EWEN: A COVARIATE-AWARE FOUNDATION MODEL FOR EEG UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG foundation models have improved transferable neural representation learning, but existing EFM–LLM frameworks still largely use language models as closed-set text classifiers. We propose Ewen, a covariate-aware EEG foundation model that incorporates physiological context into EEG representations for decoding and natural-language description. Ewen converts EEG and available auxiliary physiological signals into descriptorized covariates, providing language-aligned context for interpreting neural signals. These covariates are integrated with EEG tokens through covariate-aware token geometry, connecting physiological context with the temporal and channel structure of EEG representations. The language backbone is then adapted through language-reference-guided gradient projection to control language drift during task learning. Experiments across six EEG benchmarks demonstrate improved closed-vocabulary decoding, while generation evaluations show that Ewen can describe spectral characteristics, physiological relationships, and temporal and spatial variations. Together, these results support the value of physiological context for EEG modeling and represent a step toward physiologically grounded EEG understanding.

1 INTRODUCTION

Electroencephalography (EEG) offers a non-invasive, high-temporal-resolution recording of brain activity, supporting a broad range of applications, including clinical diagnosis (Thijs et al., 2019), neuroscience (Pesaran et al., 2018), and brain-computer interfaces (Ding et al., 2025b). Traditional EEG analysis usually trains separate models for different tasks (Ma et al., 2025; Liu et al., 2024c; Ma et al., 2026a), limiting knowledge transfer across datasets (Zhou et al., 2025d). EEG Foundation Models (EFMs) address this by pretraining on large-scale unlabeled EEG resources to learn transferable representations (Pradeepkumar et al., 2026b; Ma et al., 2026b; Wang et al., 2026b). While these representations improve generalization, their downstream use is limited to assigning each EEG segment to a predefined class, collapsing neural details into a single label.

Inspired by the success of large language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2023; Achiam et al., 2023), recent EFMs have begun to couple with LLMs for signal analysis (Jiang et al., 2025c; Lu et al., 2025; Jiang et al., 2025a; Liu et al., 2026a). However, **these LLMs largely operate as text-formatted classifiers**, mapping EEG segments to predefined textual labels (Figure 1(a)). This contrasts with their role in other modalities, where LLMs have demonstrated semantic understanding beyond label prediction (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023). Huang et al. (2026) assess visual understanding in multimodal LLMs (MLLMs) through object relationships and temporal changes. Following this account, Figure 1(b–c) connects *object–environment* understanding to *EEG understanding*: describing EEG patterns and semantic environment, and comparing earlier and later signal patterns within the evaluated tasks. Realizing these capabilities in EFM-LLM frameworks poses three challenges:

1) Lacking a semantic context for understanding. Prior EFM-LLM frameworks provide the LLM with EEG-derived tokens alone (Jiang et al., 2025c; Lu et al., 2025; Jiang et al., 2025a; Liu et al., 2026a), ignoring the physiological signals that are routinely co-recorded with EEG in clinical and experimental settings (Shu et al., 2018). Without access to such surrounding context, the LLM has no basis to understand about *the broader environment in which the signal arises or how the two interact*.

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

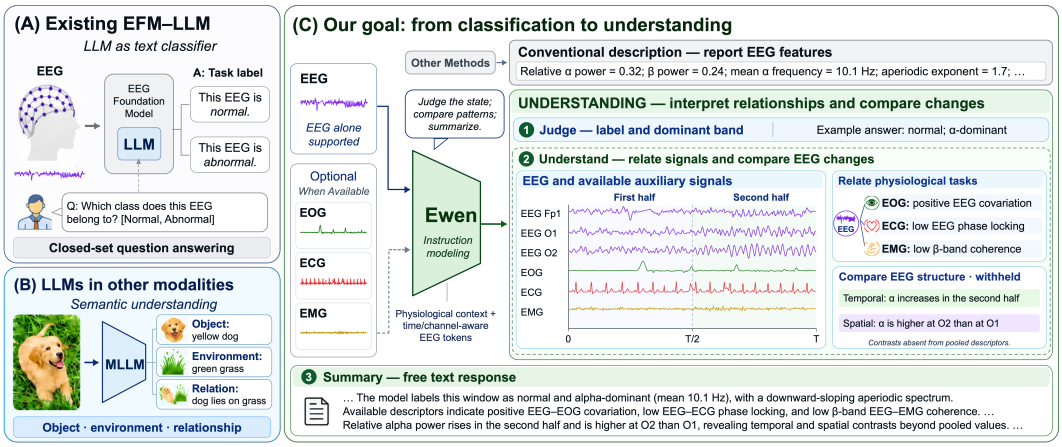

Figure 1: Conceptual motivation of our work. (a) Existing EFM-LLM frameworks use the LLM as a text-formatted classifier. (b) In contrast, MLLMs support semantic understanding the object itself, the surrounding environment, and their relationships. (c) Our goal is to shift LLM in EFM toward EEG understanding.

2) Aligning semantic information with EEG waveforms. Pairing an EEG segment with semantic information provides sample-level supervision, but leaves the correspondence between textual claims and the underlying channels and temporal patterns unspecified. This makes it difficult to learn *which waveform evidence supports a particular description*.

3) Adapting to classification tasks can cause language drift. Existing EFM-LLM frameworks adapt EEG representations to the language space through unconstrained fine-tuning or trainable projections (Jiang et al., 2025c; Lu et al., 2025; Jiang et al., 2025a; Liu et al., 2026a), optimizing primarily for EEG decoding, which is acceptable under closed-vocabulary classification. However, such unconstrained updates may shift the LLM away from its pretrained linguistic representations, raising the risk of catastrophic forgetting and degradation of general language and reasoning abilities (Luo et al., 2025; Kotha et al., 2024; Lobo et al., 2025).

As illustrated in Figure 1(c), **to shift the LLM in EFM from closed-vocabulary EEG decoding to a new paradigm of semantic EEG understanding**, we propose *Ewen*, the EEG foundation model that grounds EEG in physiological context for EEG understanding. *Ewen* first transforms auxiliary physiological signals into *descriptorized covariates*, couples them with EEG tokens through *covariate-aware token geometry*, and then adapts the language backbone with an *orthogonal update constraint* for generation. Our contributions can be summarized below.

- **Language-aligned physiological reasoning environment.** *Ewen* introduces language-aligned physiological descriptors as a semantic bridge to make heterogeneous auxiliary biosignals legible to the LLM, providing it with environment for understanding.
- **Signal-environment aligned EEG token geometry.** Rather than treating physiological descriptors as external context, *Ewen* injects them into EEG representations through covariate-aware token geometry.
- **EEG understanding with controlled language drift.** *Ewen* reformulates EEG decoding as conditional generation through language-reference projected adaptation, attenuating reference-aligned components of the core task gradient to reduce interference with pretrained semantics.
- **Comprehensive evaluation from EEG decoding to understanding.** We evaluate *Ewen* on eight EEG datasets under standard closed-vocabulary settings, where it achieves strong performance. We further assess EEG understanding through label, relation, band, and summary generation, showing its ability to produce interpretations beyond discrete task labels.

2 Related Work

Self-supervised learning has been widely explored to learn from large-scale unlabeled EEG data (Kostas et al., 2021; Mohammadi Foumani et al., 2024; Yang et al., 2023), leading to EEG foundation

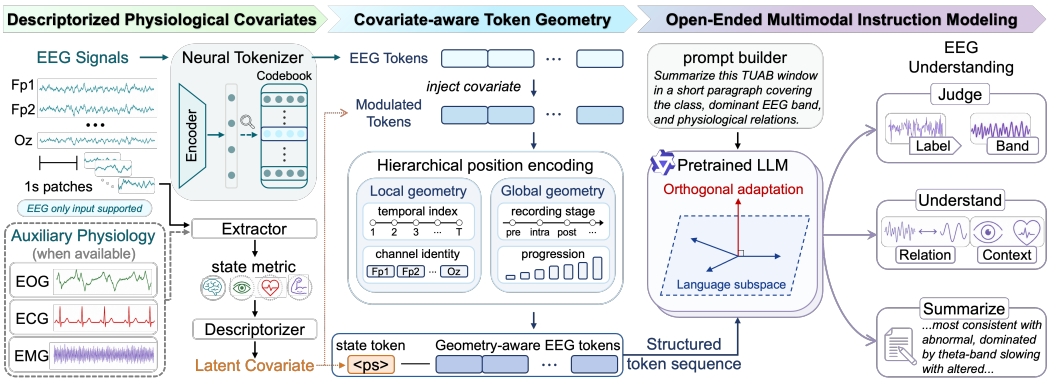

Figure 2: Architecture of *Ewen*. EEG descriptors and optional auxiliary descriptor modules provide semantic covariates that modulate EEG tokens. Local and global geometry connects signal structure to window-level physiological context. A language backbone adapted through online language-reference projection generates a physiological description.

models (EFMs) with improved transferability across tasks (Jiang et al., 2024). Subsequent work has developed a variety of architectures and pretraining strategies (Wang et al., 2024a; 2025a; Zhou et al., 2025e; Wang et al., 2025b; Chen et al., 2026b), scaling to more diverse datasets (El Ouahidi et al., 2025), accommodating heterogeneous electrode layouts (Döner et al., 2025), and incorporating neuroscience insights to enhance generalization(Ding et al., 2025a).

Motivated by the success of large language models (LLMs) (Kojima et al., 2022; Zhao et al., 2023; Zhou et al., 2025a), recent work have begun to connect EFM with LLMs. NeuroLM (Jiang et al., 2025c) treats EEG as a foreign language through neural tokenization and instruction tuning, UniMind (Lu et al., 2025) introduces a neuro-language connector with task-aware queries, LEAF (Jiang et al., 2025a) incorporates instruction-conditioned alignment with label semantics, and ECHO (Liu et al., 2026a) reformulates EEG decoding as a contextual sequence-to-sequence problem. However, these approaches remain largely label-centric. In parallel, clinically oriented models extend beyond classification. CELM (Pradeepkumar et al., 2026a) and NeuroNarrator (Wang et al., 2026a) generate clinical reports and narratives from EEG signals, and CerebraGloss (Gu et al., 2026) uses EEG visualizations for fine-grained interpretation. More detailed related work is provided in Appendix A.

3 Methodology

3.1 Problem Formulation and Solution Roadmap

Let $X^{(0)} \in \mathbb{R}^{C\times T}$ denote an EEG segment with C channels and T samples, and $\{X^{(k)}\}_{k\in\mathcal{K}}$ candidate ocular, cardiac, and muscular recordings. Availability is indicated by $\delta_k \in \{0,1\}$. The EEG is tokenized as $\mathbf{z}_{1:T_z} = \mathcal{Q}(X^{(0)})$, and the target $\mathbf{y} = (y_1, \ldots, y_N)$ is a description of its neural and physiological state. We summarize the available signals as

$$\mathbf{m} = \big[g_0(X^{(0)});\ \|_{k\in\mathcal{K}}\delta_k g_k(X^{(k)});\ \boldsymbol{\delta}\big],$$

where g_0 extracts EEG-intrinsic descriptors and g_k modality-specific indicators. Unobserved blocks are masked, with $\boldsymbol{\delta}$ distinguishing missing measurements from low values. EEG descriptors provide the base covariates; auxiliary families are plug-and-play modules. With EEG alone, $g_0(X^{(0)})$ supplies the same conditioning pathway. Appendix F lists the 18 metric types: 22 numerical values plus three auxiliary-availability indicators give $\mathbf{m} \in \mathbb{R}^{25}$.

