BALANCED LATENT SEMANTICS AND SIGNAL FIDELITY FOR EEG REPRESENTATION LEARNING

Van-Chien Nguyen[1], **Trung-Hieu Tran**[1], **Tuan-Kiet Doan**[1], **Quang Hung Pham**[2], **Ngoc-Son Vu**[3], **Duc Han Le**[1], **Huy Phan**[4], **Phi Le Nguyen**[2], **Nikola Simidjievski**[1], **Samuel Tardieu**[1], **Van-Tam Nguyen**[1]
[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2]SoICT, Hanoi University of Science and Technology, Vietnam
[3]Université de Technologie de Troyes, France
[4]Reality Labs, Meta
{name.surname}@telecom-paris.fr

ABSTRACT

Electroencephalography (EEG) is critical for neurological diagnosis but suffers from low SNR and subject variability. Current foundation models relying on raw signal reconstruction often overfit to local noise. We propose STELAR, a foundation model with a dual-space objective combining patch-level masked latent prediction for semantic stability with masked reconstruction for raw signal fidelity. To balance these objectives, we introduce MTPE-GB, a validation-driven gradient balancer that adaptively weights tasks without manual tuning or computational overhead. STELAR achieves state-of-the-art linear probing performance across diverse EEG benchmarks, demonstrating robust generalization. Source code is available at: https://github.com/nvc45421/stelar-tsalm-2026.

Track: Research

1 INTRODUCTION

Electroencephalography (EEG) is critical for neurological diagnostics, yet its inherent non-stationarity, low signal-to-noise ratio (SNR), and subject variability pose significant challenges for machine learning. While task-specific pipelines have achieved isolated successes, they lack the scalability required for general-purpose application. This has motivated a shift toward EEG foundation models. However, most existing self-supervised approaches rely on masked raw signal reconstruction (Wang et al., 2024b; Zhou et al., 2025; Döner et al., 2025). While this objective maintains local waveform fidelity, it often biases the encoder toward high-frequency noise and acquisition artifacts rather than the high-level semantic representations necessary for broad generalization.

To address this, we propose STELAR (Spatio-Temporal EEG Latent Alignment & Reconstruction), a framework that employs a dual-space self-supervised objective. STELAR decouples semantic abstraction from physical grounding by performing patch-level masked prediction in the *latent space*, forcing the encoder to capture stable semantic structures, while simultaneously using *signal reconstruction* as a fidelity regularizer. This ensures the learned features retain physical waveform characteristics without overfitting to trivial low-level noise.

Optimizing these conflicting objectives is non-trivial; naive multi-task weighting often leads to task competition, while existing gradient-balancing methods impose Chen et al. (2018); Yu et al. (2020); Sener & Koltun (2018) prohibitive computational overheads, often requiring computing second-order derivatives, storing high-dimensional gradient vectors, or performing multiple backward passes per iteration. Consequently, we introduce the Multi-Task Pre-training EEG Gradient Balancer (MTPE-GB). This validation-driven scheme dynamically adjusts task weights based on relative learning progress, achieving the stability of gradient alignment with negligible computational cost.

Our **contributions** are: (1) A dual-space objective harmonizing semantic and signal fidelity; (2) The almost zero-overhead MTPE-GB algorithm for adaptive loss balancing; and (3) Extensive validation showing SOTA linear probing performance.

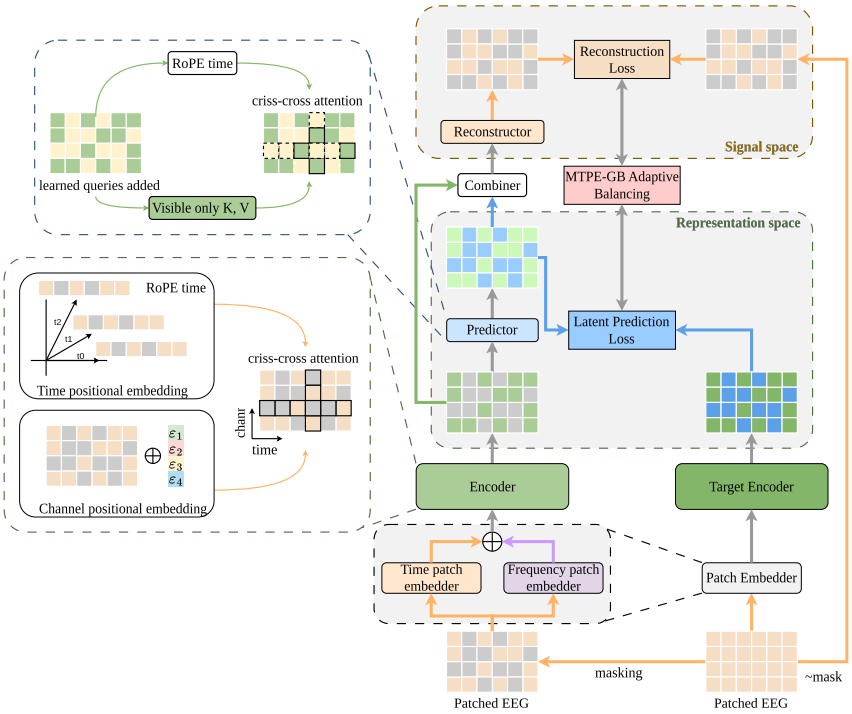


Figure 1: Schematic of the STELAR framework: teacher-guided patch-level masked latent prediction supervised by a target encoder (EMA updated); lightweight masked signal reconstruction as a fidelity regularizer; and validation-driven adaptive balancing between objectives.

