Bal.Acc.	AUC-PR	AUROC
EEGPT	34.48 ± 0.65	26.04 ± 0.76	34.10 ± 0.88	23.86 ± 0.52	4.91 ± 0.62	23.57 ± 0.76	**89.81** ± 1.64	**97.86** ± 0.71	**97.84** ± 0.73
BrainOmni	22.94 ± 0.82	13.24 ± 0.91	22.50 ± 0.77	24.06 ± 0.51	5.32 ± 0.57	24.08 ± 0.69	80.95 ± 0.69	88.00 ± 1.58	89.58 ± 1.04
LaBraM	55.75 ± 1.06	50.10 ± 1.21	56.53 ± 0.97	38.99 ± 0.43	23.86 ± 0.62	39.42 ± 0.66	89.53 ± 0.62	97.16 ± 0.94	96.78 ± 1.31
CBraMod	57.09 ± 0.68	51.48 ± 0.75	57.47 ± 0.67	38.68 ± 0.97	23.85 ± 1.26	39.48 ± 0.99	88.47 ± 0.89	97.35 ± 0.65	97.16 ± 0.78
USEA	53.22 ± 0.47	47.11 ± 0.54	53.48 ± 0.47	38.49 ± 1.40	22.83 ± 2.26	38.23 ± 2.55	88.38 ± 3.81	97.20 ± 1.10	97.09 ± 0.89
BLPM	56.65 ± 0.99	50.98 ± 0.98	56.98 ± 0.65	38.66 ± 0.97	23.84 ± 1.29	39.39 ± 1.02	88.51 ± 0.91	97.38 ± 0.63	97.18 ± 0.76
EEG-CLIP	53.40 ± 0.59	47.46 ± 0.68	54.10 ± 0.59	38.71 ± 0.74	23.62 ± 1.09	39.30 ± 0.79	87.87 ± 0.50	97.28 ± 0.67	97.25 ± 0.80
Ours	**57.96** ± 0.48	**52.34** ± 0.57	**58.02** ± 0.48	**39.35** ± 0.53	**24.86** ± 0.71	**40.31** ± 0.64	88.97 ± 0.39	97.56 ± 0.45	97.44 ± 0.51
ΔIMP ↑	**+1.52%**	**+1.67%**	**+0.96%**	**+0.92%**	**+4.19%**	**+2.10%**	**-0.94%**	**-0.31%**	**-0.41%**
Methods	MentalArithmetic, 2-class			BCIC2020-3, 5-class			TUEV, 6-class		
	Bal.Acc.	AUC-PR	AUROC	Bal.Acc.	Kappa	F1-w	Bal.Acc.	Kappa	F1-w
EEGPT	53.33 ± 2.03	37.97 ± 2.82	63.86 ± 1.68	22.13 ± 1.33	2.67 ± 1.66	21.42 ± 1.43	62.40 ± 2.88	63.47 ± 4.43	80.67 ± 2.22
BrainOmni	61.67 ± 4.45	50.61 ± 7.48	70.76 ± 5.78	27.60 ± 1.51	9.50 ± 1.89	27.36 ± 1.56	30.24 ± 2.28	23.22 ± 4.52	57.59 ± 4.19
LaBraM	63.06 ± 2.87	51.12 ± 4.70	73.84 ± 4.40	26.16 ± 1.04	7.70 ± 1.30	25.66 ± 1.38	59.66 ± 2.75	58.65 ± 4.29	78.20 ± 2.17
CBraMod	59.79 ± 6.01	43.70 ± 4.03	70.72 ± 2.91	32.51 ± 1.80	15.63 ± 2.25	32.48 ± 1.84	61.11 ± 3.14	59.50 ± 4.43	78.69 ± 2.25
USEA	58.33 ± 4.98	39.82 ± 3.34	66.46 ± 1.59	30.37 ± 2.00	12.97 ± 2.50	30.07 ± 2.35	51.38 ± 4.18	51.34 ± 6.87	74.90 ± 3.48
BLPM	59.93 ± 4.69	43.88 ± 4.25	70.85 ± 3.01	30.11 ± 4.37	12.63 ± 5.46	26.53 ± 6.31	61.05 ± 2.82	59.54 ± 4.55	78.71 ± 2.27
EEG-CLIP	60.76 ± 1.22	44.48 ± 3.47	72.01 ± 1.22	29.60 ± 1.77	12.00 ± 2.21	29.55 ± 1.81	61.15 ± 2.08	55.77 ± 2.42	77.08 ± 1.32
Ours	**65.56** ± 4.44	**51.52** ± 4.90	**74.51** ± 1.60	**33.25** ± 2.18	**16.57** ± 2.73	**33.22** ± 2.21	**63.44** ± 1.98	**63.63** ± 3.51	**80.87** ± 1.77
ΔIMP ↑	**+3.96%**	**+0.78%**	**+0.91%**	**+2.28%**	**+6.01%**	**+2.28%**	**+1.67%**	**+0.25%**	**+0.25%**

Table 2: Performance comparison across diverse EEG foundation model backbones.

Methods	LaBraM								
	BCIC-2B, 2-class			SEED-V, 5-class			MentalArithmetic, 2-class		
	Bal.Acc.	AUC-PR	AUROC	Bal.Acc.	Kappa	F1-w	Bal.Acc.	AUC-PR	AUROC
LaBraM	63.27 ± 2.62	68.18 ± 2.05	70.21 ± 2.14	38.99 ± 0.43	23.86 ± 0.62	39.42 ± 0.66	63.06 ± 2.87	51.12 ± 4.70	73.84 ± 4.40
USEA	65.77 ± 1.36	68.97 ± 1.24	71.52 ± 1.15	39.82 ± 0.81	24.66 ± 1.12	39.73 ± 1.13	56.11 ± 7.27	52.15 ± 5.40	71.29 ± 1.54
BLPM	62.01 ± 3.98	68.34 ± 0.95	71.65 ± 0.35	41.13 ± 0.38	26.70 ± 0.50	41.69 ± 0.46	62.99 ± 2.93	51.55 ± 4.58	73.89 ± 4.40
EEG-CLIP	63.19 ± 1.80	66.16 ± 0.50	69.91 ± 0.31	39.80 ± 0.41	24.79 ± 0.37	40.24 ± 0.37	60.83 ± 2.27	51.68 ± 7.35	69.71 ± 3.93
Ours	**66.58** ± 0.48	**69.02** ± 0.31	**72.27** ± 0.31	**41.25** ± 0.40	**26.87** ± 0.52	**41.87** ± 0.42	**65.00** ± 3.01	**52.31** ± 3.97	**74.01** ± 2.92
ΔIMP ↑	**+1.23%**	**+0.07%**	**+0.87%**	**+0.29%**	**+0.64%**	**+0.43%**	**+3.08%**	**+0.31%**	**+0.16%**
Methods	EEGPT								
	BCIC-2B, 2-class			SEED-V, 5-class			MentalArithmetic, 2-class		
	Bal.Acc.	AUC-PR	AUROC	Bal.Acc.	Kappa	F1-w	Bal.Acc.	AUC-PR	AUROC
EEGPT	62.19 ± 5.85	67.15 ± 8.82	68.76 ± 7.83	23.86 ± 0.52	4.91 ± 0.62	23.57 ± 0.76	53.33 ± 2.03	37.97 ± 2.82	63.86 ± 1.68
USEA	62.08 ± 5.90	68.02 ± 4.91	69.43 ± 4.84	21.67 ± 0.75	2.45 ± 1.15	16.45 ± 2.52	51.39 ± 4.09	35.90 ± 5.78	63.25 ± 4.83
BLPM	56.03 ± 6.16	56.93 ± 7.94	56.92 ± 9.43	23.78 ± 0.43	4.82 ± 0.49	23.35 ± 0.94	53.40 ± 1.97	37.98 ± 2.81	63.87 ± 1.66
EEG-CLIP	61.01 ± 1.53	63.55 ± 1.62	67.29 ± 1.55	23.87 ± 0.48	4.86 ± 0.49	23.77 ± 0.79	53.26 ± 2.90	36.85 ± 4.08	64.38 ± 4.18
Ours	**64.20** ± 2.81	**68.33** ± 2.99	**69.71** ± 3.61	**24.09** ± 0.53	**5.23** ± 0.65	**24.57** ± 0.68	**56.39** ± 2.61	**39.10** ± 1.03	**64.98** ± 2.02
ΔIMP ↑	**+3.23%**	**+0.46%**	**+0.40%**	**+0.92%**	**+6.52%**	**+3.37%**	**+5.60%**	**+2.95%**	**+0.93%**