Figure 2 summarizes our response to three challenges. ***Challenge 1: Lacking a semantic environment to reason over*** is addressed in Section 3.2 by encoding EEG and available auxiliary descriptors into a shared semantic covariate. ***Challenge 2: Aligning semantic supervision with EEG waveforms during training*** is addressed in Section 3.3 by preserving time and channel structure in EEG tokens and coupling them to the physiological context described by the paired text. ***Challenge 3: Adapting for classification with resultant language drift*** is addressed in Section 3.4 through generation and

online language-reference projected adaptation, constraining core task gradients while retaining residual EEG adaptation capacity.

3.2 Descriptorized Physiological Covariates

C1. *Lacking semantic grounding for reasoning.* EEG tokens capture neural activity, but the physiological meaning of spectral patterns and any accompanying biosignals is not readily accessible to the language backbone.

Method. Descriptor covariates make this context available in a common semantic form: EEG descriptors summarize spectral and state-related statistics, while optional EOG, ECG, and EMG modules add ocular dynamics, cardiac state, and muscular activation, respectively, together with their coupling to EEG.

Descriptor text $\mathbf{d} = \mathcal{T}(\mathbf{m})$ provides offline paired supervision for tokenizer pretraining (Appendix B.4). During downstream training and inference, a two-layer MLP maps the numerical descriptor vector directly to the semantic covariate

$$\hat{\mathbf{c}} = E_\phi(\mathbf{m}) \in \mathbb{R}^d. \tag{1}$$

This window-level context modulates each EEG token embedding $\mathbf{e}_t \in \mathbb{R}^d$ through a feature-wise affine transformation,

$$\tilde{\mathbf{e}}_t = \Gamma(\hat{\mathbf{c}}, \mathbf{r}_t) \odot \mathbf{e}_t + \mathrm{B}(\hat{\mathbf{c}}, \mathbf{r}_t), \tag{2}$$

where $\mathbf{r}_t$ is local EEG context and Γ, B produce learned scales and shifts. Their dependence on $\mathbf{r}_t$ allows the same covariate to modulate different neural patterns differently. This formulation relates to language-grounded neural descriptors (Wei et al., 2025) and covariate homogenization in time-series models (Han et al., 2026).

We encourage agreement between descriptor and neural representations through

$$\mathcal{L}_{\mathrm{cov}} = \mathbb{E}\left[\left\| h_{\mathrm{desc}}(\mathbf{m}) - h_{\mathrm{neural}}(\mathbf{z}_{1:T_z}) \right\|_2^2\right], \tag{3}$$

where h_{desc} and h_{neural} project paired inputs from the same window into a common feature space. The covariate supplies compact context while the token sequence retains detailed waveform information.

3.3 Covariate-Aware Token Geometry

C2. *Paired semantic information does not specify where its claims are supported in the waveform.* Dataset labels and narratives provide supervision at the sample level, but do not by themselves establish correspondence with particular EEG channels or temporal patterns. Training must learn this correspondence across discrete text and continuous signals.

Method. We introduce a hierarchical encoding that preserves the waveform structure needed to ground paired text, combining token-level position with window-level physiological context. For a token z_t in the input window w, τ_t and u_t denote its temporal position and source channel. The local branch is $\mathbf{p}_t^{\mathrm{loc}} = \mathrm{PE}(\tau_t) + \mathrm{IE}(u_t)$, while the global branch $\mathbf{p}_w^{\mathrm{glob}} = \Psi(\hat{\mathbf{c}}_w)$ projects the window covariate into a context shared by its EEG tokens. The resulting input state is

$$\mathbf{h}_t^{(0)} = \tilde{\mathbf{e}}_t + \mathbf{p}_t^{\mathrm{loc}} + \mathbf{p}_w^{\mathrm{glob}}. \tag{4}$$

Local geometry preserves fine-grained temporal order and channel identity; global geometry conditions these representations on the physiological state of the window. The projected covariate can also be prepended as one conditioning token. This design draws on hierarchical multimodal position handling (Xu et al., 2025b) and covariate-conditioned temporal adaptation (Han et al., 2026).

For the EEG token sequence, we compute Q, K, and V from the covariate-conditioned states $\mathbf{h}^{(0)}$ and encode temporal and channel relationships through a structured attention bias:

$$\Delta_{ij} = b_\tau(\tau_i, \tau_j) + b_u(u_i, u_j), \tag{5}$$

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + \Delta\right). \tag{6}$$

Physiological context enters attention through the token representations, while the bias preserves their temporal and channel relationships.

3.4 MULTIMODAL INSTRUCTION MODELING

C3. *Improved EEG predictions can mask changes in language behavior.* Classification-oriented adaptation may improve decoding while compromising pretrained language competence. Freezing an unchanged language pathway retains its text behavior but limits EEG adaptation to the trainable interface; updating the backbone or active adapters can change its language behavior.

Method (conditional generation). Task semantics ℓ define a prompt $\mathbf{x}^{\text{prompt}} = \mathcal{P}(\ell)$ requesting a physiological account. Descriptor values enter through $\hat{\mathbf{c}}$ rather than prompt text, while EEG tokens retain the waveform evidence. The paired target $\mathbf{y}$ describes neural patterns and available physiological relationships, including waveform-supported details beyond the supplied descriptors. We model

$$p_\theta(\mathbf{y} \mid \mathbf{z}_{1:T_z}, \mathbf{x}^{\text{prompt}}, \hat{\mathbf{c}}) = \prod_{n=1}^{N} p_\theta(y_n \mid \mathbf{z}_{1:T_z}, \mathbf{x}^{\text{prompt}}, \hat{\mathbf{c}}, y_{<n}), \tag{7}$$

and minimize

$$\mathcal{L}_{\text{gen}} = -\frac{1}{N} \sum_{n=1}^{N} \log p_\theta(y_n \mid \mathbf{z}_{1:T_z}, \mathbf{x}^{\text{prompt}}, \hat{\mathbf{c}}, y_{<n}). \tag{8}$$

Here N counts valid target tokens; batches use the corresponding token-weighted mean. This objective supervises the full physiological description.

Method (online language-reference projected adaptation). Text-only reference gradients guide a structured trainable core, complemented by a low-rank residual:

$$W = W_0 + \underbrace{\text{Diag}(\mathbf{c})\overline{W}_0}_{\Delta W_{\text{core}}} + BA, \quad \mathbf{c} \in \mathbb{R}^{d_{\text{out}}}, \quad B \in \mathbb{R}^{d_{\text{out}} \times r}, \quad A \in \mathbb{R}^{r \times d_{\text{in}}}, \tag{9}$$

Here $W_0, \overline{W}_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, $\overline{W}_{0,i:} = W_{0,i:}/\|W_{0,i:}\|_2$ is the frozen row-normalized base, and $\text{rank}(BA) \le \min(r, d_{\text{in}}, d_{\text{out}})$. The pretrained base W_0 is frozen. On every configured target in every decoder layer, θ_{core} contains one coefficient c_i per output row, and θ_{res} contains only A, B. Initialize $\mathbf{c} = \mathbf{0}$, $B = 0$, and Kaiming-uniform A to recover the pretrained decoder. EEG-side parameters θ_{EEG} receive unprojected task gradients. We use $\alpha_t = \rho\,\text{tr}(\mathbf{G}_t^\top \mathbf{G}_t)/K$ with fixed dimensionless $\rho > 0$; Appendix C.4 specifies the configuration. Both decoder branches learn from

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{gen}} + \lambda_{\text{cov}} \mathcal{L}_{\text{cov}}, \tag{10}$$

where λ_{cov} weights descriptor alignment. At step t, the task gradient with respect to core parameters θ_{core} is $\mathbf{g}_{\text{task},t} = \nabla_{\theta_{\text{core}}} \mathcal{L}_{\text{task}}$. We sample K text-only micro-batches from OpenWebText (OWT) and compute their next-token losses at the current model, forming

$$\mathbf{G}_t = [\mathbf{g}_{\text{lang},t}^{(1)}, \ldots, \mathbf{g}_{\text{lang},t}^{(K)}], \qquad \mathbf{g}_{\text{lang},t}^{(k)} = \nabla_{\theta_{\text{core}}} \mathcal{L}_{\text{lang},t}^{(k)}. \tag{11}$$

The reference gradients jointly define local directions of text-loss sensitivity for a single projection of the core task gradient, followed by one optimizer update. We modify the task gradient through

$$\mathbf{g}_{\perp,t} = \left[\mathbf{I} - \mathbf{G}_t(\mathbf{G}_t^\top \mathbf{G}_t + \alpha \mathbf{I})^{-1} \mathbf{G}_t^\top\right] \mathbf{g}_{\text{task},t}, \tag{12}$$

where $\alpha = \alpha_t > 0$ for a nonzero reference regularizes the Gram system (computed by a linear solve), yielding attenuation rather than exact removal of reference-aligned components. The core optimizer receives $\mathbf{g}_{\perp,t}$; A and B follow $\mathcal{L}_{\text{task}}$ without projection. Refreshing $\mathbf{G}_t$ online couples EEG learning to the current decoder's language reference, while the residual accommodates task structure restricted by the core constraint.

The constraint concerns the supplied core gradient and a local, first-order reference. Ridge regularization, AdamW preconditioning, momentum, weight decay, and unprojected residual updates can change language behavior.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We use eight EEG benchmarks spanning seven directions: TUAB (Abnormal Screening), TUEV (Epileptiform Events), Mental Arithmetic (Workload), SHU–MI (Motor Imagery), Mumtaz 2016 (Depression), BCIC IV 2a (Motor Imagery), BCIC 2020–T3 (Imagined Speech), and

Table 1: Downstream task evaluation results on six EEG benchmarks. Best is in **bold**.