2 METHODOLOGY

2.1 ARCHITECTURE OVERVIEW

The STELAR framework (Figure 1) introduces a dual-space pre-training paradigm designed to reconcile high-level semantic abstraction with low-level physical fidelity. The architecture consists of four primary components: (i) an Online Encoder (E_θ) processing visible data; (ii) a Target Encoder (E_ξ, updated via EMA) acting as a stable teacher; (iii) a lightweight Latent Predictor (P_ϕ); and (iv) a Signal Reconstructor (R_ψ).

Embeddings and Masking. Input EEG signals $S \in \mathbb{R}^{C \times T}$ are segmented into non-overlapping patches and projected into d-dimensional embeddings. Following Jiang et al. (2024), we employ a dual-branch embedder (temporal convolutions + frequency FFT-MLP) to capture both morphological and oscillatory features. We apply a 50% random masking ratio; E_θ processes only visible patches, while E_ξ encodes the full unmasked signal to generate semantic targets.

Encoder Backbone. To efficiently model long-range dependencies, we utilize the Criss-Cross Attention architecture (Wang et al., 2024b) augmented with Rotary Positional Embeddings (RoPE) and learnable channel embeddings. This decomposes global attention into orthogonal spatial (cross-channel) and temporal (cross-time) operations, significantly reducing computational complexity while preserving spatio-temporal context.

2.2 DUAL-SPACE OBJECTIVE

To reconcile semantic stability with physical fidelity, STELAR optimizes a joint objective comprising two complementary tasks. **Masked Latent Prediction ($\mathcal{L}_{\text{lat}}$)** enforces semantic abstraction by regressing masked patches to their EMA teacher representations, preventing the model from learning trivial low-level shortcuts. Simultaneously, **Masked Signal Reconstruction ($\mathcal{L}_{\text{rec}}$)** acts as a fidelity

regularizer by recovering raw waveforms to prevent dimensional collapse. The total loss is a dynamic weighted sum: $\mathcal{L}_{\text{total}} = w_{\text{lat}}\mathcal{L}_{\text{lat}} + w_{\text{rec}}\mathcal{L}_{\text{rec}}$.

2.3 Multi-Task Pre-training EEG Gradient Balancer (MTPE-GB)

Balancing semantic and reconstruction objectives is non-trivial, as the "easier" reconstruction task often dominates gradient flow. Manual tuning is inefficient, while existing gradient-alignment methods (Chen et al., 2018; Sener & Koltun, 2018) impose prohibitive computational costs by requiring second-order derivatives, storing high-dimensional gradient vectors or multiple backward passes per step. To address this, we introduce **MTPE-GB**, a validation-driven weighting scheme with **near-zero computational overhead**, operates solely on scalar validation loss already available in the training loop. This ensures the balancing process adds negligible cost to the training loop.

Algorithm 1 Multi-Task Pre-training EEG Gradient Balancer (MTPE-GB)

Input: Model Θ, k tasks, M updates, smoothing c.
Initialize $w_i(0) = 1/k$. Compute initial losses $\mathcal{L}_i^V(0)$.
for update $t = 1$ **to** M **do**
  Train epoch; validate to get current losses $\mathcal{L}_i^V(t)$.
  Compute progress: $u_i(t) = (\mathcal{L}_i^V(0) - \mathcal{L}_i^V(t))/\mathcal{L}_i^V(0)$
  Compute raw weights: $\overline{w_i}(t+1) = \frac{\sum_{j=1}^{k} u_j^{-1}(t)}{u_i^{-1}(t)}$ s.t. $\sum \overline{w_i} = 1$
  Smooth updates: $w_i(t+1) \leftarrow c \cdot \overline{w_i}(t+1) + (1-c) \cdot w_i(t)$
end for

3 Experiments

3.1 Experimental Setup

Pre-training & Protocol. STELAR was pre-trained on a curated 500-hour subset of the TUEG dataset (Obeid & Picone, 2016b) for 30 epochs using a 50% masking ratio. We employed a standard pre-processing pipeline (0.3–75 Hz filtering, notch filter at 60Hz, 200 Hz resampling). To ensure rigorous evaluation, we utilized strictly subject-wise cross-validation, preventing data leakage. Our primary metric is **linear probing** (freezing the encoder) to isolate intrinsic representation quality, reporting balanced accuracy and standard task-specific metrics.

Datasets & Baselines. We evaluated performance across six tasks on seven datasets: *Motor Imagery* (BCIC-2A Brunner et al. (2008), PhysioNet-MI Goldberger et al. (2000); Schalk et al. (2004)), *Emotion Recognition* (FACED Chen et al. (2023)), *Abnormality Detection* (TUAB Obeid & Picone (2016a)), *Sleep Staging* (Sleep-EDF Kemp et al. (2000)), and *Error-Related Potentials* (Kaggle-ERN Margaux et al. (2012)). All recordings were mapped to a unified 19-channel montage via a 1×1 convolutional adapter. We compared STELAR against state-of-the-art EEG foundation models including EEGPT (Wang et al., 2024a), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2024b), and CSBrain (Zhou et al., 2025). Details are provided in Appendix E, F.