dent on a specific EEG foundation model, supporting its generalizability across diverse backbone architectures.

4.4 Ablation Study

Setup. We conduct a systematic ablation study with eight variants. **(1) Semantic Grounding (Only-SG)** retains only semantic grounding, without Semantic Interaction, MVSR, or EISP. **(2) Semantic Interaction (Only-SI)** retains only semantic interaction, without Semantic Grounding, MVSR, or EISP. **(3) Full EGSA** combines both mechanisms, without MVSR or EISP. **(4) Unguided Refinement** builds upon Full EGSA and performs semantic refinement without target-model feedback or EISP. **(5) Utility-Guided Refinement (Utility-Guided)** builds upon Full EGSA and uses only descriptor-level utility feedback, without EISP. **(6) Performance-Guided Refinement**

Table 3: Ablation study of the ground–validate–refine–internalize design choices in SemLoop.

	BCIC-2A, 4-class			FACED, 9-class			MentalArithmetic, 2-class		
Methods	**Bal.Acc.**	**Kappa**	**F1-w**	**Bal.Acc.**	**Kappa**	**F1-w**	**Bal.Acc.**	**AUC-PR**	**AUROC**
Only-SG	46.63 ± 1.72	28.84 ± 2.30	44.36 ± 1.92	56.99 ± 0.69	51.39 ± 0.84	57.45 ± 0.77	60.69 ± 1.89	47.74 ± 4.92	71.90 ± 2.49
Only-SI	44.41 ± 1.78	25.88 ± 2.38	41.39 ± 2.44	56.46 ± 0.55	50.67 ± 0.62	56.61 ± 0.70	57.71 ± 1.14	43.72 ± 3.09	70.91 ± 3.49
Full EGSA	47.15 ± 0.48	29.54 ± 0.64	44.96 ± 0.92	57.09 ± 0.87	51.51 ± 0.87	57.56 ± 0.57	62.29 ± 3.84	47.77 ± 5.84	71.16 ± 2.59
Unguided Refinement	47.48 ± 0.80	29.98 ± 1.07	45.12 ± 1.11	57.13 ± 0.63	51.52 ± 0.73	57.52 ± 0.69	62.15 ± 3.72	47.56 ± 6.20	70.11 ± 3.08
Utility-Guided	48.07 ± 1.20	30.76 ± 1.60	46.27 ± 1.57	57.59 ± 0.75	51.97 ± 0.85	57.75 ± 0.70	63.06 ± 5.10	47.91 ± 7.90	72.17 ± 3.51
Performance-Guided	48.00 ± 1.49	30.67 ± 1.99	46.12 ± 1.88	57.24 ± 0.79	51.61 ± 0.92	57.53 ± 0.79	61.74 ± 4.33	47.34 ± 7.39	70.99 ± 3.96
Full MVSR	48.66 ± 1.10	31.55 ± 1.47	46.85 ± 1.54	57.79 ± 0.72	52.14 ± 0.82	57.84 ± 0.69	63.47 ± 4.60	48.44 ± 7.70	73.25 ± 2.93
SemLoop	**49.08** ± 1.17	**32.11** ± 1.56	**47.34** ± 1.52	**57.96** ± 0.48	**52.34** ± 0.57	**58.02** ± 0.48	**65.56** ± 4.44	**51.52** ± 4.90	**74.51** ± 1.60

(Performance-Guided) builds upon Full EGSA and uses only global validation performance feedback, without EISP. **(7) Full MVSR** builds upon Full EGSA and uses both feedback signals, without EISP. Finally, **(8) SemLoop** builds upon Full MVSR and incorporates EISP. **Results and Analysis.** As shown in Table 3, we analyze the contribution of each component in SemLoop.**(i)** Full EGSA consistently outperforms Only-SG and Only-SI, demonstrating that semantic grounding and semantic interaction provide complementary benefits for semantic-guided adaptation. **(ii)** Using model-validated feedback improves refinement performance over Unguided Refinement. Moreover, Full MVSR achieves superior overall performance by jointly leveraging descriptor-level utility feedback and global validation performance feedback. These two signals provide complementary guidance, where utility feedback captures fine-grained semantic contributions and validation performance feedback reflects the overall adaptation outcome. **(iii)** SemLoop further improves upon Full MVSR, demonstrating that EISP successfully internalizes effective refinement experiences into reusable planning knowledge and enables more efficient semantic refinement. Overall, these results validate the ground–validate–refine–internalize design of SemLoop.

4.5 Analysis of Model-Validated Semantic Refinement

(1) Feedback-guided Semantic Refinement: Setup. We further analyze the refinement variants introduced in Sec. 4.4: **Unguided Refinement**, **Utility-Guided Refinement**, **Performance-Guided Refinement**, and **Full MVSR**. All variants build upon Full EGSA and exclude EISP, thereby isolating the effect of model-validated semantic refinement. We use **Initial Semantics (Full EGSA)** as the common starting point. All variants use the same refinement budget. We report validation performance across refinement rounds, consistent with our protocol in which the validation set provides model feedback and is used for refinement-round selection. **Results and Analysis.** As shown in Fig. 3(A-C), semantic refinement exhibits non-monotonic validation trajectories as semantic configurations are explored. Compared with Unguided Refinement, model-guided variants more effectively navigate the semantic search space, with Full MVSR identifying the best-performing configuration under the same refinement budget. Moreover, the advantage of Full MVSR over the single-feedback variants indicates that descriptor-level utility and global performance provide complementary fine-grained and configuration-level guidance, respectively. Together, these results demonstrate how target-model feedback turns predefined, one-way semantic guidance into a closed-loop process in which semantics are iteratively validated and refined according to their model utility.