Methods	TUAB			Mental Arithmetic		
	ACC-B	ROC AUC	PR AUC	ACC-B	ROC AUC	PR AUC
EEGNet	0.7712 ± 0.0056	0.8231 ± 0.0029	0.8501 ± 0.0027	0.5138 ± 0.0471	0.5395 ± 0.0109	0.5302 ± 0.0441
EEGPT	0.7922 ± 0.0046	0.7455 ± 0.0050	0.8662 ± 0.0079	0.5117 ± 0.0317	0.5612 ± 0.0198	0.5041 ± 0.0384
BIOT	0.7959 ± 0.0057	0.8792 ± 0.0023	0.8815 ± 0.0043	0.5281 ± 0.0384	0.5970 ± 0.0468	0.5567 ± 0.0249
LaBraM	0.8140 ± 0.0019	$\mathbf{0.9022 \pm 0.0009}$	$\mathbf{0.8965 \pm 0.0016}$	0.5793 ± 0.0631	0.5700 ± 0.0232	0.6341 ± 0.0422
CBraMod	0.7891 ± 0.0030	0.8636 ± 0.0063	0.8606 ± 0.0057	0.5906 ± 0.0531	0.5045 ± 0.0519	0.7047 ± 0.0428
NeuroLM	0.7969 ± 0.0091	0.7219 ± 0.0082	0.7884 ± 0.0194	0.6345 ± 0.0442	0.6130 ± 0.0764	0.5889 ± 0.0423
ECHO	0.7386 ± 0.0017	0.8412 ± 0.0024	0.8351 ± 0.0011	0.6851 ± 0.0032	0.7500 ± 0.0062	0.7530 ± 0.0015
Ewen-S	0.7986 ± 0.0015	0.8835 ± 0.0135	0.8690 ± 0.0177	0.7813 ± 0.0168	0.7001 ± 0.0135	0.8913 ± 0.0177
Ewen-B	0.8068 ± 0.0006	0.8922 ± 0.0124	0.8791 ± 0.0069	0.8417 ± 0.0033	0.7468 ± 0.0094	0.9235 ± 0.0052
Ewen-L	$\mathbf{0.8149 \pm 0.0186}$	0.8990 ± 0.0050	0.8867 ± 0.0046	$\mathbf{0.8790 \pm 0.0125}$	$\mathbf{0.9317 \pm 0.0080}$	$\mathbf{0.9635 \pm 0.0095}$

Methods	SHU-MI			Mumtaz 2016		
	ACC-B	ROC AUC	PR AUC	ACC-B	ROC AUC	PR AUC
EEGNet	0.6258 ± 0.0073	0.6643 ± 0.0087	0.6781 ± 0.0076	0.9113 ± 0.0104	0.9512 ± 0.0096	0.9632 ± 0.0045
EEGPT	0.5778 ± 0.0162	0.6241 ± 0.0071	0.6266 ± 0.0133	0.8475 ± 0.0233	0.9669 ± 0.0069	0.9695 ± 0.0076
BIOT	0.5605 ± 0.0252	0.5984 ± 0.0292	0.5980 ± 0.0252	0.8789 ± 0.0190	0.9664 ± 0.0136	0.9744 ± 0.0083
LaBraM	0.6084 ± 0.0175	0.6697 ± 0.0363	0.6643 ± 0.0338	0.8986 ± 0.0028	0.9754 ± 0.0050	0.9791 ± 0.0041
CBraMod	0.6196 ± 0.0048	0.6922 ± 0.0085	0.6865 ± 0.0052	0.8946 ± 0.0047	0.9800 ± 0.0045	0.9765 ± 0.0061
NeuroLM	0.5416 ± 0.0106	0.5503 ± 0.0137	0.5599 ± 0.0099	0.8935 ± 0.0048	0.9657 ± 0.0063	0.9704 ± 0.0051
ECHO	0.5406 ± 0.0009	0.5537 ± 0.0021	0.5508 ± 0.0013	0.9056 ± 0.0012	0.9745 ± 0.0023	0.9748 ± 0.0015
Ewen-S	0.6201 ± 0.0011	0.6768 ± 0.0199	0.6671 ± 0.0132	0.9146 ± 0.0125	0.9738 ± 0.0037	0.9672 ± 0.0094
Ewen-B	0.6296 ± 0.0024	0.6852 ± 0.0105	0.6774 ± 0.0130	0.9392 ± 0.0147	0.9893 ± 0.0036	0.9831 ± 0.0065
Ewen-L	$\mathbf{0.6362 \pm 0.0167}$	$\mathbf{0.6929 \pm 0.0057}$	$\mathbf{0.6869 \pm 0.0048}$	$\mathbf{0.9415 \pm 0.0124}$	$\mathbf{0.9903 \pm 0.0016}$	$\mathbf{0.9893 \pm 0.0024}$

Methods	BCIC IV 2a			SEED-V		
	ACC-B	Cohen's Kappa	Weighted F1	ACC-B	Cohen's Kappa	Weighted F1
EEGNet	0.4583 ± 0.0281	0.2937 ± 0.0612	0.4265 ± 0.0498	0.2413 ± 0.0021	0.0592 ± 0.0054	0.2317 ± 0.0017
EEGPT	0.4676 ± 0.0304	0.2889 ± 0.0529	0.4312 ± 0.0391	0.2202 ± 0.0044	0.0496 ± 0.0036	0.2301 ± 0.0096
BIOT	0.4421 ± 0.0415	0.2768 ± 0.0387	0.4180 ± 0.0634	0.2245 ± 0.0061	0.0432 ± 0.0010	0.2153 ± 0.0038
LaBraM	0.4538 ± 0.0468	0.3011 ± 0.0432	0.4147 ± 0.0587	0.2372 ± 0.0053	0.0562 ± 0.0028	0.2237 ± 0.0078
CBraMod	0.4816 ± 0.0355	0.3088 ± 0.0473	0.4571 ± 0.0543	0.2432 ± 0.0046	0.0586 ± 0.0059	0.2452 ± 0.0043
NeuroLM	0.4193 ± 0.0129	0.2199 ± 0.0184	0.4046 ± 0.0077	0.2336 ± 0.0069	0.0532 ± 0.0176	0.2167 ± 0.0038
ECHO	0.4763 ± 0.0011	0.3015 ± 0.0012	0.4632 ± 0.0002	0.2484 ± 0.0021	0.0640 ± 0.0008	0.2456 ± 0.0010
Ewen-S	0.4864 ± 0.0140	0.3178 ± 0.0047	0.4839 ± 0.0044	0.2437 ± 0.0119	0.0598 ± 0.0182	0.2418 ± 0.0078
Ewen-B	$\mathbf{0.5036 \pm 0.0110}$	0.3245 ± 0.0117	0.4914 ± 0.0055	0.2509 ± 0.0104	0.0660 ± 0.0087	0.2489 ± 0.0091
Ewen-L	0.5014 ± 0.0176	$\mathbf{0.3312 \pm 0.0173}$	$\mathbf{0.4971 \pm 0.0156}$	$\mathbf{0.2558 \pm 0.0155}$	$\mathbf{0.0711 \pm 0.0183}$	$\mathbf{0.2562 \pm 0.0004}$

SEED-V (Emotion Recognition), using the dataset-specific benchmark partitions in Appendix D. They cover binary through six-way classification with both within-subject and cross-subject splits. Six form the main comparison; TUEV and BCIC 2020–T3 appear in Appendix K. Splits and preprocessing are detailed in Appendix D. All auxiliary modality data are derived from the raw data of public datasets.

Baselines. Five domain-specific models (EEGNet, EEGPT, BIOT, LaBraM, CBraMod) and two LLM-aligned foundation models (NeuroLM, ECHO), reproduced from official codebases under matched preprocessing (Appendix E).

Model setting. *Ewen-S/B/L* use 0.8B/2B/4B Qwen3.5 backbones; the EEG-side tokenizer, covariate set, position encoder, and orthogonal adapter are shared (Appendix O). Using the same backbone family across scales keeps the adaptation comparison consistent.

Training and environment. Same instruction-tuning recipe across scales (Appendix N); training uses PyTorch DDP with bfloat16 on $8\times$A100 40GB and $4\times$H200 141GB GPUs.

Evaluation metrics. Closed-vocabulary evaluation uses Balanced Accuracy and ROC/PR AUC for binary tasks, and Cohen's Kappa and Weighted F1 for multiclass tasks. Generation evaluation reports Total F-Acc, the mean exact-match accuracy over *label*, *relation*, *band*, and *field summary*, together with BERTScore against OWT, dataset narratives, and their balanced mixture. Figure 3 additionally reports Direct F1 for explicit numeric claims, Withheld F1 for physiological attributes not explicitly supplied in descriptors, and Human overall. Table 3 reports Withheld F1, summary ROUGE-L, and OWT ΔPPL. These metrics use different targets and aggregation rules (Appendix C).

4.2 Closed Vocabulary Benchmarks

Table 1 shows that *Ewen-L* leads 15 of 18 metrics and ranks second in the remaining three. It improves all Mental Arithmetic metrics over prior baselines and leads all metrics on Mumtaz 2016.

On TUAB, Ewen-L leads ACC-B, while LaBraM has the highest ROC and PR AUC. Larger Ewen variants improve SHU–MI and SEED-V performance; on BCIC IV 2a, Ewen-B retains the best ACC-B. Appendix K.2 provides an EEG-only comparison with NeuroLM and ECHO.

4.3 Ablation Study

We ablate hierarchical position encoding (PE) and descriptorized covariate fusion (Cov.) on *Ewen-B* (Table 2). Their combination performs best across all reported metrics. PE contributes more on BCIC IV 2a, while Cov. gives larger ACC-B and ROC AUC gains on Mumtaz 2016. PE alone reduces performance on SHU–MI, but combining it with Cov. improves all metrics, suggesting that the two components provide complementary information.

Table 2: Ablation of hierarchical position encoding (PE) and descriptorized window covariate (Cov.) on *Ewen-B*.

Dataset	PE	Cov.	ACC-B ↑	ROC AUC / Cohen's Kappa ↑	PR AUC / Weighted F1 ↑
SHU-MI	✓	✓	$\mathbf{0.6296 \pm 0.0024}$	$\mathbf{0.6852 \pm 0.0105}$	$\mathbf{0.6774 \pm 0.0130}$
	✓	×	0.5700 ± 0.0142	0.5984 ± 0.0116	0.6254 ± 0.0183
	×	✓	0.6032 ± 0.0067	0.6414 ± 0.0128	0.6480 ± 0.0095
	×	×	0.5820 ± 0.0171	0.6259 ± 0.0074	0.6370 ± 0.0139
Mumtaz 2016	✓	✓	$\mathbf{0.9392 \pm 0.0147}$	$\mathbf{0.9893 \pm 0.0036}$	$\mathbf{0.9831 \pm 0.0065}$
	✓	×	0.9148 ± 0.0164	0.9765 ± 0.0121	0.9802 ± 0.0062
	×	✓	0.9293 ± 0.0039	0.9849 ± 0.0088	0.9765 ± 0.0057
	×	×	0.9078 ± 0.0108	0.9730 ± 0.0046	0.9758 ± 0.0152
BCIC IV 2a	✓	✓	$\mathbf{0.5036 \pm 0.0110}$	$\mathbf{0.3245 \pm 0.0117}$	$\mathbf{0.4914 \pm 0.0055}$
	✓	×	0.4514 ± 0.0079	0.2685 ± 0.0148	0.4477 ± 0.0112
	×	✓	0.3932 ± 0.0186	0.1910 ± 0.0063	0.3869 ± 0.0194
	×	×	0.3490 ± 0.0127	0.1319 ± 0.0091	0.3213 ± 0.0169

4.4 Understanding Physiological Signals

Protocol. We assess physiological signal understanding through generated descriptions of temporal and spatial spectral relationships on Mumtaz 2016, Mental Arithmetic, and SHU–MI. Direct targets are explicit descriptor facts; withheld targets are waveform-derived states not explicitly supplied by descriptors. Figure 3 compares input conditions, and Table 3 compares adaptation methods under fixed targets and scoring rules. Both report means over three independently trained checkpoints with different seeds; error bars and ± indicate sample SD (Appendices C and G).