3.2 Results and Analysis

State-of-the-Art Performance. As shown in Table 1, STELAR achieves the best performance on most of metric points across the seven datasets, creates a clear gap (up to 6%) with SOTA in BCIC-2A, PhyioMI, FACED, and Sleep-EDFx datasets while staying competitive on others. Notably, purely raw-reconstruction models (CBraMod, CSBrain) struggle significantly on complex cognitive tasks like Emotion Recognition (FACED) under linear probing, while STELAR maintains high accuracy. This strengthens our hypothesis that signal-space objectives prioritize local waveform fidelity at the expense of the high-level semantic abstraction required for tasks with subtle cognitive signatures.

Granularity Matters. STELAR outperforms EEGPT on 5/7 datasets while remaining competitive on the others. We attribute this to our **patch-level** latent supervision, which enforces dense spatio-temporal modeling compared to EEGPT's coarse summary-token approach.

Table 1: Linear probe (lp) results across various downstream tasks (average of 3 seeds)

Datasets	Metrics	EEGPT-lp (25M)	LaBraM-lp (5.8M)	CBraMod-lp (3.9M)	CSBrain-lp (8.9M)	STELAR-lp (0.1M)
BCIC-2A	B-Acc	0.5327 ± 0.0151	0.3130 ± 0.0004	0.2753 ± 0.0196	0.4719 ± 0.0005	**0.5905 ± 0.0016**
	C-Kappa	0.3770 ± 0.0202	0.0840 ± 0.0005	0.0338 ± 0.0262	0.2958 ± 0.0007	**0.4540 ± 0.0021**
	W-F1	0.5141 ± 0.0164	0.2810 ± 0.0008	0.1630 ± 0.0474	0.4555 ± 0.0006	**0.5778 ± 0.0031**
PhysioMI	B-Acc	0.5467 ± 0.0221	0.2782 ± 0.0055	0.3978 ± 0.0562	0.5114 ± 0.0325	**0.5998 ± 0.0297**
	C-Kappa	0.3956 ± 0.0295	0.0376 ± 0.0073	0.1971 ± 0.0751	0.3485 ± 0.0433	**0.4663 ± 0.0395**
	W-F1	0.5487 ± 0.0208	0.2683 ± 0.0084	0.3570 ± 0.0870	0.5097 ± 0.0314	**0.6004 ± 0.0303**
KaggleERN	B-Acc	**0.5575 ± 0.0074**	0.4998 ± 0.0001	0.5002 ± 0.0014	0.5428 ± 0.0014	0.5460 ± 0.0114
	AUC-PR	**0.7847 ± 0.0044**	0.7019 ± 0.0075	0.7223 ± 0.0087	0.7692 ± 0.0009	0.7840 ± 0.0060
	AUROC	**0.6194 ± 0.0042**	0.4830 ± 0.0124	0.5081 ± 0.0199	0.5814 ± 0.0019	0.6105 ± 0.0046
FACED	B-Acc	0.4580 ± 0.0037	0.1779 ± 0.0042	0.1206 ± 0.0051	0.1533 ± 0.0322	**0.4963 ± 0.0026**
	C-Kappa	0.3873 ± 0.0044	0.0773 ± 0.0048	0.0119 ± 0.0056	0.0512 ± 0.0383	**0.4302 ± 0.0033**
	W-F1	0.4548 ± 0.0044	0.1687 ± 0.0066	0.0530 ± 0.0085	0.0936 ± 0.0414	**0.4933 ± 0.0031**
Sleep-EDFx	B-Acc	0.6010 ± 0.0172	0.6252 ± 0.0470	0.5091 ± 0.1162	0.5817 ± 0.0143	**0.6806 ± 0.0052**
	C-Kappa	0.5930 ± 0.0310	0.6023 ± 0.0575	0.4692 ± 0.1378	0.5430 ± 0.0114	**0.6682 ± 0.0047**
	W-F1	0.6794 ± 0.0225	0.6988 ± 0.0507	0.5741 ± 0.1126	0.6456 ± 0.0130	**0.7463 ± 0.0035**
TUEV	B-Acc	0.4213 ± 0.0122	0.4409 ± 0.0107	0.4422 ± 0.0122	**0.5236 ± 0.0100**	0.5148 ± 0.0299
	C-Kappa	0.4197 ± 0.0062	0.5224 ± 0.0113	0.4985 ± 0.0189	**0.5684 ± 0.0202**	0.5506 ± 0.0251
	W-F1	0.6805 ± 0.0062	0.7476 ± 0.0071	0.7358 ± 0.0069	0.7727 ± 0.0098	**0.7745 ± 0.0124**
TUAB	B-Acc	**0.7981 ± 0.0190**	0.7748 ± 0.0007	0.6455 ± 0.0765	0.7601 ± 0.0023	0.7958 ± 0.0108
	AUC-PR	**0.8824 ± 0.0200**	0.8603 ± 0.0006	0.6551 ± 0.1225	0.8462 ± 0.0012	0.8676 ± 0.0131
	AUROC	**0.8772 ± 0.0183**	0.8579 ± 0.0007	0.6942 ± 0.1006	0.8409 ± 0.0011	0.8745 ± 0.0073

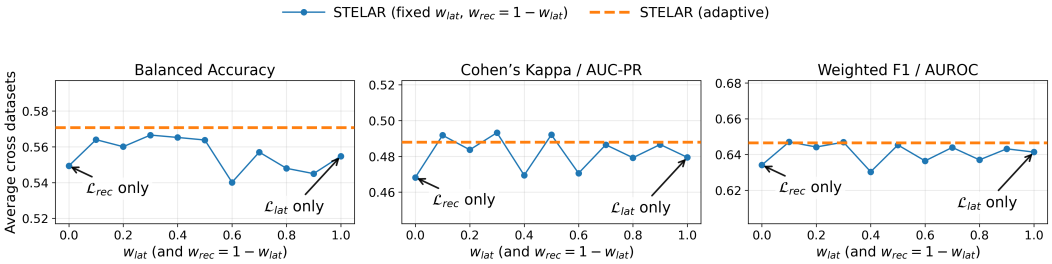


Figure 2: Ablation on manual losses tuning and adaptive losses tuning MTPE-GB across 3 datasets BCIC2A, PhysioMI, and TUEV

3.3 Efficacy of MTPE-GB

We validated the Dual-Space + MTPE-GB strategy by comparing it against fixed-weight combinations (w_{lat}, w_{rec}) and single-objective baselines.