(2) Need for Semantic Reasoning in Refinement. Setup. Although descriptor-level utility provides a signal for measuring model utility, it remains unclear whether utility-based selection is sufficient for semantic refinement. We therefore compare **Full MVSR** with two utility-based strategies: **Utility Filtering**, which removes descriptors with negative utility, and **Utility Ranking**, which retains the top-k descriptors ranked by their estimated utility scores. Unlike these selection-based strategies, Full MVSR employs evaluator and planner agents to diagnose semantic limitations and generate alternative semantic configurations based on model feedback. All variants start from the same initial semantics and use the same refinement budget. **Results and Analysis.** As shown in Table 4, utility-based strategies provide limited improvements over the initial semantics, demonstrating that model feedback can identify useful descriptors. However, both strategies remain inferior to Full MVSR. These results indicate that utility signals alone are insufficient for effective semantic refinement: they can evaluate existing descriptors but cannot determine how semantic knowledge should be modified. By leveraging evaluator and planner agents to reason over model feedback, MVSR enables broader exploration of the semantic space and discovers semantic configurations with higher model utility.

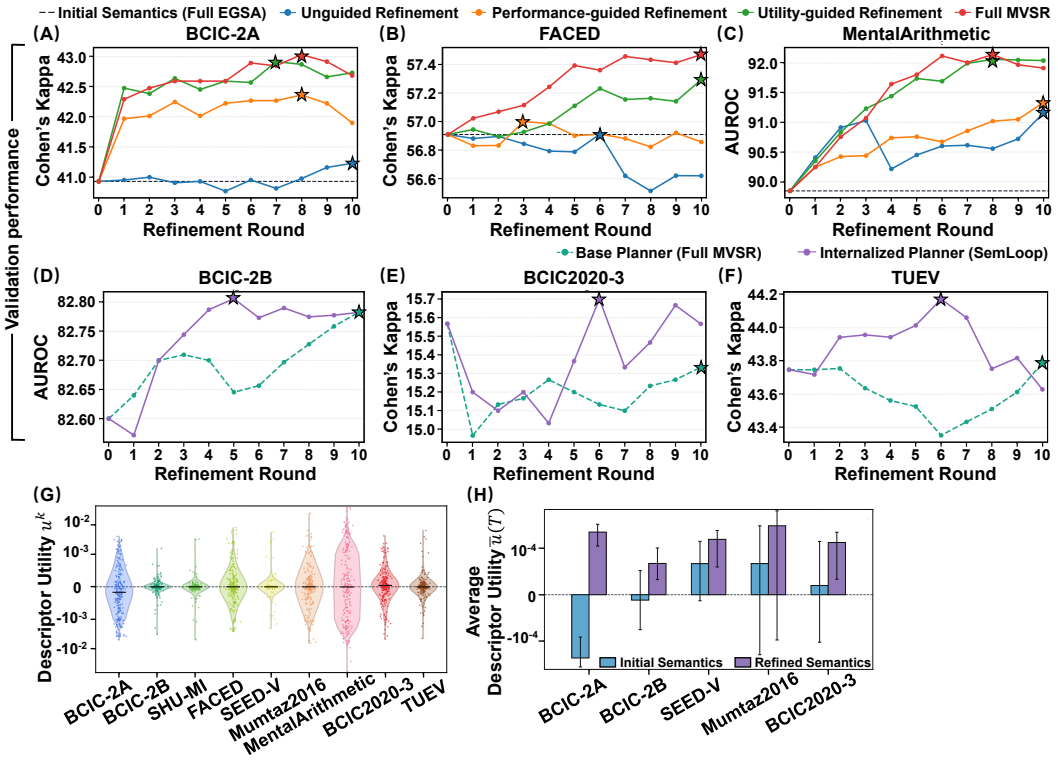

Figure 3: Analysis of model-validated refinement, experience internalization, and model utility in SemLoop. (A–C) Validation performance across refinement rounds shows that Full MVSR outperforms unguided and single-feedback refinement under the same refinement budget. (D–F) Validation performance across refinement rounds demonstrates that the Internalized Planner achieves stronger early-round performance than the Base Planner by leveraging successful refinement experiences. (G) Task-relevant descriptors exhibit substantial variation in model utility, showing that semantic relevance does not necessarily imply usefulness. (H) Model-validated refinement produces semantic banks with higher average utility than the initial semantics.

Table 4: Ablation study of utility-based semantic refinement strategies.

	SHU-MI, 2-class			FACED, 9-class			BCIC2020-3, 5-class		
Methods	**Bal.Acc.**	**AUC-PR**	**AUROC**	**Bal.Acc.**	**Kappa**	**F1-w**	**Bal.Acc.**	**Kappa**	**F1-w**
Full EGSA	60.76 ± 1.50	66.87 ± 1.30	66.59 ± 2.01	57.09 ± 0.87	51.51 ± 0.87	57.56 ± 0.57	32.69 ± 2.44	15.87 ± 3.04	32.70 ± 2.43
Utility Filtering	60.56 ± 1.74	66.66 ± 1.50	66.04 ± 2.72	57.40 ± 0.79	51.76 ± 0.81	57.63 ± 0.47	32.45 ± 2.45	15.57 ± 3.06	32.45 ± 2.47
Utility Ranking	60.61 ± 1.59	66.61 ± 1.54	66.01 ± 2.74	57.34 ± 0.75	51.66 ± 0.83	57.44 ± 0.64	32.43 ± 2.47	15.53 ± 3.09	32.41 ± 2.50
Full MVSR	**61.23** ± 2.05	**67.18** ± 1.47	**67.24** ± 2.34	**57.79** ± 0.72	**52.14** ± 0.82	**57.84** ± 0.69	**32.99** ± 2.85	**16.23** ± 3.56	**32.99** ± 2.85

4.6 Analysis of Experience-Internalized Semantic Planning

Setup. We further analyze the effect of EISP by comparing the settings corresponding to **Full MVSR** and **SemLoop** in Sec. 4.4: **Base Planner (Full MVSR)** and **Internalized Planner (SemLoop)**. Both settings share the same Full MVSR refinement protocol, including the initial semantics, target-model feedback, and refinement budget, differing only in whether successful refinement experiences are internalized through EISP. We report validation performance across refinement rounds to examine how experience internalization affects subsequent semantic planning. **Results and Analysis.** As shown in Fig. 3(D-F), the Internalized Planner achieves stronger performance in early refinement rounds and identifies a better-performing semantic configuration under the same refinement budget. These results indicate that EISP converts successful refinement experiences into reusable planning knowledge, reducing repeated exploration and enabling the planner to identify better-performing semantic configurations with fewer refinement rounds.

Table 5: Cross-task transfer of internalized experience. Cross-Task Planner is compared with Initial Semantics after one refinement round and with the Base Planner under the same refinement budget.