Contribution of input sources. In Figure 3, Waveform and Full inputs recover withheld physiological relationships better than Descriptor or Prompt alone. Full performs best overall, while its gain over

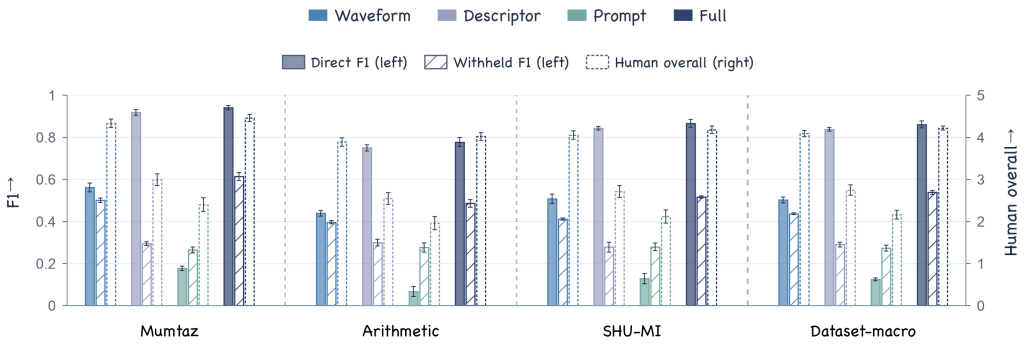

Figure 3: Physiological signal understanding across input sources on Mumtaz 2016, Mental Arithmetic, and SHU–MI, with an equally weighted dataset-macro summary. Solid outlines indicate Direct and Withheld F1 (left axis); dashed outlines indicate Human overall (right axis). Higher is better for all three metrics.

Table 3: Physiological understanding and language behavior across adaptation methods on Mumtaz 2016, Mental Arithmetic, and SHU–MI. External systems are task-aligned on the same narratives. Best means within each dataset or aggregate are bold; lower OWT ΔPPL is better.

Method	Dataset	Withheld F1 ↑	Summary R-L ↑	OWT ΔPPL ↓
NeuroLM†	Mumtaz	0.4982 ± 0.0247	0.8727 ± 0.0117	N/A
	Arithmetic	0.3625 ± 0.0164	0.8604 ± 0.0173	
	SHU–MI	0.4272 ± 0.0154	0.8757 ± 0.0061	
	Dataset-macro	0.4293 ± 0.0093	0.8696 ± 0.0101	
CELM†	Mumtaz	0.3463 ± 0.0208	0.8120 ± 0.0201	N/A
	Arithmetic	0.2111 ± 0.0179	0.5299 ± 0.0209	
	SHU–MI	0.2959 ± 0.0081	0.7482 ± 0.0135	
	Dataset-macro	0.2844 ± 0.0116	0.6967 ± 0.0109	
LoRA	Mumtaz	0.5716 ± 0.0241	$\mathbf{0.8868 \pm 0.0213}$	$\mathbf{0.0345 \pm 0.0008}$
	Arithmetic	0.4274 ± 0.0233	0.8961 ± 0.0067	
	SHU–MI	0.5013 ± 0.0162	$\mathbf{0.8947 \pm 0.0193}$	
	Dataset-macro	0.5001 ± 0.0212	0.8925 ± 0.0102	
Full FT	Mumtaz	0.5919 ± 0.0059	0.8854 ± 0.0195	0.1983 ± 0.0016
	Arithmetic	0.4293 ± 0.0095	0.8988 ± 0.0154	
	SHU–MI	0.4578 ± 0.0244	0.8891 ± 0.0247	
	Dataset-macro	0.4930 ± 0.0096	0.8911 ± 0.0096	
Ewen	Mumtaz	$\mathbf{0.6137 \pm 0.0155}$	0.8857 ± 0.0091	0.0302 ± 0.0011
	Arithmetic	$\mathbf{0.4861 \pm 0.0172}$	$\mathbf{0.8998 \pm 0.0064}$	
	SHU–MI	$\mathbf{0.5163 \pm 0.0135}$	0.8931 ± 0.0237	
	Dataset-macro	$\mathbf{0.5387 \pm 0.0120}$	$\mathbf{0.8929 \pm 0.0113}$	

†NeuroLM and CELM are our task-adapted versions, modified for the same narrative-generation task.

Waveform varies by dataset. These results support waveform-dependent interpretation of temporal and spatial spectral structure beyond restating supplied descriptors.

EEG understanding across methods. Ewen has the highest mean Withheld F1 across the evaluated datasets in Table 3, supporting more accurate recovery of waveform-dependent physiological relationships. Summary ROUGE-L is similar among LoRA, Full FT, and Ewen, indicating that text overlap alone does not reflect these differences in signal understanding. OWT ΔPPL assesses changes in text prediction separately. Expert evaluation of freely generated summaries in Appendix Q provides complementary evidence on the factual accuracy and coverage of the descriptions.

Complementary analysis. Figures 4–6 examine update geometry and generation behavior for LoRA, Full FT, and Ewen. For the task-field evaluation in Figure 5, each test window is queried on *label*, *relation*, *band*, and *summary*; answers are parsed using Appendix G. Their mean field accuracy is $\text{Total F-Acc} = \frac{1}{4}\sum_{\tau} \text{F-Acc}_{\tau}$. The three methods share the same backbone, data, and step budget; detailed results and human evaluation are provided in Appendices I and J.

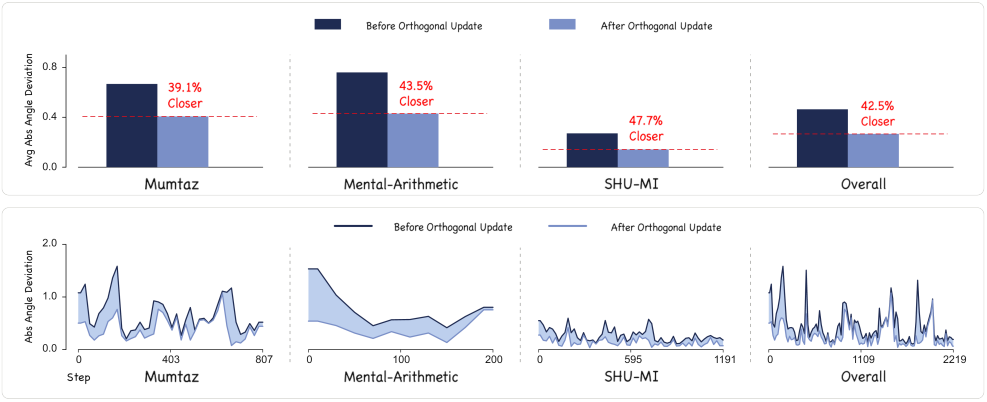

Figure 4: Absolute angular deviation from orthogonality to the language-reference direction: training-wide means (top) and per-step values (bottom).

Controlling language drift (Figure 4). Smaller angular deviations under projection indicate less alignment with the language reference, consistent with the intended constraint on language drift. This geometric diagnostic is complemented by generation and OWT evaluations (Appendix C.4).

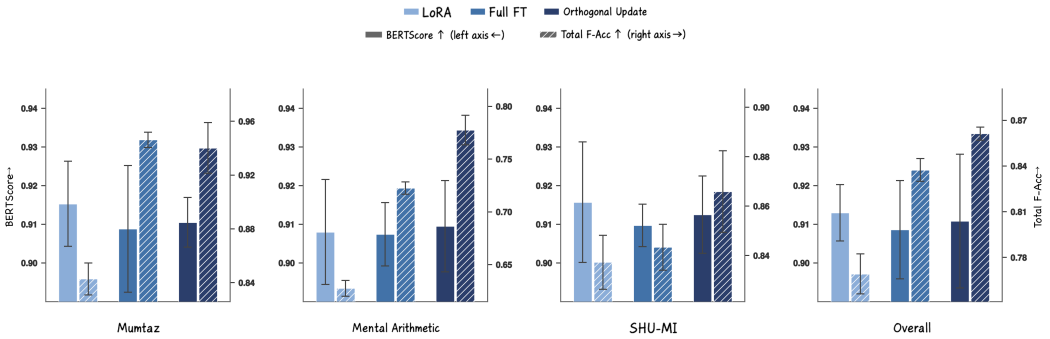

Figure 5: Generation trade-off between Mix-BERTScore (left axis) and Total F-Acc (right axis) across the three EEG datasets and their dataset-macro aggregate.

Task accuracy and language behavior trade-off (Figure 5). Ewen achieves the highest overall task-field accuracy, while LoRA leads overall Mix-BERT. This contrast illustrates that stronger physiological task performance does not necessarily improve agreement with the combined language references.

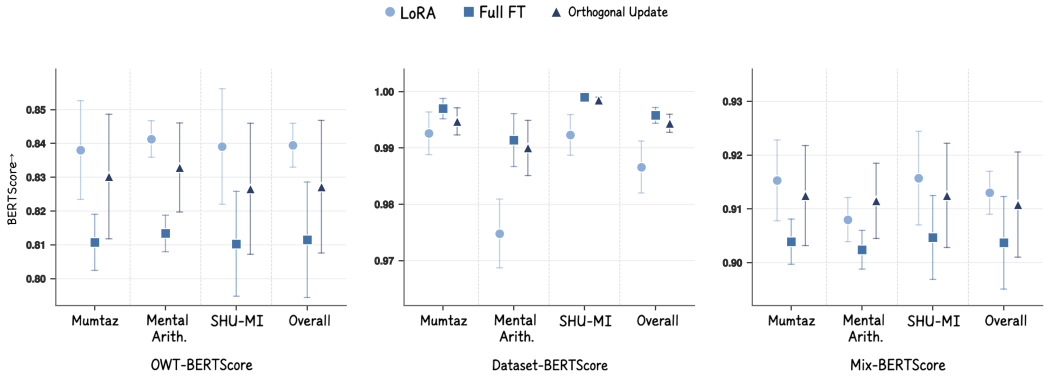

Figure 6: BERTScore across OWT text continuations, dataset-specific EEG narration, and their balanced mixture.

Language behavior beyond EEG narration (Figure 6). EEG-conditioned generations provide only a task-specific view of language behavior. We therefore distinguish OWT-BERT for non-task-specific text continuations, Dataset-BERT for dataset-specific EEG narration, and Mix-BERT for their balanced mixture. Ewen remains close to Full FT on Dataset-BERT but scores higher on OWT-BERT, highlighting the need to evaluate EEG narration and language behavior on the OWT reference domain separately.

5 Conclusion

Ewen combines descriptorized physiological covariates, covariate-aware token geometry, and language-reference projected adaptation to connect EEG signals with a pretrained language backbone. The projection attenuates reference-aligned core gradients, while the residual branch provides task-adaptation capacity. Across the listed benchmarks, Ewen achieves the strongest reported results in most comparisons. Generation experiments show a trade-off between task accuracy and reference language adaptation. Future work will extend the physiological inputs and scale the EEG–language interface.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing and LaTeX preparation. The use of language models as components of Ewen is described in Sections 3 and 4.1.

ETHICS STATEMENT

This work uses previously released EEG datasets and does not collect new EEG recordings. Dataset sources and evaluation protocols are described in Appendix D. Ewen is a research prototype; its generated descriptions are not intended for standalone clinical diagnosis. Limitations and broader impacts are discussed in Appendix P.