Failure of Single Objectives. Extreme settings ($w_{lat} = 1$ or $w_{rec} = 1$) yield suboptimal results (Fig. 2 in Appendix).

MTPE-GB vs. Manual Tuning. MTPE-GB matches or exceeds the best manually tuned fixed-weight configurations across BCIC-2A, PhysioMI, and TUEV. This demonstrates that dynamically balancing objectives based on validation progress effectively captures the synergy of dual-space learning without expensive hyperparameter search.

4 Conclusion

We introduced STELAR, an EEG foundation model combining teacher-guided masked latent prediction with signal reconstruction to learn semantic abstraction while keeping signal fidelity. To stabilize this dual-space optimization, we proposed MTPE-GB, a validation-driven balancing strategy that eliminates manual tuning with negligible overhead. Extensive linear probing across diverse EEG tasks confirms that STELAR yields robust, transferable representations, consistently outperforming larger state-of-the-art models. Our findings highlight that prioritizing patch-level latent supervision and adaptive task balancing is key to efficient general-purpose EEG modeling.

REFERENCES

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, 2008.

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.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Federico Del Pup, Andrea Zanola, Louis Fabrice Tshimanga, Alessandra Bertoldo, Livio Finos, and Manfredo Atzori. The role of data partitioning on the performance of eeg-based deep learning models in supervised cross-subject analysis: A preliminary study. *Computers in Biology and Medicine*, 196:110608, 2025. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2025.110608. URL https://www.sciencedirect.com/science/article/pii/S001048252500959X.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. Luna: Efficient and topology-agnostic foundation model for eeg signal analysis. *arXiv preprint arXiv:2510.22257*, 2025.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. Eeg2rep: Enhancing self-supervised eeg representation through informative masked inputs. *arXiv preprint arXiv:2402.17772*, 2024.

Navid Mohammadi Foumani, Soheila Ghane, Nam Nguyen, Mahsa Salehi, Geoffrey I Webb, and Geoffrey Mackellar. Eeg-x: Device-agnostic and noise-robust foundation model for eeg. *arXiv preprint arXiv:2511.08861*, 2025.

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.

Weibang Jiang, Liming Zhao, and Bao liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

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.

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.

Xiang Lan, Hanshu Yan, Shenda Hong, and Mengling Feng. Towards enhancing time series contrastive learning: A dynamic bad pair mining approach. *arXiv preprint arXiv:2302.03357*, 2023.

Cheol-Hui Lee, Hakseung Kim, Hyun-jee Han, Min-Kyung Jung, Byung C Yoon, and Dong-Joo Kim. Neuronet: A novel hybrid self-supervised learning framework for sleep stage classification using single-channel eeg. *arXiv preprint arXiv:2404.17585*, 2024.

Perrin Margaux, Maby Emmanuel, Daligault Sébastien, Bertrand Olivier, and Mattout Jérémie. Objective and subjective evaluation of online error correction during p300-based spelling. *Advances in Human-Computer Interaction*, 2012(1):578295, 2012.

Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In *Machine learning for health*, pp. 238–253. PMLR, 2020.

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

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in Neuroscience*, 10:196, 2016b. doi: 10.3389/fnins.2016.00196.

Mathis Rezzouk, Fabrice Gagnon, Alyson Champagne, Mathieu Roy, Philippe Albouy, Michel-Pierre Coll, and Cem Subakan. Towards generalizable learning models for eeg-based identification of pain perception, 2025. URL `https://arxiv.org/abs/2508.11691`.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

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*, 2024a. 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. *arXiv preprint arXiv:2412.07236*, 2024b.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12695–12705, 2020.

Wei Xiong, Jiangtong Li, Jie Li, and Kun Zhu. Eeg-fm-bench: A comprehensive benchmark for the systematic evaluation of eeg foundation models. *arXiv preprint arXiv:2508.17742*, 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, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. *arXiv preprint arXiv:2506.23075*, 2025.

A Related Work

SSL for EEG has evolved across three primary paradigms: (i) Generative Masked Modeling, which utilizes masked signal reconstruction. While effective for capturing local dependencies, they often over-index on low-level noise and acquisition artifacts at the expense of high-level semantics; (ii) Discriminative and Alignment Objectives, which aim to learn invariant features but they can be brittle when faced with the non-stationary, low-SNR nature of raw EEG; and (iii) Hybrid Schemes, which combine multiple losses but typically rely on static weighting or expensive auxiliary networks, making them difficult to scale across diverse datasets. STELAR addresses these limitations by decoupling representation learning (predicting masked latent features) from signal preservation (masked reconstruction). Unlike prior hybrid methods, STELAR uses MTPE-GB to dynamically balance these two objectives with minimal extra computation.