	BCIC-2A, 4-class			MentalArithmetic, 2-class			TUEV, 6-class		
Methods	**Bal.Acc.**	**Kappa**	**F1-w**	**Bal.Acc.**	**AUC-PR**	**AUROC**	**Bal.Acc.**	**Kappa**	**F1-w**
Initial Semantics	47.15 ± 0.48	29.54 ± 0.64	44.96 ± 0.92	62.29 ± 3.84	47.77 ± 5.84	71.16 ± 2.59	61.63 ± 2.27	59.54 ± 2.79	78.80 ± 1.28
Cross-Task (1 Round)	48.59 ± 1.14	31.46 ± 1.52	46.73 ± 1.64	64.72 ± 5.34	47.99 ± 5.90	72.39 ± 2.80	62.10 ± 1.83	60.42 ± 2.41	79.27 ± 1.10
Base Planner	48.66 ± 1.10	31.55 ± 1.47	46.85 ± 1.54	63.47 ± 4.60	48.44 ± 7.70	73.25 ± 2.93	61.98 ± 1.96	60.25 ± 2.68	79.18 ± 1.24
Cross-Task Planner	**48.70** ± 0.98	**31.60** ± 1.31	**46.93** ± 1.15	**65.69** ± 7.45	**52.51** ± 9.72	**75.19** ± 5.53	**62.39** ± 1.89	**61.22** ± 2.09	**79.66** ± 1.00

4.7 Cross-Task Transfer of Internalized Experience

Setup. We further examine whether experience internalized by EISP transfers to unseen tasks. For each target dataset, we adopt a leave-one-task-out protocol, internalizing only successful refinement experiences from the remaining tasks and evaluating the resulting planner on the held-out target task. We refer to this planner as the **Cross-Task Planner** and evaluate it under two settings. First, we perform a single refinement round and compare against Initial Semantics (Full EGSA) to assess whether transferred experience can directly guide effective semantic refinement. Second, we perform multiple refinement rounds and compare against the Base Planner (Full MVSR) under the same refinement budget to assess the benefit of transferred experience during iterative refinement. **Results and Analysis.** As shown in Table 5, the Cross-Task Planner outperforms Initial Semantics after a single refinement round, showing that experience internalized from other tasks can directly benefit semantic planning on an unseen task. Under the same refinement budget, it further outperforms the Base Planner, demonstrating the benefit of transferred experience for iterative semantic refinement. Together, these results show that EISP captures reusable refinement knowledge beyond task-specific trajectories, supporting the cross-task transferability of internalized experience.

4.8 Analysis of Semantic Relevance and Model Utility

(1) Heterogeneous Utility of Task-Relevant Semantics. We examine whether task-relevant semantics are equally useful to the target model by computing descriptor-level utility before model-validated refinement. As shown in Fig. 3(G), task-relevant descriptors exhibit substantial utility variation, including positive, near-zero, and negative values, demonstrating that semantic relevance alone is insufficient to determine useful knowledge for model adaptation. **(2) Utility Improvement after Model-Validated Refinement.** We next examine whether MVSR steers semantic knowledge toward higher model utility by comparing the average descriptor utility before and after refinement: $\bar{u}(T) = \frac{1}{|T|}\sum_{t^k \in T} u^k$. As shown in Fig. 3(H) and Fig. 6, refined semantic banks consistently exhibit higher average utility across tasks, demonstrating that MVSR improves both adaptation performance and semantic utility. **(3) Qualitative Analysis of Semantic Refinement.** To understand how MVSR refines semantic guidance, we provide initial and refined descriptor sets in Appendix H. Compared with the initial semantic banks, which mainly contain task-relevant spectral and statistical descriptors, the refined banks additionally incorporate region-specific activity, functional connectivity, and interhemispheric interaction patterns. This shift suggests that model feedback complements task-relevant semantic knowledge with descriptors better aligned with target-model utility.

5 Conclusion

We present SemLoop, a ground–validate–refine–internalize paradigm for closed-loop semantic adaptation of EEG foundation models that bridges semantic relevance and model utility. SemLoop grounds semantics in EEG representations, iteratively validates and refines them through target-model feedback, and internalizes successful refinements for subsequent semantic planning. Across nine EEG benchmarks and multiple EEG foundation models, SemLoop consistently outperforms standard full fine-tuning and predefined semantic-guided approaches. Further analyses reveal substantial utility variation among task-relevant semantics and show that model-validated refinement yields semantics with higher model utility. Overall, SemLoop shifts semantic-guided adaptation from predefined, one-way knowledge injection toward an adaptive process shaped by target-model feedback. **Limitations and Future Work.** While this work focuses on semantic adaptation of EEG foundation models, extending the closed-loop paradigm to other neural modalities and broader adaptation settings remains an important direction.

AI USE STATEMENT

We did not use generative AI tools to generate scientific claims, conduct experiments, or analyze experimental results. The authors are responsible for the final content of this work.

ETHICS STATEMENT

This work uses publicly available EEG datasets and does not involve new human subject experiments or data collection.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed framework, implementation details, and dataset preprocessing procedures. The anonymous source code repository is available at: `https://anonymous.4open.science/r/SemLoop-75F0`.

REFERENCES

Reza Abiri, Soheil Borhani, Eric W. Sellers, Yang Jiang, and Xiaopeng Zhao. A comprehensive review of eeg-based brain–computer interface paradigms. *Journal of Neural Engineering*, 16, 2019. URL https://api.semanticscholar.org/CorpusID:54559538.

Swati Aggarwal and Nupur Chugh. Review of machine learning techniques for eeg based brain computer interface. *Archives of Computational Methods in Engineering*, 29:3001 – 3020, 2022. URL https://api.semanticscholar.org/CorpusID:255397903.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16(1-6):1, 2008.

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.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10(1):740, 2023.

Myeong-Ju Cho, Hye-Bin Shin, Seo-Hyun Lee, and Seong-Whan Lee. Continuous-latent predictive modeling with semantic alignment for eeg-language foundation models. *arXiv preprint arXiv:2608.11656*, 2026.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. Domain-specific denoising diffusion probabilistic models for brain dynamics. *arXiv preprint arXiv:2305.04200*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Ji-Hoon Jeong, Jeong-Hyun Cho, Young-Eun Lee, Seo-Hyun Lee, Gi-Hwan Shin, Young-Seok Kweon, José del R Millán, Klaus-Robert Müller, and Seong-Whan Lee. 2020 international brain–computer interface competition: A review. *Frontiers in human neuroscience*, 16:898300, 2022.

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*, volume 2024, pp. 16405–16426, 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*, volume 2025, pp. 55436–55457, 2025.

Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, and Yogatheesan Varatharajah. Eeg foundation models: a critical review of current progress and future directions. *Journal of neural engineering*, 23(2):021001, 2026.

Junhong Lai, Jiyu Wei, Lin Yao, and Yueming Wang. A simple review of eeg foundation models: Datasets, advancements and future perspectives. *arXiv preprint arXiv:2504.20069*, 2025.

Weibin Li, Wendu Li, Xihua Yin, Yushan You, Xinke Shen, Zongxiang Tan, and Quanying Liu. Cbp: Learning shared cognitive basis space and connectivity patterns for cross-cognitive-task brain dynamics modeling. *Neurocomputing*, 700:134353, 2026a. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.134353. URL https://www.sciencedirect.com/science/article/pii/S0925231226017510.

Weibin Li, Wendu Li, Yushan You, Chen Wei, and Quanying Liu. Neuroonline: Bridging pretraining and online adaptation for eeg foundation models. *arXiv preprint arXiv:2607.03925*, 2026b.

Bo Liu, Changdong Liu, Yanshan Xiao, Laiwang Liu, Weibin Li, and Xiaodong Chen. Adaboost-based transfer learning method for positive and unlabelled learning problem. *Knowledge-Based Systems*, 241:108162, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.108162. URL https://www.sciencedirect.com/science/article/pii/S0950705122000260.