REPRODUCIBILITY STATEMENT

The methodology and experimental setup are described in Sections 3 and 4. The appendix provides the tokenizer, evaluation metrics, dataset protocols, descriptor construction, prompt and answer templates, training algorithm, and model configuration. Code access is described in Appendix H.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi. Emotion recognition based on eeg using lstm recurrent neural network. *International Journal of Advanced Computer Science and Applications*, 8(10), 2017.

Naseem Babu, Jimson Mathew, and AP Vinod. Large language models for eeg: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2506.06353*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Chi-Sheng Chen, Ying-Jung Chen, and Aidan Hung-Wen Tsai. Large cognition model: Towards pretrained eeg foundation model. *arXiv preprint arXiv:2502.17464*, 2025.

Yan Chen, Jie Peng, Moajjem Hossain Chowdhury, Tianlong Chen, and Yunmei Liu. Neurocanvas: Vllm-powered robust seizure detection by reformulating multichannel eeg as image. *arXiv preprint arXiv:2602.04769*, 2026a.

Zhisheng Chen, Yingwei Zhang, Qizhen Lan, Tianyu Liu, Huacan Wang, Yi Ding, Ziyu Jia, Ronghao Chen, Kun Wang, and Xinliang Zhou. Uni-ntfm: A unified foundation model for eeg signal representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher Michael Sandino, and Joseph Yitan Cheng. Maeeg: Masked auto-encoder for eeg representation learning. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023.

Mingzhe Cui, Tao Chen, Yang Jiao, Yiqin Wang, Lei Xie, Yi Pan, and Luca Mainardi. Brainrvq: A high-fidelity eeg foundation model via dual-domain residual quantization and hierarchical autoregression. *arXiv preprint arXiv:2602.16951*, 2026.

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

Xia Deng, Shen Chen, Jiale Zhou, and Lei Li. Mind2matter: Creating 3d models from eeg signals. *arXiv preprint arXiv:2504.11936*, 2025.

Yi Ding, Yong Li, Hao Sun, Rui Liu, Chengxuan Tong, Chenyu Liu, Xinliang Zhou, and Cuntai Guan. Eeg-deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 29(3):1909–1918, 2024.

Yi Ding, Muyun Jiang, Weibang Jiang, Shuailei Zhang, Xinliang Zhou, Chenyu Liu, Shanglin Li, Yong Li, and Cuntai Guan. Brainpro: Towards large-scale brain state-aware eeg representation learning. *arXiv preprint arXiv:2509.22050*, 2025a.

Yidan Ding, Chalisa Udompanyawit, Yisha Zhang, and Bin He. Eeg-based brain-computer interface enables real-time robotic hand control at individual finger level. *Nature Communications*, 16(1): 5401, 2025b.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. Luna: Efficient and topology-agnostic foundation model for eeg signal analysis. In *Advances in Neural Information Processing Systems*, 2025.

Zijian Dong, Li Ruilin, Joanna Su Xian Chong, Niousha Dehestani, Yinghui Teng, Yi Lin, Zhizhou Li, Yichi Zhang, Yapei Xie, Leon Qi Rong Ooi, et al. Brain harmony: A multimodal foundation model unifying morphology and function into 1d tokens. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Danny Dongyeop Han, Ahhyun Lucy Lee, Taeyang Lee, Yonghyeon Gwon, Sebin Lee, Seongjin Lee, David Keetae Park, Shinjae Yoo, Jiook Cha, and Chun Kee Chung. Diver-0: A fully channel equivariant eeg foundation model. *arXiv e-prints*, pp. arXiv–2507, 2025.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg-adapting to any setup with large-scale pretraining on 25,000 subjects. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zitao Fang, CHENXUAN LI, Zhou Hongting, Shuyang Yu, Guodong DU, Ashwaq Qasem, Yang Lu, Jing Li, Junsong Zhang, and Sim Kuan Goh. Neuript: Foundation model for neural interfaces. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Sam Gijsen and Kerstin Ritter. Eeg-language pretraining for highly label-efficient clinical phenotyping. In *Forty-second International Conference on Machine Learning*, 2025.

Wei Gu, Tianming Luo, Qiran Zhang, Mohan Ye, Xiao Shen, Wenxin Chen, Yunhuan Li, Yichen Zhang, Jing Hong, Bao-liang Lu, and Wei-Long Zheng. Cerebragloss: Instruction-tuning a large vision-language model for fine-grained clinical eeg interpretation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Haokun Gui, Xiucheng Li, and Xinyang Chen. Vector quantization pretraining for eeg time series with random projection and phase alignment. In *ICML*, 2024a.

Yiyu Gui, MingZhi Chen, Yuqi Su, Guibo Luo, and Yuchao Yang. Eegmamba: Bidirectional state space model with mixture of experts for eeg multi-task classification. *arXiv preprint arXiv:2407.20254*, 2024b.

Hanning Guo, Farah Abdellatif, Hanwen Bi, Andrei Galbenus, Jon Shah, Abigail Morrison, Jürgen Dammers, et al. Brain-of: An omnifunctional foundation model for fmri, eeg and meg. *arXiv preprint arXiv:2602.23410*, 2026.

Lu Han, Yu Liu, Lan Li, Qiwen Deng, Jian Jiang, Yinbo Sun, Zhe Yu, Binfeng Wang, Xingyu Lu, Lintao Ma, Han-Jia Ye, and De-Chuan Zhan. Unica: Unified covariate adaptation for time series foundation model. In *International Conference on Learning Representations*, 2026. OpenReview preprint.

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

Yuzhi Huang, Kairun Wen, Rongxin Gao, Dongxuan Liu, Yibin Lou, Jie Wu, Jing Xu, Jian Zhang, Zheng Yang, Yunlong Lin, Chenxin Li, Panwang Pan, Junbin Lu, Jingyan Jiang, Xinghao Ding, Yue Huang, and Zhi Wang. Thinking in dynamics: How multimodal large language models perceive, track, and reason dynamics in physical 4d world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33446–33456, June 2026.

Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In *Ijcai*, volume 2021, pp. 1324–1330, 2020.

Muyun Jiang, Shuailei Zhang, Zhenjie Yang, Mengjun Wu, Weibang Jiang, Zhiwei Guo, Wei Zhang, Rui Liu, Shangen Zhang, Yong Li, et al. Leaf: Language-eeg aligned foundation model for brain-computer interfaces. *arXiv preprint arXiv:2509.24302*, 2025a.

Wei-Bang Jiang, Xi Fu, Yi Ding, and Cuntai Guan. Towards robust multimodal physiological foundation models: Handling arbitrary missing modalities. *arXiv preprint arXiv:2504.19596*, 2025b.

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.

Weibang 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 *The Thirteenth International Conference on Learning Representations*, 2025c.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

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.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*, 2024.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

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

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL `https://aclanthology.org/W04-1013/`.

Aoyang Liu, Haodong Jing, Yulong Liu, Yongqiang Ma, and Nanning Zheng. Hidden states in llms improve eeg representation learning and visual decoding. In *ECAI*, volume 392, pp. 2130–7, 2024a.

Chenyu Liu, Xinliang Zhou, Jiaping Xiao, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. Vsgt: Variational spatial and gaussian temporal graph models for eeg-based emotion recognition. In *IJCAI*, pp. 3078–3086, 2024b.

Chenyu Liu, Xinliang Zhou, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. Vbh-gnn: Variational bayesian heterogeneous graph neural networks for cross-subject emotion recognition. In *The Twelfth International Conference on Learning Representations*, 2024c.

Chenyu Liu, Yuqiu Deng, Tianyu Liu, Jinan Zhou, Xinliang Zhou, Ziyu Jia, and Yi Ding. Echo: Toward contextual seq2seq paradigms in large eeg models. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. Eeg foundation models: Progresses, benchmarking, and open problems. *arXiv preprint arXiv:2601.17883*, 2026b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. On the impact of fine-tuning on chain-of-thought reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11679–11698, 2025.

Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1–R13, 2007.

Weiheng Lu, Chunfeng Song, Jiamin Wu, Pengyu Zhu, Yuchen Zhou, Weijian Mai, Qihao Zheng, and Wanli Ouyang. Unimind: Unleashing the power of llms for unified multi-task brain decoding. *arXiv preprint arXiv:2506.18962*, 2025.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Jingying Ma, Qika Lin, Ziyu Jia, and Mengling Feng. St-usleepnet: a spatial-temporal coupling prominence network for multi-channel sleep staging. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4182–4190, 2025.

Jingying Ma, Qika Lin, Feng Wu, Yucheng Xing, Ziyu Jia, and Mengling Feng. Logicsleep: a neurosymbolic-guided framework for explainable and personalized sleep staging. *npj Artificial Intelligence*, 2026a.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Bridging decoupled tokenizer and multi-scale architecture for eeg foundation model. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka, and Shounak Roychowdhury. Thought2text: text generation from eeg signal using large language models (llms). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3747–3759, 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. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5544–5555, 2024.

Nikhil Shivakumar Nayak, Krishnateja Killamsetty, Ligong Han, Abhishek Bhandwaldar, Prateek Chanda, Kai Xu, Hao Wang, Aldo Pareja, Oleg Silkin, Mustafa Eyceoz, and Akash Srivastava. Sculpting subspaces: Constrained full fine-tuning in llms for continual learning. *arXiv preprint arXiv:2504.07097*, 2025. doi: 10.48550/arXiv.2504.07097.

Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. U-sleep: resilient high-frequency sleep staging. *NPJ digital medicine*, 4(1):72, 2021.

Bijan Pesaran, Martin Vinck, Gaute T Einevoll, Anton Sirota, Pascal Fries, Markus Siegel, Wilson Truccolo, Charles E Schroeder, and Ramesh Srinivasan. Investigating large-scale brain dynamics using field potential recordings: analysis and interpretation. *Nature neuroscience*, 21(7):903–919, 2018.

Jathurshan Pradeepkumar, Zheng Chen, and Jimeng Sun. Neural signals generate clinical notes in the wild. *arXiv preprint arXiv:2601.22197*, 2026a. URL https://arxiv.org/abs/2601.22197.

Jathurshan Pradeepkumar, Xihao Piao, Zheng Chen, and Jimeng Sun. Tokenizing single-channel eeg with time-frequency motif learning. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Mahboobeh Jafari, Parisa Moridian, Roohallah Alizadehsani, Maryam Panahiazar, Fahime Khozeimeh, Assef Zare, Hossein Hosseini-Nejad, et al. Epileptic seizures detection using deep learning techniques: a review. *International journal of environmental research and public health*, 18(11):5780, 2021.

Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7):2074, 2018.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Roland D Thijs, Rainer Surges, Terence J O'Brien, and Josemir W Sander. Epilepsy in adults. *The lancet*, 393(10172):689–701, 2019.

Kostas M Tsiouris, Vasileios C Pezoulas, Michalis Zervakis, Spiros Konitsiotis, Dimitrios D Koutsouris, and Dimitrios I Fotiadis. A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals. *Computers in biology and medicine*, 99:24–37, 2018.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*, 2023.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024a.