Generative Masked Model. Early efforts such as BENDR (Kostas et al., 2021) pioneered masked signal reconstruction for EEG, a paradigm later refined by LaBraM (Jiang et al., 2024) through neural tokenization and discretized targets. Subsequent works have introduced specialized inductive biases to handle the complexity of multi-channel neural data: CBraMod (Wang et al., 2024b) utilizes Criss-Cross attention to capture spatio-temporal interactions, while CSBrain Zhou et al. (2025) employs cross-scale tokenization (CST) and structured sparse attention (SSA) to model regional brain dependencies. To address montage variability, LUNA Döner et al. (2025) proposes a topology-invariant encoder that projects arbitrary channel sets into a unified latent space. Although these architectural refinements yield strong full-model fine-tuning performance, their reliance on masked reconstruction objectives can lead the model to overly focus on local waveform details. As noted in recent literature (Jiang et al., 2024; Wang et al., 2024b; Xiong et al., 2025), this emphasis on low-level signal fidelity often fails to yield the robust, transferable abstractions required for high-performance linear probing.

Discriminative and Alignment Objectives. Alignment frameworks aim to learn invariant representations by projecting EEG into a unified latent space (Mohsenvand et al., 2020; Yang et al., 2023). Recently, EEG2Rep (Foumani et al., 2024) advanced this paradigm by introducing masked latent prediction. By aligning masked segments with their latent representations in a self-prediction task, EEG2Rep bypasses the low SNR of raw signals and avoids the need for explicitly curated negative samples. Despite these advancements, defining stable alignment targets in non-stationary EEG remains challenging, as the absence of clear semantic boundaries can lead to representational collapse (Lan et al., 2023). STELAR builds on the patch-level latent prediction concept but differs by grounding these abstractions with raw signal reconstruction in a dual-space framework to prevent degenerate solutions.

Hybrid Schemes. Recent efforts have sought to synthesize generative and alignment advantages by combining multiple pre-training objectives (Lee et al., 2024; Wang et al., 2024a). For instance, EEG-X (Foumani et al., 2025) employs a noise-aware masking-reconstruction strategy in both raw and latent spaces, while EEGPT (Wang et al., 2024a) utilizes a momentum teacher for representation matching alongside signal reconstruction. Despite their promise, these frameworks face two critical limitations: (i) *Coarse-Grained Supervision*: Methods like EEGPT often align representations only at a summary token level, which can overlook the fine-grained spatio-temporal dynamics essential for capturing transient neural events. (ii) *Optimization Bottlenecks*: Existing hybrid schemes typically rely on fixed loss-weighting, which necessitates exhaustive manual tuning of hyperparameters. Such static approaches cannot account for the shifting scales and competing convergence rates of reconstruction versus alignment tasks throughout the pre-training phase, which may lead to suboptimal shared representations. While Gradient Blending (Wang et al., 2020) addressed a similar problem in multi-modal classification by weighting modalities based on their standalone optimization functions (the ratio of generalization to overfitting), it focuses on balancing input types rather than SSL objectives.

STELAR re-frames this process by prioritizing patch-level masked latent prediction as the primary driver of transfer while incorporating masked signal reconstruction as a fidelity regularizer. To harmonize these objectives, we introduce MTPE-GB, an original adaptive balancing scheme. Unlike the standalone approach of Gradient Blending, MTPE-GB is an online, behavior-based algorithm that adjusts weights by monitoring the model's actual learning speed across objectives. This ensures stable joint optimization and balanced representations without the need for manual weight search.

Table 2: Pretraining computation cost

Method	Total Params	Encoder Params	MFLOPs/step
LaBraM	5.992M	5.8M	2,583
EEGPT	76M	25M	25,000
CBraMod	3.95M	3.9M	16,679
CSBrain	8.902M	8.9M	30,913
STELAR (ours)	**263K**	**112K**	**635**