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, 2021.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018.

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.

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

John W. Pratt. Remarks on zeros and ties in the wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54:655–667, 1959. URL https://api.semanticscholar.org/CorpusID:120225998.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

David Steyrl, Reinhold Scherer, Josef Faller, and Gernot R Müller-Putz. Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: a practical and convenient non-linear classifier. *Biomedical Engineering/Biomedizinische Technik*, 61(1):77–86, 2016.

Kaido Värbu, Naveed Muhammad, and Yar Muhammad. Past, present, and future of eeg-based bci applications. *Sensors*, 22(9):3331, 2022.

Mumtaz Wajid. Mdd patients and healthy controls eeg data (new). 2016. URL https://api.semanticscholar.org/CorpusID:151858102.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=lvS2b8CjG5.

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 *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=NPNUHgHF2w.

Suli Wang, Yangshen Deng, Zhenghua Bao, Xinyu Zhan, and Yiqun Duan. Neurottt: Bridging pretraining-downstream task misalignment in eeg foundation models via test-time training. *arXiv preprint arXiv:2509.26301*, 2025b.

Jiamin Wu, Zichen Ren, Junyu Wang, Pengyu Zhu, Yonghao Song, Mianxin Liu, Qihao Zheng, Lei Bai, Wanli Ouyang, and Chunfeng Song. Adabrain-bench: Benchmarking brain foundation models for brain-computer interface applications. *arXiv preprint arXiv:2507.09882*, 2025.

Qinfan Xiao, Ziyun Cui, Chi Zhang, SiQi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified EEG and MEG signals. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=cjHQj0tCy6.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. Eeg-fm-bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of eeg foundation models. *arXiv preprint arXiv:2508.17742*, 2025.

Ziman Ye, Muyun Jiang, Jiaqi Zhu, Hao Zheng, Mengchi Rong, and Fang Deng. A unified foundation model for heterogeneous eeg signal modeling via language-aligned semi-supervised learning. *Neural Networks*, pp. 109544, 2026.

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.

A Additional Details on Datasets and Preprocessing

We provide additional details on the datasets and preprocessing procedures used in downstream tasks (Table 6).

A.1 Motor Imagery Classification

BCIC-2A includes EEG data collected from 9 subjects through 22 channels. During data collection, each subject performed four motor imagery tasks: imagining movements of the left hand, right hand, both feet, and tongue. We divide the continuous recordings into 4-second segments and obtain 5,088 trials in total. The signals are filtered within 0.3–50 Hz and resampled to 200 Hz. Following CBraMod, we use subjects 1–5 for training, subjects 6–7 for validation, and subjects 8–9 for testing under a cross-subject setting.

BCIC-2B is a binary motor imagery EEG dataset collected from 9 subjects using 3 EEG channels. Participants imagined moving either their left or right hand. The recordings are divided into 4-second trials, resulting in 6,520 samples. We filter the signals to 0–38 Hz and resample them to 200 Hz. The EOG channels are then removed, and common average referencing (CAR) is applied across the EEG channels. For cross-subject evaluation, we use subjects 1–5 for training, subjects 6–7 for validation, and subjects 8–9 for testing.

SHU-MI is a binary motor imagery EEG dataset comprising 25 subjects and 32 EEG channels. It also focuses on imagined movements of the left hand and the right hand. The recordings are divided into 4-second trials and resampled to 200 Hz, yielding 11,988 samples. Subjects 1–15 constitute the training set, while subjects 16–20 and 21–25 form the validation and test sets.

A.2 Emotion Recogntion

FACED is a fine-grained emotion recognition dataset with 32-channel EEG recordings from 123 subjects. It covers nine emotional states: amusement, inspiration, joy, tenderness, anger, fear, disgust, sadness, and neutral. We divide the recordings into 10-second trials and resample them to 200

Table 6: Overview of downstream BCI tasks and datasets.

BCI Tasks	Datasets	Rate	# Channels	Duration	# Samples	Label
I. Motor Imagery Classification	BCIC-2A	250Hz	22	4s	5,088	4-class
	BCIC-2B	250Hz	3	4s	6,520	2-class
	SHU-MI	250Hz	32	4s	11,988	2-class
II. Emotion Recognition	FACED	250Hz	32	10s	10,332	9-class
	SEED-V	1000Hz	62	1s	115,001	5-class
III. Mental Disorder Diagnosis	Mumtaz2016	256Hz	19	5s	7,143	2-class
IV. Mental Stress Detection	MentalArithmetic	500Hz	20	5s	1,707	2-class
V. Imagined Speech Classification	BCIC2020-3	256Hz	64	3s	6,000	5-class
VI. Event Type Classification	TUEV	250Hz	16	5s	113,353	6-class

Hz, obtaining 10,332 samples. For cross-subject evaluation, subjects 1–80 are used for training, subjects 81–100 for validation, and subjects 101–123 for testing.

SEED-V is an emotion recognition dataset collected from 16 subjects using 62 EEG channels. It contains five classes: happiness, sadness, neutral, disgust, and fear. Each subject completed three sessions, with 15 trials recorded in each session. One corrupted session is excluded from our experiments. We remove the M1, M2, VEO, and HEO channels and extract individual trials using the provided onset and offset annotations. To suppress low-frequency drift and high-frequency noise, we filter the signals within 0.3–75 Hz and set the sampling rate to 200 Hz. We then extract non-overlapping 1-second segments from each trial, resulting in 115,001 samples. Within each session, five trials are used for training, five for validation, and five for testing.

A.3 Mental Disorder Diagnosis

Mumtaz2016 is a clinical EEG benchmark for binary depression recognition. It includes 34 patients diagnosed with major depressive disorder (MDD) and 30 healthy controls. EEG signals were collected through 19 electrodes at 256 Hz under three conditions: eyes open, eyes closed, and task states. To maintain the same experimental setting as CBraMod, we retain only the eyes-open and eyes-closed recordings. The signals are processed with a 0.3–75 Hz band-pass filter and a 50 Hz notch filter before being resampled to 200 Hz. We then divide the recordings into 5-second trials, obtaining 7,143 samples. At the subject level, the training set includes 21 MDD and 20 NC subjects, the validation set includes 5 MDD and 4 NC subjects, and the test set includes 8 MDD and 6 NC subjects.

A.4 Mental Stress Detection

MentalArithmetic is an EEG dataset for binary mental stress recognition. It contains recordings from 36 subjects before and during a mental arithmetic task. The pre-task recordings are labeled as non-stress, while the recordings collected during the task are labeled as stress. EEG signals were recorded from 20 electrodes at 500 Hz. After filtering the signals within 0.5–45 Hz and resampling them to 200 Hz, we divide them into 5-second trials. This results in 1,707 samples. Subjects 1–28 form the training set, subjects 29–32 form the validation set, and subjects 33–36 form the test set.

A.5 Imagined Speech Classification

BCIC2020-3 is a five-class imagined speech EEG dataset. It contains 64-channel recordings from 15 subjects. The five imagined commands are "Hello," "Help me," "Stop," "Thank you," and "Yes." Each subject–command pair contains 80 trials, giving 6,000 trials in total. Every trial covers a 3-second EEG segment, which is resampled to 200 Hz. We use the official split of 60 trials for training, 10 for validation, and 10 for testing.