Guoan Wang, Shihao Yang, Jun-en Ding, Hao Zhu, and Feng Liu. Neuronarrator: A generalist eeg-to-text foundation model for clinical interpretation via spectro-spatial grounding and temporal state-space reasoning. *arXiv preprint arXiv:2603.16880*, 2026a.

Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. Enhancing eeg-to-text decoding through transferable representations from pre-trained contrastive eeg-text masked autoencoder. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7278–7292, 2024b.

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.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. Eegmamba: An eeg foundation model with mamba. *Neural Networks*, pp. 107816, 2025b.

Jiquan Wang, Sha Zhao, Yangxuan Zhou, Yiming Kang, Shijian Li, and Gang Pan. Deeperbrain: A neuro-grounded eeg foundation model towards universal bci. *arXiv preprint arXiv:2601.06134*, 2026b.

Yuxiang Wei, Yanteng Zhang, Xi Xiao, Chengxuan Qian, Tianyang Wang, and Vince D. Calhoun. fmri-lm: Towards a universal foundation model for language-aligned fmri understanding. *arXiv preprint arXiv:2511.21760*, 2025. doi: 10.48550/arXiv.2511.21760.

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.

Haoyu Xu, Pengxiang Lan, Enneng Yang, Guibing Guo, Jianzhe Zhao, Linying Jiang, and Xingwei Wang. Knowledge decoupling via orthogonal projection for lifelong editing of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13194–13213, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b. doi: 10.48550/arXiv.2509.17765.

Lichao Xu, Minpeng Xu, Yufeng Ke, Xingwei An, Shuang Liu, and Dong Ming. Cross-dataset variability problem in eeg decoding with deep learning. *Frontiers in human neuroscience*, 14:103, 2020.

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

Shihao Yang, Xiying Huang, Danilo Bernardo, Jun-En Ding, Andrew Michael, Jingmei Yang, Patrick Kwan, Ashish Raj, and Feng Liu. Foundation and large-scale ai models in neuroscience: A comprehensive review. *arXiv preprint arXiv:2510.16658*, 2025.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36:26304–26321, 2023.

Daoze Zhang, Zhizhang Yuan, Junru Chen, Kerui Chen, and Yang Yang. Brant-x: A unified physiological signal alignment framework. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4155–4166, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Animashree Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124, 2023.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, 16(12):9851–9915, 2025a.

Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4489–4504, 2025b.

Jianheng Zhou, Chenyu Liu, Jinan Zhou, Yi Ding, Yang Liu, Haoran Luo, Ziyu Jia, and Xinliang Zhou. Introducing multimodal paradigm for learning sleep staging psg via general-purpose model. *arXiv preprint arXiv:2509.22810*, 2025c.

Xinliang Zhou, Chenyu Liu, Zhisheng Chen, Kun Wang, Yi Ding, Ziyu Jia, and Qingsong Wen. Brain foundation models: A survey on advancements in neural signal processing and brain discovery. *IEEE Signal Processing Magazine*, 42(5):22–35, 2025d.

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. In *Advances in Neural Information Processing Systems*, 2025e.

A Detailed Related Work

A.1 Scalp EEG Foundation Models

Electroencephalography (EEG) modeling has undergone a substantial paradigm shift, evolving from task-specific supervised pipelines toward general-purpose representation learning, and more recently toward integration with large language models (LLMs) (Yang et al., 2025; Zhou et al., 2025d; Liu et al., 2026b).

Early EEG analysis was dominated by task-specific supervised methods, where each downstream application, such as sleep staging (Jia et al., 2020; Perslev et al., 2021), seizure detection (Tsiouris et al., 2018; Shoeibi et al., 2021), or emotion recognition (Zheng & Lu, 2015; Liu et al., 2024b), required independently designed models. These methods initially relied on handcrafted features combined with classical classifiers (Lotte et al., 2007). With the rise of deep learning, End-to-end architectures have emerged, including convolutional neural networks (CNNs) (Lawhern et al., 2018), recurrent neural networks (RNNs) (Alhagry et al., 2017), and Transformer models (Song et al., 2022; Ding et al., 2024). While these approaches improved performance, they remained task-specific since models trained for one task rarely generalized across tasks, subjects, acquisition protocols, or electrode configurations (Xu et al., 2020). Their reliance on costly EEG annotations further limited scalability.

To address these limitations, the field has moved toward self-supervised representation learning. These methods train EEG encoders on unlabeled recordings through pretext objectives such as contrastive learning (Kostas et al., 2021), masked reconstruction (Chien et al., 2022), or masked segment prediction (Mohammadi Foumani et al., 2024), thereby learning reusable neural representations before downstream adaptation.

Building on these developments, recent work has introduced EEG foundation models (EFMs), aiming to learn unified and transferable representations from diverse EEG datasets. At the architectural level, existing EFMs have explored Transformer-based backbones (Yang et al., 2023; Wang et al., 2024a; 2025a; Zhou et al., 2025e; Chen et al., 2025), state-space architectures (Wang et al., 2025b; Gui et al., 2024b), and modular routing mechanisms such as mixture-of-experts (Fang et al., 2025; Chen et al., 2026b). In parallel, several works investigate how EEG should be tokenized before pretraining (Jiang et al., 2024; Gui et al., 2024a; Ma et al., 2026b; Pradeepkumar et al., 2026b; Cui et al., 2026). Beyond architecture, the field has also moved toward scaling pretraining to diverse EEG corpora (El Ouahidi et al., 2025), accommodating heterogeneous electrode layouts (Döner et al., 2025; Dongyeop Han et al., 2025), and incorporating neuroscience-inspired priors (Ding et al., 2025a; Wang et al., 2026b).

The emergence of large language models (LLMs) further introduces a fundamentally different capability, which is semantic reasoning over sequences (Babu et al., 2025). Recent studies have begun to explore LLMs for EEG-related problems, including EEG-to-language decoding (Mishra et al., 2025; Wang et al., 2024b; Gijsen & Ritter, 2025), cross-modal generation from neural activity (Liu et al., 2024a; Deng et al., 2025), and vision-language-guided EEG interpretation (Zhou et al., 2025c; Chen et al., 2026a). More recently, this direction has extended to coupling EEG foundation models with LLMs. For example, neural tokenization for multi-task decoding (Jiang et al., 2025c), alignment between EEG representations and language embeddings (Lu et al., 2025; Jiang et al., 2025a), and decoder-centric formulations (Cui et al., 2024; Liu et al., 2026a). These decoding-oriented frameworks primarily target predefined task labels. In parallel, clinically oriented models move beyond label prediction: CELM (Pradeepkumar et al., 2026a) and NeuroNarrator (Wang et al., 2026a) generate clinical reports and narratives from EEG signals, while CerebraGloss (Gu et al., 2026) supports fine-grained interpretation of EEG visualizations. Although these advances broaden the role of language in EEG analysis, they remain primarily focused on clinical interpretation.

A.2 Multimodal Brain Foundation Models

Beyond scalp EEG foundation models, recent work has begun to explore broader brain foundation models. Some extend foundation modeling to invasive electrophysiological recordings, which often provide higher signal fidelity but are harder to acquire (Wang et al., 2023; Zhang et al., 2023), while others integrate richer but more expensive neuroimaging modalities, such as EEG, MEG, fMRI, and structural MRI (Xiao et al., 2025; Guo et al., 2026; Jiang et al., 2025b; Dong et al., 2025; Zhang et al., 2024). Beyond modality integration, fMRI-LM (Wei et al., 2025) explores

language-aligned tokenization and structured neural descriptors for fMRI, highlighting efforts toward semantic grounding in brain foundation models. However, these additional modalities might require invasive acquisition or specialized imaging systems. Complementary to this line of work, we focus on enriching scalp EEG foundation models with contextual information that is naturally paired with EEG recordings to improve semantic grounding.

A.3 Language Model Adaptation

Language drift emerges during non-linguistic fine-tuning. Adapting a pretrained LLM on domain-specific, non-linguistic token sequences introduces two compounding risks. The first is an *input-distribution gap*: discrete tokens produced by a domain tokenizer differ statistically from natural-language tokens, pushing activations into directions that were never reinforced during pretraining. The second is a *gradient-direction conflict*: the task objective drives weight updates toward regions useful for the target domain but destructive to language representations—a form of catastrophic forgetting documented across continual fine-tuning settings (Kotha et al., 2024; Luo et al., 2025; Zheng et al., 2025) that impairs not only surface fluency but also instruction-following and reasoning (Lobo et al., 2025). Contrastive alignment objectives (Wei et al., 2025) narrow but do not eliminate this drift, and recent EEG–language systems (Jiang et al., 2025a; Lu et al., 2025) face the same conflict during downstream adaptation regardless of tokenizer alignment quality.

Parameter-efficient fine-tuning can avoid this problem but suffers from inherent limitations. Low-rank adaptation (LoRA) (Hu et al., 2022) limits trainable parameter count by freezing all backbone weights and routing adaptation through a low-rank perturbation $\Delta W = BA$. The frozen base weights remain unchanged, but the active adapter can still alter language behavior; freezing alone does not eliminate forgetting. The same bottleneck arises in vision-language models: Zhou et al. (Zhou et al., 2025b) show that parameter-freezing limits a VLM's ability to absorb genuinely new visual concepts and propose projection-based fusion as a partial remedy. Gradient-projected schemes such as GaLore (Zhao et al., 2024) improve backbone-update efficiency by confining gradient steps to a low-rank subspace, but do not specifically protect language-sensitive directions.

A principled middle ground is to restrict adaptation to weight subspaces orthogonal to the directions that encode representations one wishes to preserve (Nayak et al., 2025; Xu et al., 2025a). Xu et al. (Xu et al., 2025a) formalize this for lifelong LLM editing: orthogonal projection in parameter space decouples sequentially acquired knowledge and prevents inter-task interference. Nayak et al. (Nayak et al., 2025) extend orthogonal subspace carving to continual full fine-tuning, showing that backbone weights can be updated without disturbing a statically defined reference subspace. A key open question is whether the protected direction should be fixed at initialization or updated dynamically as the model evolves, a distinction that matters when the language manifold itself shifts during training.

Task-accuracy metrics alone cannot reveal language erosion: a model that memorizes domain-specific narration templates can score high on in-domain generation measures regardless of whether its general language competence has degraded (Luo et al., 2025). Measuring language drift control against a held-out general-text stream, such as OpenWebText, provides a complementary probe that is orthogonal to in-domain accuracy and captures the forgetting that task metrics would otherwise mask (Zhang et al., 2020).

B Text-Aligned EEG Tokenizer

This appendix details the EEG tokenizer used in Section 3. It converts multichannel waveforms into discrete tokens and aligns their representations with descriptor text while retaining information needed to reconstruct the signal.

B.1 Problem Statement and Notation

Let $\mathbf{X}^{(0)} \in \mathbb{R}^{C\times T}$ denote a primary EEG segment with C channels and T time samples, where all signals have been preprocessed to a uniform sampling rate of 200 Hz and band-pass filtered to $[0.5, 45]$ Hz. The tokenizer is realized by a text-aligned vector-quantized autoencoder

$$\mathcal{Q}\colon \mathbf{X}^{(0)} \;\longmapsto\; \mathbf{z}_{1:T_z} \in \mathcal{V}^{T_z}, \tag{13}$$

where $\mathcal{V}$ is a learned discrete codebook of size K (we use $K = 8192$). The tokenizer is composed of an encoder $\mathcal{E}_\theta$, a quantizer $\mathcal{Q}_\xi$, and a decoder $\mathcal{D}_\phi$, jointly trained by reconstruction, vector-quantization, contrastive alignment, and domain-confusion objectives.