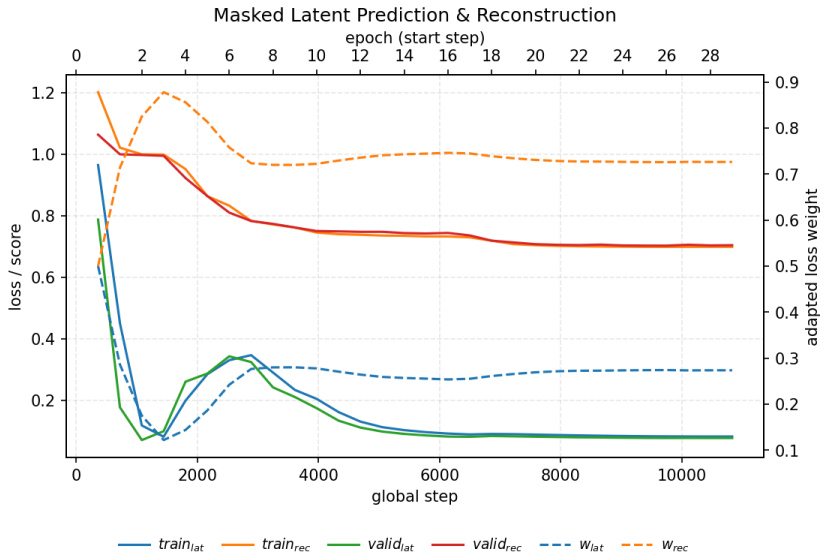


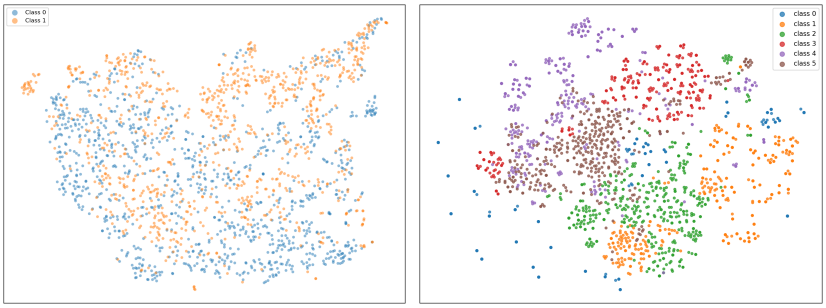
Figure 3: Pre-training loss curve of the STELAR over 30 epochs.

B Efficiency

Unlike contemporary EEG foundation models that rely on deep, large encoders and decoders, STELAR achieves state-of-the-art results with a significantly smaller total footprint (Table 2). Both our predictor and reconstructor are restricted to a single Transformer layer, and our core encoder is deliberately optimized to be more lightweight than existing baselines. Despite this overall reduction in parameter count, STELAR consistently outperforms bulkier models in linear-probing evaluation. This result demonstrates high representational density: our dual-space objective effectively concentrates semantic and morphological information into a lean architecture without requiring the massive scale of prior hybrid models. Furthermore, the MTPE-GB balancing rule enables these gains with negligible computational overhead, eliminating the need for manual loss tuning.

C Visualization

Pre-training Dynamics and Adaptive Balancing. The pre-training curves in Figure 3 illustrate the interaction between the dual-space objectives and the adaptive weighting of MTPE-GB. The model achieves rapid and stable convergence around epoch 15. A behavioral analysis of the loss curves reveals that the two objectives exhibit different convergence rates. To prevent the "easier" task from dominating the shared encoder—a phenomenon observed in multi-modal learning where signals with different learning speeds can cause suboptimality (Wang et al., 2020)—MTPE-GB dynamically recalibrates the loss weights. Unlike methods that rely on standalone optimization functions or fixed ratios, MTPE-GB monitors the real-time learning speed of each objective. As shown in Figure 3, the balancer automatically adjusts the pressure between latent prediction and signal reconstruction,



(a) STELAR embeddings on TUAB (b) STELAR embeddings on TUEV

Figure 4: t-SNE visualizations of STELAR embeddings on (a) TUAB and (b) TUEV.

ensuring both semantic transfer and physical grounding are optimized in harmony without manual intervention.

Latent space analysis. We provide a visualization of the STELAR encoder representations. As shown in Figure 4b, pretrained STELAR successfully clusters the representations of the TUEV and TUAB datasets, demonstrating an effective pre-training.

D Limitations

While STELAR demonstrates state-of-the-art representational quality with a compact model size, several areas for further investigation exist. First, our pre-training was limited to a curated 500-hour subset of EEG data. While this scale was probably sufficient to prove the effectiveness of the dual-space objective, it remains a modest fraction of the massive, diverse corpora used in other domains, like natural language processing. Future work should investigate the "scaling laws" of the STELAR framework, evaluating whether our adaptive balancing scheme and latent alignment objectives continue to perform well when scaled to, for example, tens of thousands of hours of data.

Second, the upper limits of STELAR scalability remain an open question. It is yet to be determined if significantly increasing the model's depth and width will allow the dual-space objective to capture even more complex neural dynamics. Additionally, while our channel adapter effectively handles varying electrode montages, exploring strict permutation-equivariant architectures could further enhance robustness to non-standard channel configurations.

Third, while our proposed MTPE-GB algorithm effectively stabilizes dual-space pre-training and eliminates the burden of manual weight tuning, its evaluation is currently focused on the two-task setting (reconstruction and latent prediction). We have not yet conducted a comprehensive benchmarking of MTPE-GB against computationally intensive multi-task optimization (MTO) frameworks in other domains (Chen et al., 2018; Yu et al., 2020; Sener & Koltun, 2018). Although MTPE-GB offers a significant architectural advantage by operating with nearly zero additional computational overhead, leveraging existing validation metrics rather than expensive gradient-norm calculations, further investigation is required to assess its convergence stability in high-complexity scenarios. Future work will explore its robustness in settings involving a larger number of heterogeneous objectives or more severe gradient interference.

E Pretraining setup

STELAR was implemented with PyTorch 2.1.2 and CUDA 11.8. After preprocessing, each 30s EEG segment, sampled at 200 Hz with 19 channels, is regarded as one sample. We define the length of each patch as 200 time points, equivalent to 1s of EEG data. Patches were randomly masked with a ratio of 50%. 10% of the pretraining dataset was held out for validation. Pretrained model was optimized by AdamW optimizer with learning rate 1e-4 for 30 training epochs using a batch size of 128. STELAR is trained on NVIDIA A100 40GB GPUs.

Table 3: Hyperparameters for STELAR pre-training.