A.6 Event Type Classification

TUEV is a clinical EEG dataset for six-class neurological event recognition. The event categories include spike and sharp wave (SPSW), generalized periodic epileptiform discharges (GPED), periodic lateralized epileptiform discharges (PLED), eye movement (EYEM), artifacts (ARTF), and

background activity (BCKG). EEG signals were recorded from 23 channels at 250 Hz. Following CBraMod, we retain 16 commonly used bipolar channels defined by the international 10–20 system. The selected signals undergo 0.3–75 Hz band-pass filtering and 60 Hz notch filtering, followed by resampling to 200 Hz. We then divide them into 5-second segments and obtain 112,491 samples. We follow the official train-test split and reserve 20% of the training data for validation.

B Additional Details on Baselines

B.1 EEG Foundation Models

We evaluate four representative foundation models through full fine-tuning on each downstream EEG task.

CBraMod introduces a criss-cross Transformer to model channel-wise and temporal interactions in EEG signals. It also uses asymmetric conditional positional encoding to retain the position information of EEG patches.

LaBraM converts channel-wise EEG patches into discrete neural tokens through a vector-quantized tokenizer, which is trained using neural spectrum prediction.

EEGPT replaces raw signal reconstruction with spatial and temporal alignment in the representation space. To perform this alignment efficiently, it adopts a hierarchical encoder for spatial and temporal feature extraction.

BrainOmni is jointly pretrained on EEG and MEG recordings. BrainTokenizer converts continuous neural signals into discrete tokens, while a sensor encoder captures information about different sensors and acquisition devices.

B.2 Semantic-Guided EEG Models

We further compare SemLoop with three representative methods that introduce language-derived supervision into EEG representation learning. To ensure a fair comparison, we adapt these methods to the downstream setting using the same backbone and evaluation protocol as SemLoop.

EEG-CLIP aligns EEG recordings with their paired textual descriptions through contrastive learning. The two modalities are mapped into a shared embedding space, where the text provides fixed semantic supervision for learning EEG representations.

USEA integrates semantic alignment and autoregressive EEG modeling within a semi-supervised framework. For labeled samples, descriptions of class labels are encoded by a frozen language model and used as alignment targets. Unlabeled samples are learned through autoregressive prediction of EEG representations.

BLPM learns continuous EEG representations through latent target prediction. Its multi-query semantic decomposition module extracts task-related components from these representations and aligns them with language-derived semantics in a shared latent space.

Although these methods incorporate language knowledge in different ways, their semantic guidance remains fixed during adaptation. As a result, they cannot refine semantic guidance based on model feedback.

C More Details for Backbone Models

We evaluate our framework with three pretrained EEG backbones to examine its generalizability across different EEG foundation model designs. The evaluation settings remain unchanged when switching between backbones.

CBraMod explicitly models interactions across EEG channels and temporal patches. Its criss-cross Transformer captures dependencies along both dimensions, while asymmetric conditional positional encoding provides location information for individual patches.

LaBraM follows a tokenization-based pretraining paradigm. It maps channel-wise EEG patches into discrete neural tokens using a vector-quantized tokenizer trained through neural spectrum prediction.

EEGPT performs self-supervised learning in the representation space rather than reconstructing raw EEG signals. Its hierarchical encoder aligns spatial and temporal features at different levels to learn informative EEG representations.

Together, these backbones allow us to determine whether the observed gains depend on a particular architecture or pretraining objective.

D Details of Evaluation Metrics

Consistent with the evaluation protocol of CBraMod, we report five metrics that capture complementary aspects of classification performance, including class-wise recall, ranking quality, agreement beyond chance, and the balance between precision and recall.

Balanced accuracy averages the recall obtained for each class. Since every class contributes equally, this metric is less affected by differences in class frequency and can be used for both binary and multi-class tasks.

AUC-PR measures how precision and recall change as the decision threshold varies. It is particularly informative for imbalanced binary tasks because it focuses on the model's ability to identify positive samples.

AUROC evaluates how well a model distinguishes between the positive and negative classes in binary classification across different thresholds. A higher AUROC indicates that positive samples are more consistently assigned higher scores than negative samples.

Cohen's Kappa measures the agreement between predicted and ground-truth labels beyond what would be expected by chance. By accounting for chance agreement, it provides a more reliable evaluation than raw accuracy for imbalanced multi-class tasks.

Weighted F1-score evaluates multi-class classification by first computing the F1-score of each class and then weighting it by the number of samples in that class. It summarizes precision and recall while accounting for the class distribution of the dataset.

E Details of Implementation Details

All experiments are conducted with Python 3.10.20 and PyTorch 2.5.1 under CUDA 12.1. The experiments run on a Linux server equipped with an NVIDIA GeForce RTX L40 GPU (48 GB memory) and an AMD EPYC 9554 64-core CPU.

F Details of Hyperparameter Settings

We adopt the training configuration of CBraMod for all post-training experiments, as detailed in Table 7. The same settings are used across all datasets and backbone models to ensure a consistent comparison. Additional hyperparameters for MVSR and EISP are reported in Table 8.

G Prompt Templates for Semantic Refinement

We provide the complete prompt templates for the evaluator and planner agents. The two prompt templates are presented in Figs. 4 and 5.

H Initial and Refined Semantic Descriptor Sets

Tables 9–12 present the initial and refined semantic descriptor sets for representative downstream datasets. The initial sets contain task-relevant neurophysiological descriptors, while the refined sets result from iterative model-validated refinement. Newly introduced descriptors are underlined.

Table 7: Hyperparameter settings for downstream optimization.

Hyperparameter	Setting
Epochs	50
Batch size	64
Dropout rate	0.1
Optimizer	AdamW
Learning rate	1e-4
Weight decay	5e-2
Scheduler	CosineAnnealingLR
Cosine cycle epochs	50
Minimum learning rate	1e-6
Gradient clipping norm	1.0
Label smoothing	0.1

Table 8: Hyperparameter settings for MVSR and EISP.

Hyperparameter	Setting
Initial semantic descriptor set size (K)	50
Candidate semantic revisions per round (P)	3
Number of refinement rounds	10
Planner fine-tuning epochs	5

I Additional Results of Semantic Utility Improvement

Fig. 6 supplements Fig. 3(H) with results on the remaining four datasets. Consistent improvements in average descriptor utility are observed after model-validated refinement across SHU-MI, FACED, MentalArithmetic, and TUEV. Notably, SHU-MI shows a transition from negative to positive average utility, indicating that MVSR effectively adjusts semantic guidance toward descriptors that are more beneficial to the target model. For the other three datasets, the initially positive average utility is further improved after refinement. These results demonstrate that MVSR consistently shifts semantic guidance toward higher model utility across diverse downstream tasks.