B.2 Encoder, Patchification, and Codebook

We use a Transformer encoder $\mathcal{E}_\theta$ over channel-wise temporal patches. The input segment is patchified along the temporal dimension by a patch size P (we use $P = 200$, i.e., one second per patch), while keeping every channel as a separate spatial token. This yields a token grid of size

$$M \;=\; \big\lfloor \tfrac{T}{P} \big\rfloor \times C \;=\; T' \times C, \tag{14}$$

which is preserved through all encoder layers so that channel identity remains addressable. The encoder produces a continuous latent

$$\mathbf{h} \;=\; \mathcal{E}_\theta(\mathbf{X}^{(0)}) \in \mathbb{R}^{M \times D}, \tag{15}$$

where D is the encoder embedding dimension. A lightweight projection head $\pi_{\text{enc}} : \mathbb{R}^D \to \mathbb{R}^{d_q}$ maps each token to the codebook dimension d_q ($d_q = 128$).

The discretization step uses an EMA-updated normalized vector quantizer: every projected token $\mathbf{u}_m = \pi_{\text{enc}}(\mathbf{h}_m)$ is matched to the nearest codeword $\mathbf{e}_{k(m)}$ on the unit sphere,

$$k(m) \;=\; \arg\min_{k \in \{1,\dots,K\}} \left\| \frac{\mathbf{u}_m}{\|\mathbf{u}_m\|_2} - \frac{\mathbf{e}_k}{\|\mathbf{e}_k\|_2} \right\|_2^2, \qquad \hat{\mathbf{u}}_m \;=\; \mathbf{e}_{k(m)}. \tag{16}$$

Codebook entries are updated by an exponential moving average of the assigned features (decay 0.99) and re-initialized via k-means on the first batch to avoid codebook collapse.

B.3 Reconstruction and Quantization Objectives

The decoder $\mathcal{D}_\phi$ shares the same Transformer template as $\mathcal{E}_\theta$ and reconstructs the raw segment $\hat{\mathbf{X}}^{(0)} = \mathcal{D}_\phi(\hat{\mathbf{u}}_{1:M})$ from the quantized representation. The autoencoding loss combines reconstruction error with the standard VQ-VAE commitment term

$$\mathcal{L}_{\text{rec}} \;=\; \mathbb{E}\Big[\big\| \mathbf{X}^{(0)} - \hat{\mathbf{X}}^{(0)} \big\|_{1,\text{Huber}} \Big], \tag{17}$$

$$\mathcal{L}_{\text{VQ}} \;=\; \mathbb{E}\Big[\big\| \text{sg}[\mathbf{u}_m] - \mathbf{e}_{k(m)} \big\|_2^2 \;+\; \beta \big\| \mathbf{u}_m - \text{sg}[\mathbf{e}_{k(m)}] \big\|_2^2 \Big], \tag{18}$$

where $\text{sg}[\cdot]$ is the stop-gradient operator, $\beta = 1.0$ is the commitment weight, and the reconstruction loss is the smoothed-ℓ_1 (Huber) variant of the mean absolute error to limit the influence of outlier amplitudes. Gradients flow through the quantizer with the standard straight-through estimator $\partial \hat{\mathbf{u}}_m / \partial \mathbf{u}_m = \mathbf{I}$.

B.4 Contrastive Alignment with Descriptor Text

For tokenizer pretraining, we construct offline EEG–text pairs using the descriptorizer $\mathcal{T}$ introduced in Section 3.2: the structured metric state $\mathbf{m}$ is mapped to a descriptor sentence $\mathbf{d}$ through a deterministic template, which is then refined into fluent text by a strong off-the-shelf LLM. The descriptor sentence is encoded by a frozen text encoder into a pooled text embedding $\mathbf{t} \in \mathbb{R}^{d_t}$. We pool the EEG side either by mean pooling, a `[CLS]` token, or an attention-pooling head h_{MAP}; in the final configuration we use mean pooling for stability:

$$\mathbf{f} \;=\; \frac{1}{M} \sum_{m=1}^{M} \mathbf{h}_m, \qquad \mathbf{f}^{\text{eeg}} \;=\; \pi_{\text{eeg}}(\mathbf{f}), \qquad \mathbf{f}^{\text{txt}} \;=\; \pi_{\text{txt}}(\mathbf{t}), \tag{19}$$

where π_{eeg} and π_{txt} are linear projections to a common d_p-dimensional space ($d_p = 512$).

We employ a SigLIP-style sigmoid contrastive loss to align the two modalities. With L_2-normalized embeddings $\bar{\mathbf{f}}^{\text{eeg}}$ and $\bar{\mathbf{f}}^{\text{txt}}$, the per-batch logit matrix is

$$\mathbf{S}_{ij} \;=\; \sigma \left\langle \bar{\mathbf{f}}_i^{\text{eeg}}, \bar{\mathbf{f}}_j^{\text{txt}} \right\rangle, \qquad \sigma \;=\; \exp(\tau) \le 100, \tag{20}$$

where τ is a learnable log-temperature initialized to $\log(1/0.07)$. Letting $\mathbf{Y} = \mathbf{I}_B$ on the diagonal and $\mathbf{0}$ elsewhere, the symmetric SigLIP loss is

$$\mathcal{L}_{\text{contr}} \;=\; \frac{1}{2}\Big[\mathrm{BCE}(\mathbf{S}, \mathbf{Y}; w^+) + \mathrm{BCE}(\mathbf{S}^\top, \mathbf{Y}; w^+)\Big], \qquad w^+ = B - 1, \tag{21}$$

with $\mathrm{BCE}(\cdot, \cdot; w^+)$ the positive-weighted binary cross-entropy that balances 1 positive vs. $B-1$ negatives per row/column. We optionally substitute Eq. 21 with the symmetric InfoNCE loss

$$\mathcal{L}_{\text{CLIP}} \;=\; \tfrac{1}{2}\big[\mathrm{CE}(\mathbf{S}, \mathbf{y}) + \mathrm{CE}(\mathbf{S}^\top, \mathbf{y})\big], \qquad \mathbf{y} = (1, \ldots, B), \tag{22}$$

or a soft-target distillation variant in which intra-text similarity $\mathbf{S}_{\text{tea}} = \bar{\mathbf{f}}^{\text{txt}}(\bar{\mathbf{f}}^{\text{txt}})^\top/\tau$ provides a teacher distribution $\mathbf{P} = \mathrm{softmax}(\mathbf{S}_{\text{tea}})$ used as soft labels in a symmetric KL divergence; this regularizes alignment when batches contain semantically near-duplicate descriptors.

B.5 Domain-Adversarial Confusion

Because EEG and text inhabit different statistical manifolds, naive contrastive learning can leave a residual gap that hurts downstream language modeling. We additionally pass random text token sequences through the same projection head and require a small classifier C_ψ to fail to distinguish EEG tokens from text tokens. With a gradient-reversal layer $\mathcal{R}_\alpha(\cdot)$ between the encoder and the classifier and binary indicator $t_m \in \{0, 1\}$ (1 for EEG, 0 for text), the domain-confusion loss is

$$\mathcal{L}_{\text{dom}} \;=\; -\frac{1}{M}\sum_{m=1}^{M}\Big[t_m \log C_\psi\big(\mathcal{R}_\alpha(\mathbf{h}_m)\big) \;+\; (1 - t_m)\log\big(1 - C_\psi(\mathcal{R}_\alpha(\mathbf{h}_m))\big)\Big], \tag{23}$$

where α is annealed from 0 to 1 via the schedule $\alpha = 2/(1 + \exp(-10\,p)) - 1$ in training progress $p \in [0, 1]$.

B.6 Total Tokenizer Objective

The tokenizer is trained end-to-end with

$$\mathcal{L}_{\text{tok}} \;=\; \mathcal{L}_{\text{rec}} \;+\; \mathcal{L}_{\text{VQ}} \;+\; \lambda_{\text{c}}\,\mathcal{L}_{\text{contr}} \;+\; \lambda_{\text{d}}\,\mathcal{L}_{\text{dom}}, \tag{24}$$

with $\lambda_{\text{c}} = 1.0$ and $\lambda_{\text{d}} = 0.5$. Optimization uses fused AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate 1×10^{-4} and cosine annealing with warm restarts, weight decay 1×10^{-4}, and global gradient clipping disabled. Codebook re-initialization (k-means on the first batch) is performed once at the beginning of training. After convergence, the encoder $\mathcal{E}_\theta$, projection π_{enc}, and quantizer codebook $\{\mathbf{e}_k\}_{k=1}^K$ are frozen and exposed to the downstream language model as a fixed mapping $\mathcal{Q} : \mathbf{X}^{(0)} \to \mathbf{z}_{1:T_z}$ Training converges stably, as shown in Figures 7 and 8.

Why a language-aligned tokenizer. The combined objective in Eq. 24 jointly enforces three properties: (i) information fidelity, so that the EEG signal can be reconstructed from its tokens; (ii) discrete addressability, so that downstream language modeling operates over a finite, well-conditioned vocabulary; and (iii) cross-modal compatibility, so that EEG tokens occupy a semantic geometry close to that of the frozen language model. Property (iii) motivates combining tokenizer alignment with the projected adaptation in Section 3; neither alignment nor the local gradient constraint guarantees unchanged linguistic competence.

C Evaluation Metrics

This appendix lists every metric used in the main results (Table 1), the ablation study (Table 2), the trade-off figures (Figures 4–5), and the additional results reported in the supplementary tables. We follow the standard EEG-classification reporting convention used in recent EEG foundation models (Jiang et al., 2025c; Liu et al., 2026a) for the primary classification metrics and for the BERTScore-based language metric.

Notation. Let $y_n \in \{1, \ldots, K\}$ be the ground-truth label and $\hat{y}_n$ the predicted label of sample $n = 1, \ldots, N$. We denote by TP_k, TN_k, FP_k, FN_k the per-class confusion-matrix counts, by $\hat{p}_n \in [0, 1]^K$ the predicted class-probability vector, and by $w_k = \sum_n \mathbf{1}\{y_n = k\}/N$ the empirical class frequency.