Component	Hyperparameters	Settings
EEG sample	Channels	19
	Time points	6000 (30s @ 200Hz)
	Patch dimension	200
	Sequence length	30 patches
	Mask ratio	50%
	Mask token	Constant
Encoder	Layers (depth)	2
	Hidden dimension	64
	Attention heads	2
	MLP ratio	4.0
	qkv bias	True
	Init std	0.02
	Norm ϵ	1×10^{-6}
	Attention pattern	Criss-Cross
Predictor	Depth	1 layer
	KV mode	Global visible-only K/V
	Attention	Criss-Cross
Reconstructor	Depth	1 layer
	Attention	Full
Momentum Encoder	EMA momentum	Cosine ramp $0.996 \rightarrow 1$
Pre-training	Epochs	30
	Batch size	64 (global)
	Optimizer	AdamW
	Learning rate	1×10^{-4}
	LR schedule	Warmup + Cosine decay
	Warmup steps	$\max(500, 0.1T)$
	Weight decay	1×10^{-4} (cosine schedule)
	Adam β	(0.9, 0.999)
	Adam ϵ	1×10^{-8}
	Gradient clipping	1.0

We next describe the experimental settings used for STELAR pre-training. In line with CBraMod (Wang et al., 2024b), all EEG recordings are segmented into 30-second samples. This window length is notably longer than that employed in prior work, such as BIOT (Yang et al., 2023) (10 seconds) and LaBraM (Jiang et al., 2024) (4–8 seconds). We choose 30-second segments for two main reasons: (1) they provide the model with longer temporal contexts, enabling the capture of long-term dependencies that have been shown to improve downstream performance (Kostas et al., 2021); and (2) the 30-second duration closely matches the segment lengths commonly used in the downstream tasks evaluated in this study, ensuring consistency between pre-training and fine-tuning. A complete list of pre-training hyperparameters is provided in Table 3.

F Downstream, Evaluation & Setup

Comprehensive Evaluation. We observe that the performance of EEG models is strongly affected by subject partition during evaluation, which leads to considerable variability across different selections (Del Pup et al., 2025; Rezzouk et al., 2025). For fair evaluation, we have extensively built a **subject-wise cross-evaluation** scheme, in which all subjects are partitioned into *N* folds for the validation set or the test set. For example, we conduct *N* fine-tunings; in each of them, one fold is held out as the test set while the remaining folds are used for training and validation. After each fine-tuning time, we just once test the "best validated checkpoint", which is defined by the training checkpoint with the highest monitoring metric on the validation set (Cohen's kappa for multiclassification and

Table 4: Overview of EEG datasets and their corresponding BCI tasks.

BCI Task	Dataset	Rate	#Channels	Duration	Labels
Motor Imagery	BCIC-2A	250 Hz	22	4s	4-class
	PhysioNet-MI	160 Hz	64	4s	4-class
Event-related Potentials	KaggleERN	200 Hz	56	2s	2-class
Sleep Staging	Sleep-EDFx	100 Hz	2	30s	5-class
Emotion recognition	FACED	250 Hz	32	10s	9-class
Seizure / Event Detection	TUEV	250 Hz	16	5s	4-class
Abnormal EEG Detection	TUAB	250 Hz	16	10s	2-class

AUC-PR for binary classification). Reported performance is the average across all folds. Details about train-validation-test splits of each dataset are presented below.

Downstream BCI Tasks & Preprocessing. To comprehensively evaluate STELAR, we conducted experiments on **7 tasks** using **7 different datasets**, as presented in Table 4. These datasets were recorded at various sampling rates with varying numbers of channels. To efficiently adapt pre-trained STELAR and create a universal downstreaming framework, we resample all samples to 200 Hz and scale them to the range $[-1, 1]$ as done with the pre-training dataset. Due to mis-match channel numbers between the pretraining dataset and various downstream datasets, we constructed a linear mapping to map the dataset's channels to the pre-defined channels, similar to EEGPT. Each dataset associated with a downstreaming task has specific events with different time spans so we adaptively truncate them accordingly to capture meaningful EEG samples. More details about downstreaming datasets are also presented in Table 4

Baselines. Existing state-of-the-art EEG foundation models, such as EEGPT, LaBraM, and CBraMod, CSBrain are regarded as our baselines. We reproduce their results using their best settings reported in the original works for comparison. Firstly, we preprocess the datasets following the papers' specifications, i.e., applying their corresponding sampling rates, channels, band-pass filters, sample lengths, etc. Secondly, we apply their corresponding setups for each downstream dataset, including learning rate, masking ratio, and specific additional architectures (for example, EEGPT requires an additional adapter before the encoder).

Downstream Setup. We conducted a linear probing evaluation scheme (freezing encoder + fine-tuning linear head) on the six downstream datasets. Balanced Accuracy, Cohen's Kappa, and Weighted F1 are reported for multiclass classification (BCIC2A, PhysioNet-MI, Sleep-EDFx, and TUEV), while Balanced Accuracy, AUC-PR, and AUROC are reported for binary classification (TUAB, KaggleERN). STELAR uses a linear channel adapter to adapt the specific channels of the downstream dataset to predefined channels, similar to EEGPT. For the baselines, the setups from their original works are used.

Common settings. All experiments are performed with a global batch size of 64 with seed set to 7. Optimization uses AdamW with a maximum learning rate of 5e-4 and weight decay of 0.05, following a OneCycle schedule with 20% warm-up.

F.1 BCIC-2A

Description & Preprocessing. BCIC-2A consists of data from 9 subjects doing trials of 4 different motor imagery tasks. These tasks are motor imagery of the left hand (Class 1), right hand (Class 2), feet (Class 3), and tongue (Class 4). Each subject performs two sessions on different days, with each session consisting of 288 trials. STELAR applies a band-pass filter from 0 to 38 Hz, sampling rate at 200 Hz, and 4-second window sample (800 data points).