Evaluator Agent Prompt

```
Act as an EEG semantic evaluator for the multiclass task "{self.task}".

Task information:
- Target classes: {self.class_text}
- Current validation metric: {current_metric:.6f}

Semantic basis:
- Indexed descriptors with utility scores:
{indexed descriptors with overall and class-wise utility scores}
- Complete current descriptor basis: {current_text}

Overall utility is the validation cross-entropy loss after removing a descriptor
minus the loss obtained with the complete basis. Positive utility indicates that
the descriptor is useful because its removal increases the loss. Negative utility
suggests that the descriptor may be harmful because its removal reduces the loss.
Near-zero utility indicates that the descriptor may be weak or redundant.
Class-wise utility is calculated in the same way using the average cross-entropy
loss over samples from each target class.

Use both attribution evidence and EEG semantics to determine which descriptors
should be preserved or replaced.

Descriptor diagnosis:
- preserve_indices must contain up to three different indices of clearly useful
descriptors. Prefer descriptors with positive overall utility and consistent
utility across classes.
- replacement_indices must contain exactly five different indices of harmful, weak,
redundant, misleading, or poorly class-balanced descriptors. Negative or near-zero
utility should generally be preferred, but semantic redundancy and conflicting
utility across classes must also be considered.
- Every index must come from the indexed descriptors.
- preserve_indices and replacement_indices must not overlap.

Semantic diagnosis:
- coverage_gaps must contain exactly two missing scientific EEG directions that are
not already represented in the complete basis. Describe general and measurable
semantic directions rather than complete class-specific descriptors.
- avoid_semantics must contain one to three short semantic types that future
proposals should avoid. Do not write attribution rules or instruction sentences in
this field.
- planning_guidance must contain one short, actionable instruction for generating
nonredundant multiclass EEG concepts.

Additional constraints:
- Consider whether the current basis covers all target classes.
- Do not copy existing descriptors into coverage_gaps.
- Do not generate final delete_indices or add_targets; only provide diagnostic
recommendations for the planner.
- Do not use unsupported baseline-relative semantics such as increase, decrease,
suppression, rebound, ERD, or ERS unless baseline-relative evidence is explicitly
available.

Return only a valid JSON object with exactly these keys:
{{"preserve_indices":[0,1],
  "replacement_indices":[2,3,4,5,6],
  "coverage_gaps":["missing EEG direction one","missing EEG direction two"],
  "avoid_semantics":["misleading semantic type"],
  "planning_guidance":"one short actionable sentence"}}
```

Figure 4: Prompt template for the evaluator agent.

Planner Agent Prompt

```
Improve the EEG descriptor basis for the task "{self.task}".

Task information:
- Target classes: {self.class_text}
- Proposal index: {proposal_index + 1}
- Current validation metric: {current_metric:.6f}

Evaluation-agent diagnosis:
- Preserve: {preserve_text}
- Replace: {replacement_text}
- Missing scientific semantics: {coverage_text}
- Avoid: {avoid_text}
- Guidance: {guidance_text}

Indexed descriptors considered for replacement:
{indexed descriptors with overall and class-wise utility scores}
Complete current descriptor basis: {current_text}

Descriptor revision:
- Select up to five different delete_indices from the indexed descriptors
considered for replacement.
- Prioritize descriptors listed under Replace and those with lower or negative
utility.
- Do not select any descriptor listed under Preserve.
- Generate up to five different add_targets so that valid alternatives remain after
duplicate filtering.
- Each add target must contain 2 to 6 words and express one EEG-observable,
task-relevant scientific concept.
- Use the missing scientific semantics as general directions rather than copying
them mechanically.

Additional constraints:
- Do not add any descriptor already present in the complete current basis.
- Do not add case-only, hyphenation-only, singular-plural, abbreviation-only, or
other trivial variants of existing descriptors.
- Do not add a paraphrase of a deleted descriptor.
- Do not repeat an existing band-region, laterality, connectivity, temporal, or
morphology meaning.
- Do not use unsupported baseline-relative claims such as increase, decrease,
suppression, rebound, ERD, or ERS unless they are explicitly requested in the
diagnosis.

Return only a valid JSON object with exactly these keys:
{{"delete_indices":[0,1,2,3,4],
  "add_targets":["candidate EEG concept one","candidate EEG concept two","candidate
EEG concept three","candidate EEG concept four","candidate EEG concept five"]}}
```

Figure 5: Prompt template for the planner agent.

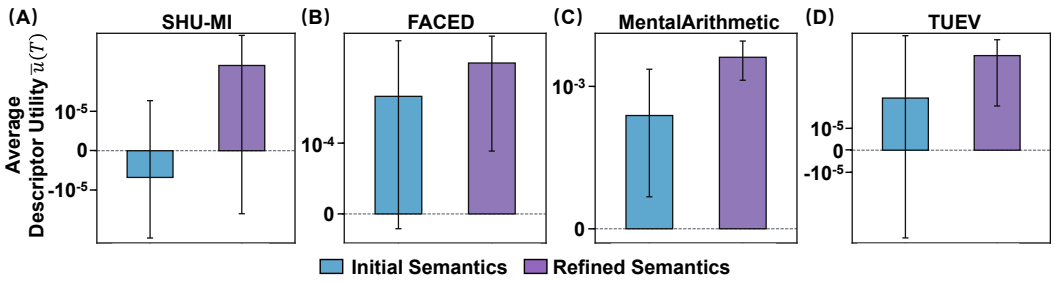

Figure 6: Additional results of model-validated semantic refinement. The refined semantic banks exhibit higher average descriptor utility than their initial counterparts.

Table 9: Initial and refined semantic descriptor sets for BCIC-2A.

Initial Descriptor Set	Refined Descriptor Set
Cz mu power	Cz mu power
C3 beta power	C3 beta power
central temporal variability	central temporal variability
C3 mu band power	C3 mu band power
C4 mu band power	C4 mu band power
Cz beta band power	Cz beta band power
central mu beta ratio	central mu beta ratio
Hjorth mobility	Hjorth mobility
Hjorth complexity	Hjorth complexity
central spectral entropy	central spectral entropy
C4 beta band power	C4 beta band power
left-right sensorimotor mu asymmetry	sensorimotor mu power
bilateral central theta ratio	central mu power
C3 delta power	sensorimotor beta power
C4 theta power	central beta power
Cz alpha power	sensorimotor alpha power
central parietal mu power	central alpha power
left-right parietal beta ratio	sensorimotor gamma power
bilateral occipital gamma power	central theta power
C3 alpha power	central gamma power
C4 beta power	parietal mu power
central temporal variance	parietal theta power
left-right temporal dynamics	parietal alpha power
bilateral alpha mu ratio	parietal beta power
C3 theta power	parietal gamma power
C4 mu power	interhemispheric sensorimotor asymmetry
Cz gamma power	central alpha asymmetry
central parietal theta ratio	central beta asymmetry
left-right parietal gamma asymmetry	central mu asymmetry
bilateral occipital mu ratio	sensorimotor alpha connectivity
C4 alpha power	sensorimotor beta connectivity
Cz parietal power	interhemispheric mu connectivity
central temporal line length	frontocentral theta connectivity
left-right temporal zero-crossing rate	central-parietal connectivity
bilateral alpha gamma ratio	central theta/mu power ratio
C3 mu power	alpha/gamma power ratio
Cz theta power	sensorimotor temporal variance
central parietal mu ratio	EEG waveform line length
left-right parietal beta asymmetry	zero-crossing rate
contralateral sensorimotor mu power	Hjorth activity
bilateral central beta power	waveform regularity
left-right beta asymmetry	waveform sharpness
bilateral sensorimotor synchronization	EEG peak amplitude
motor EEG waveform line length	EEG peak-to-trough amplitude
bilateral theta mu ratio	
central alpha gamma ratio	
left-right gamma asymmetry	
bilateral parietal theta power	
C3 alpha band power	
delta power in left hemisphere	

Table 10: Initial and refined semantic descriptor sets for SEED-V.