Evaluation. We adopt a leave-one-subject-out (LOSO) cross-validation protocol. We perform 9 fine-tunings, each involving a different subject as a testing dataset, and the remaining 8 subjects serve as the training set. We report the test result of the last checkpoint.

F.2 PHYSIONET-MI

Description & Preprocessing. PhysioNet-MI is a motor imagery dataset, which consists of data from 109 subjects doing trials of 4 different motor imagery tasks. These tasks are motor imagery of the left fist (Class 1), right fist (Class 2), both fists (Class 3), and both feet (Class 4). STELAR applies a low pass filter with a cut-off frequency at 0.3 Hz, sampling rate at 200 Hz, and 4-second window sample (800 data points).

Evaluation. As PhysioNet-MI has its own evaluation set, which we regard as the test set. We adopt the proposed cross-validation protocol for validation sets by splitting all subjects into 5 folds. We then conduct 5 fine-tunings, each involving one fold of subjects as a validation set, and the remaining subjects serve as the training set.

F.3 FACED

Description & Preprocessing. The FACED dataset consists of 32-channel EEG recordings from 123 participants who watched 28 different video clips. Nine distinct emotion classes, such as amusement, terror, and neutrality, are elicited by the stimulus. Each trial is resampled to 6,000 points and divided into three non-overlapping windows, yielding samples of 2,000 data points (32×2000).

Evaluation. We use 4-fold cross-validation that is subject-independent. One fold is kept out for testing in each fold. To track performance throughout training, the remaining folds are further divided into training and validation sets.

F.4 KAGGLEERN

Description & Preprocessing. KaggleERN is an error-related potential dataset, which requires each subject to see letters and numbers (showing 36 possible items on a matrix). Each item of the character is flashed in a random order. All subjects interact with a computer interface, which produces responses to the subject's attention over words. EEG is recorded when subjects observed whether the system correctly or incorrectly responds. This dataset consists of 2 labels: Correct feedback or Erroneous feedback. STELAR applies sampling rate of 200 Hz, and 2-second window sample (400 data points).

Evaluation. As KaggleERN has its own evaluation set, which we regard as the test set. We adopt the proposed cross-validation protocol for validation sets by splitting all subjects into 5 folds. We then conduct 5 fine-tunings, each involving one fold of subjects as a validation set, and the remaining subjects serve as the training set.

F.5 TUEV

Description & Preprocessing. TUEV is a seizure detection dataset, which is a subset of TUEG. This dataset records clinical EEG segments of 6 classes: spike and sharp wave (SPSW), generalized periodic epileptiform discharges (GPED), periodic lateralized epileptiform discharges (PLED), eye movement (EYEM), artifact (ARTF), and background (BCKG). STELAR applies a band-pass filter from 0.1 Hz to 75 Hz and a notch filter at 60Hz, sampling rate of 200 Hz, and 5-second window sample (1000 data points).

Evaluation. As TUEV has its own evaluation set, which we regard as the test set. We adopt the proposed cross-validation protocol for validation sets by splitting all subjects into 4 folds. We then conduct 4 fine-tunings, each involving one fold of subjects as a validation set, and the remaining subjects serve as the training set.

F.6 SLEEP-EDFX

Description & Preprocessing. Sleep-EDFx is a sleep stage classification dataset, consisting of data from 78 healthy subjects. This dataset contains 5 classes, corresponding to 5 stages of sleep: W, N1, N2, N3, REM. STELAR applies a low-pass filter with a cut-off frequency at 30 Hz, sampling rate: 200 Hz, and 30-second window sample (6000 data points) to Sleep-EDFx.

Evaluation. We adopt the proposed subject-wise cross-validation protocol. We split the total dataset into 5 folds with the same number of subjects. We perform 5 fine-tunings, each involving a different fold as a testing dataset, and the remaining 4 folds serve as the training and validation sets. We randomly select training and validation data from these 4 folds, with a val-train ratio of 1:9.

F.7 TUAB

Description & Preprocessing. TUAB consists of 409,455 10-second samples of subjects annotated as normal or abnormal (2-label classification). STELAR applies a band-pass filter from 0.1 to 75 Hz, a notch filter at 50 Hz, sampling rate: 200 Hz, and 10-second window sample (2000 data points).

Evaluation. As TUAB has its own evaluation set, which we consider as the test set. We adopt the proposed cross-validation protocol for validation sets. We split all subjects into 4 folds of subjects. We then conduct 4 fine-tunings, each involving one fold of subjects as a validation set, and the remaining subjects serve as the training set. Generally, the train-valid-test ratio is 6:2:2.

G METRICS DESCRIPTION

In this section, we will provide the details about all metrics we used for evaluating the model's performance.

- **Balanced Accuracy.** Balance Accuracy is usually used to measure the performance of imbalanced datasets. It is defined as the mean of recall of each class in the dataset.
- **Cohen's kappa.** Cohen's kappa is a statistical metric used to measure the level of agreement between two classifiers during classification tasks. In the experiments, one classifier is the true label of the sample.
- **Weighted F1.** Weighted F1 is the average value of the F1-score of all classes, where each class's score is weighted by its number of true instances.
- **AUROC.** AUROC stands for Area Under the Receiver Operating Characteristic curve. AUROC measures the ability of a classifier to distinguish between positive and negative classes, which is often used for binary classification.
- **AUC-PR.** AUC-PR stands for Area Under the Precision-Recall Curve. AUC-PR measures the trade-off between precision and recall across different thresholds.