Initial Descriptor Set	Refined Descriptor Set
frontal alpha band power	frontal alpha band power
central theta band power	central theta band power
parietal beta band power	parietal beta band power
sensorimotor gamma band power	sensorimotor gamma band power
frontal theta band power	frontal theta band power
frontal mu band power	frontal mu band power
parietal alpha band power	parietal alpha band power
central mu band power	central mu band power
alpha/beta power ratio	alpha/beta power ratio
frontal alpha asymmetry	frontal alpha asymmetry
central theta asymmetry	central theta asymmetry
frontal gamma asymmetry	frontal gamma asymmetry
central alpha asymmetry	central alpha asymmetry
parietal beta asymmetry	parietal beta asymmetry
gamma/theta power ratio	gamma/theta power ratio
alpha band power	alpha band power
theta band power	theta band power
beta band power	beta band power
gamma band power	gamma band power
mu band power	mu band power
alpha temporal variance	alpha temporal variance
gamma temporal variance	gamma temporal variance
zero-crossing rate	zero-crossing rate
waveform line length	waveform line length
Hjorth mobility	Hjorth mobility
waveform regularity	waveform regularity
waveform sharpness	waveform sharpness
EEG peak amplitude	EEG peak amplitude
beta/gamma power ratio	frontal theta asymmetry
theta/alpha power ratio	frontal–central spectral difference
gamma/beta power ratio	frontoparietal spectral difference
parietal mu asymmetry	central–parietal spectral difference
alpha/gamma asymmetry	frontoparietal alpha connectivity
theta/beta asymmetry	interhemispheric theta connectivity
beta/alpha power ratio	frontotemporal gamma connectivity
frontal central asymmetry	frontoparietal theta coherence
frontal parietal asymmetry	frontal mu temporal variance
central parietal asymmetry	parietal theta temporal variance
frontal C3 activity	frontal gamma temporal variance
central C4 activity	global spectral entropy
parietal Cz activity	C3 sensorimotor activity
frontal contralateral activity	C4 sensorimotor activity
central ipsilateral activity	Cz midline activity
parietal contralateral activity	
theta temporal variance	
beta temporal variance	
Hjorth activity	
Hjorth complexity	
EEG trough amplitude	
EEG peak-to-trough amplitude	

Table 11: Initial and refined semantic descriptor sets for Mumtaz2016.

Initial Descriptor Set	Refined Descriptor Set
frontal alpha band power	frontal alpha band power
frontal theta band power	frontal theta band power
parietal beta band power	parietal beta band power
central gamma band power	central gamma band power
posterior alpha band power	posterior alpha band power
theta beta power ratio	theta beta power ratio
frontal alpha asymmetry	frontal alpha asymmetry
left-right frontal imbalance	left-right frontal imbalance
frontoparietal power difference	frontoparietal power difference
global temporal variability	global temporal variability
EEG RMS amplitude	EEG RMS amplitude
EEG waveform line length	EEG waveform line length
frontal mu band power	frontal mu band power
parietal delta band power	parietal delta band power
central alpha band power	central alpha band power
posterior beta band power	posterior beta band power
theta gamma power ratio	theta gamma power ratio
parietal theta band power	parietal theta band power
alpha mu power ratio	alpha mu power ratio
frontal parietal asymmetry	frontal parietal asymmetry
posterior theta band power	posterior theta band power
alpha beta power ratio	alpha beta power ratio
posterior mu band power	posterior mu band power
theta alpha power ratio	theta alpha power ratio
beta gamma power ratio	beta gamma power ratio
parietal beta asymmetry	parietal beta asymmetry
alpha gamma power ratio	alpha gamma power ratio
central theta asymmetry	central theta asymmetry
posterior alpha asymmetry	posterior alpha asymmetry
frontal central asymmetry	frontoparietal alpha connectivity
posterior gamma band power	interhemispheric theta connectivity
beta theta power ratio	posterior alpha connectivity
central gamma asymmetry	frontocentral theta connectivity
mu theta power ratio	frontal waveform temporal complexity
frontal parietal difference	global spectral entropy
central alpha asymmetry	Hjorth mobility
frontal central difference	Hjorth complexity
mu beta power ratio	waveform regularity
theta mu power ratio	frontal–central alpha connectivity
beta mu power ratio	frontotemporal theta connectivity
gamma theta power ratio	
parietal gamma asymmetry	
alpha beta asymmetry	
mu theta asymmetry	
theta gamma asymmetry	
central mu asymmetry	
posterior beta asymmetry	
alpha gamma asymmetry	
beta mu asymmetry	
gamma theta asymmetry	

Table 12: Initial and refined semantic descriptor sets for BCIC2020-3.

Initial Descriptor Set	Refined Descriptor Set
frontal theta band power	frontal theta band power
central beta band power	central beta band power
temporal gamma band power	temporal gamma band power
left-right temporal asymmetry	left-right temporal asymmetry
frontotemporal power difference	frontotemporal power difference
central temporal variability	central temporal variability
speech imagery signal amplitude	speech imagery signal amplitude
EEG zero-crossing rate	EEG zero-crossing rate
EEG waveform line length	EEG waveform line length
frontal mu power	frontal mu power
central delta power	central delta power
temporal beta power	temporal beta power
frontocentral beta activity	frontocentral beta activity
frontal delta power	frontal delta power
central sigma power	central sigma power
frontocentral mu power	frontocentral mu power
temporal alpha power	temporal alpha power
frontocentral gamma activity	frontocentral gamma activity
central gamma power	central gamma power
temporal delta power	temporal delta power
temporal mu power	temporal mu power
frontal gamma power	frontal gamma power
central theta power	central theta power
frontal alpha band power	frontal alpha band power
frontocentral alpha suppression	C3 sensorimotor activity
theta beta power ratio	C4 sensorimotor activity
frontal C3 activity	Cz midline activity
central C4 activity	articulatory motor cortex activity
temporal Cz activity	frontotemporal theta connectivity
frontal contralateral activity	frontotemporal gamma coherence
central ipsilateral activity	temporal beta variability
temporal left-right asymmetry	temporal complexity
temporal gamma power	central spectral entropy
frontocentral theta suppression	frontocentral waveform complexity
central alpha power	waveform amplitude
frontal mu activity	Hjorth mobility
central theta activity	Hjorth complexity
temporal alpha activity	waveform regularity
central gamma activity	frontocentral beta connectivity
temporal mu activity	central–temporal theta connectivity
temporal theta power	temporal spectral entropy
frontal alpha power	frontotemporal alpha connectivity
central beta power	central mu temporal variance
frontal theta activity	temporal gamma variability
central mu activity	frontotemporal phase synchrony
temporal beta activity	EEG peak-to-trough amplitude
frontal beta power	
frontocentral alpha power	
frontal theta power	
frontocentral gamma power	