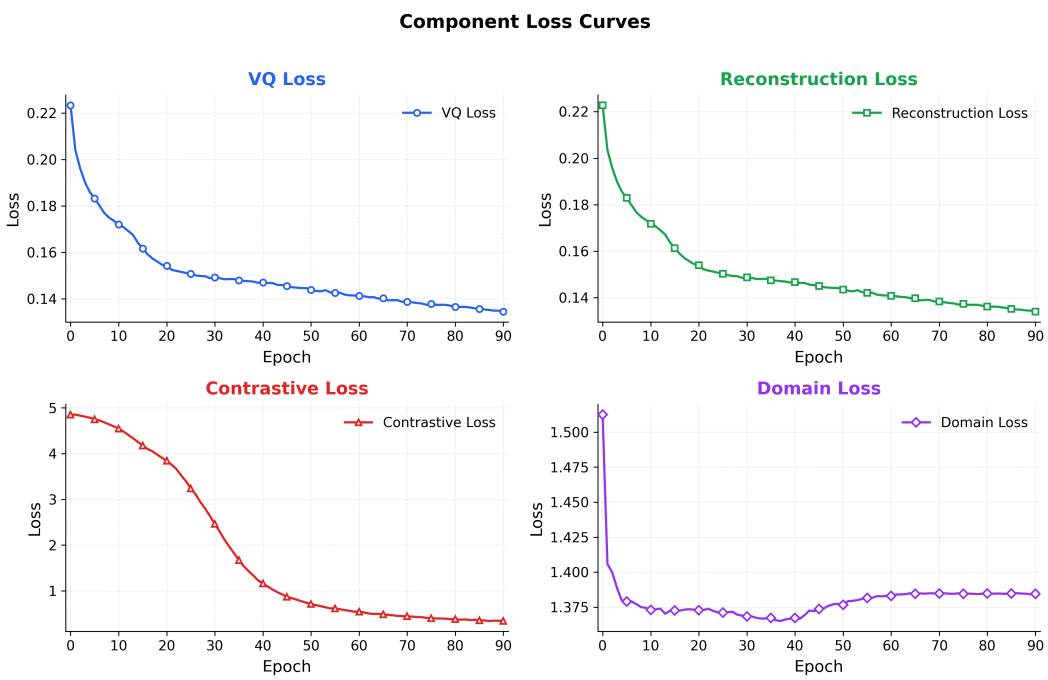

Figure 7: Component training loss curves over 90 epochs. VQ quantization loss (top) and contrastive alignment loss (bottom) both converge steadily during encoder pretraining.

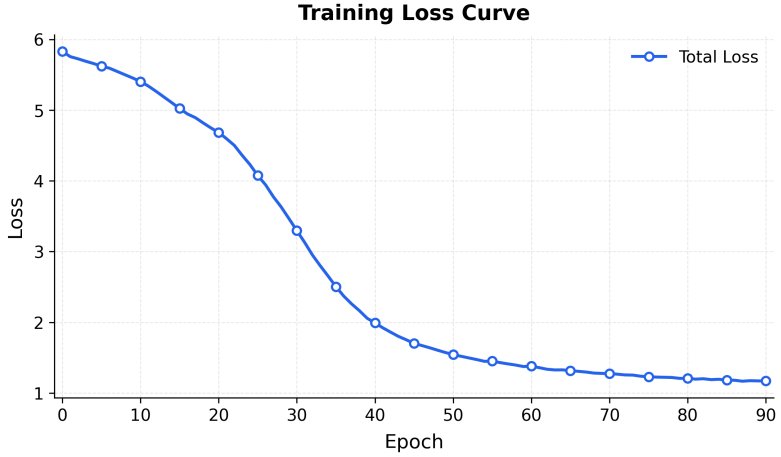

Figure 8: Total combined training loss over 90 epochs, aggregating reconstruction, VQ quantization, and contrastive alignment objectives. The loss decreases smoothly, confirming stable encoder pretraining.

C.1 Classification Metrics

Balanced Accuracy (ACC-B). The arithmetic mean of per-class recall, which corrects accuracy under class imbalance,

$$\text{ACC-B} = \frac{1}{K}\sum_{k=1}^{K}\frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}. \tag{25}$$

A random predictor scores $1/K$.

ROC AUC. For binary tasks we use the standard area under the receiver operating characteristic curve. With threshold η, define

$$\mathrm{TPR}(\eta) = \frac{\mathrm{TP}(\eta)}{\mathrm{TP}(\eta) + \mathrm{FN}(\eta)}, \qquad \mathrm{FPR}(\eta) = \frac{\mathrm{FP}(\eta)}{\mathrm{FP}(\eta) + \mathrm{TN}(\eta)}, \tag{26}$$

and integrate one over the other,

$$\mathrm{ROC\,AUC} \;=\; \int_0^1 \mathrm{TPR}\big(\mathrm{FPR}^{-1}(u)\big)\, du. \tag{27}$$

For multiclass tasks we report the one-vs-rest macro average $(1/K)\sum_k \mathrm{ROC\,AUC}_k$.

PR AUC. The area under the precision–recall curve, where

$$\mathrm{Precision}(\eta) = \frac{\mathrm{TP}(\eta)}{\mathrm{TP}(\eta) + \mathrm{FP}(\eta)}, \qquad \mathrm{Recall}(\eta) = \frac{\mathrm{TP}(\eta)}{\mathrm{TP}(\eta) + \mathrm{FN}(\eta)}, \tag{28}$$

$$\mathrm{PR\,AUC} = \int_0^1 \mathrm{Precision}\big(\mathrm{Recall}^{-1}(u)\big)\, du. \tag{29}$$

PR AUC is preferred over ROC AUC under heavy class imbalance because the negative class no longer dominates the false-positive denominator.

Cohen's Kappa. Inter-rater agreement corrected for chance,

$$\kappa \;=\; \frac{p_o - p_e}{1 - p_e}, \qquad p_o = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{\hat{y}_n = y_n\}, \qquad p_e = \sum_{k=1}^{K} \hat{w}_k \cdot w_k, \tag{30}$$

where $\hat{w}_k$ is the predicted class frequency. We report Cohen's Kappa on the multiclass benchmarks (BCIC IV 2a, SEED-V, TUEV, BCIC 2020–T3).

Weighted F1. The class-frequency-weighted average of per-class F1 scores,

$$\mathrm{F1}_k = \frac{2\,\mathrm{Precision}_k \cdot \mathrm{Recall}_k}{\mathrm{Precision}_k + \mathrm{Recall}_k}, \qquad \mathrm{Weighted\,F1} \;=\; \sum_{k=1}^{K} w_k \cdot \mathrm{F1}_k. \tag{31}$$

This metric is reported on the multiclass benchmarks alongside Cohen's Kappa.

C.2 Generation Metrics

Field accuracy (F-Acc) and Total F-Acc. For every test window the model is queried independently on the four field-aligned sub-tasks $\mathcal{T}_{\mathrm{gen}} = \{\mathrm{label}, \mathrm{relation}, \mathrm{band}, \mathrm{summary}\}$. The model's generated answer is parsed by the field-level grammar in Appendix G into a structured prediction, which is compared against the ground-truth field by exact match $\mathbf{1}\{\hat{f}_n = f_n\}$. The per sub-task field accuracy is

$$\text{F-Acc}_\tau \;=\; \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{\hat{f}_n^{\tau} = f_n^{\tau}\}, \qquad \tau \in \mathcal{T}_{\mathrm{gen}}, \tag{32}$$

and the headline scalar is the unweighted mean

$$\mathrm{Total\,F\text{-}Acc} \;=\; \tfrac{1}{4}\sum_{\tau \in \mathcal{T}_{\mathrm{gen}}} \text{F-Acc}_\tau. \tag{33}$$

Total F-Acc measures categorical field recovery. Direct F1 instead evaluates numeric claims in generated summaries, so the two metrics have different targets and aggregation rules and are reported separately.

Direct F1. We extract explicitly named relative band-power claims from the generated summary and match them to the direct reference by attribute name and numeric value, with absolute tolerance 0.005. Correct claims contribute true positives; incorrect values contribute both a false positive and a false negative; missing reference claims contribute false negatives and additional claims contribute false positives. Counts are pooled over samples within each dataset:

$$\text{Direct F1} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}. \tag{34}$$

Zero-denominator cases receive zero.

BERTScore. To examine language behavior under adaptation in Figure 5, we report BERTScore between Ewen-generated responses $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_M)$ and the corresponding reference $\mathbf{y} = (y_1, \ldots, y_N)$ using BERTScore. Let $\mathbf{e}(y_i), \mathbf{e}(\hat{y}_j) \in \mathbb{R}^d$ be contextual embeddings of the reference and candidate tokens produced by a frozen pretrained encoder. Define the cosine similarity matrix $\mathbf{R}_{ij} = \cos(\mathbf{e}(y_i), \mathbf{e}(\hat{y}_j))$. BERTScore is the F1 of greedy precision and recall,

$$\text{P}_{\text{BS}} = \frac{1}{M}\sum_{j=1}^{M} \max_i \mathbf{R}_{ij}, \quad \text{R}_{\text{BS}} = \frac{1}{N}\sum_{i=1}^{N} \max_j \mathbf{R}_{ij}, \quad \text{BERTScore} = \frac{2\,\text{P}_{\text{BS}} \cdot \text{R}_{\text{BS}}}{\text{P}_{\text{BS}} + \text{R}_{\text{BS}}}. \tag{35}$$

We use the publicly released `roberta-large` backbone and the rescaled F1 variant. Higher BERTScore indicates that the generated description is closer to the reference in the contextual semantic space induced by the pretrained language model; in our setting, BERTScore measures similarity within the evaluated reference distribution.

Summary ROUGE-L (Summary R-L). We use the F1 variant of ROUGE-L (Lin, 2004). Each generated summary and its reference are lowercased and tokenized into alphanumeric words and decimal numbers, without stemming. Let $\hat{s}_n$ and s_n denote the resulting token sequences and $\ell_n = \text{LCS}(\hat{s}_n, s_n)$ their longest common subsequence length. Then

$$P_n = \frac{\ell_n}{|\hat{s}_n|}, \qquad R_n = \frac{\ell_n}{|s_n|}, \qquad \text{Summary R-L} = \frac{1}{N}\sum_{n=1}^{N} \frac{2P_nR_n}{P_n + R_n}. \tag{36}$$

An empty sequence or zero overlap receives zero. The LCS is computed over each complete summary, and the reported score is the mean of sample-level F1 scores.

Withheld F1. Appendix Q.1 specifies the comparison targets, three-state encoding, and per-window availability. Withheld targets are waveform-derived physiological attributes whose target states are not explicitly supplied in the descriptors. The target specification and three-state encoding are fixed across methods and input conditions; a fixed target-availability mask identifies observable targets. Let $\mathcal{A}_d$ be the observable attribute set for dataset d. Using the text parser in Appendix G, we accumulate one-vs-rest counts over samples, compute three-class macro-F1 for each attribute, and average over attributes:

$$\text{Withheld F1}_d = \frac{1}{|\mathcal{A}_d|}\sum_{a \in \mathcal{A}_d} \frac{1}{3}\sum_{c=1}^{3} \frac{2\,\text{TP}_{a,c}}{2\,\text{TP}_{a,c} + \text{FP}_{a,c} + \text{FN}_{a,c}}. \tag{37}$$

Zero-denominator terms receive zero. Unavailable targets are excluded, while missing, invalid, or contradictory predictions remain in the target support and count as false negatives for the true class. This metric evaluates recovery of facts beyond the explicit descriptor content; it does not assume statistical independence between descriptors and withheld targets.

OWT relative perplexity change (OWT ΔPPL). Let θ_0 be the pretrained backbone and θ its adapted checkpoint. On a fixed OpenWebText token stream, let $\mathcal{I}$ contain the scored target-token positions and c_t their preceding context. The token-averaged next-token negative log-likelihood and perplexity are

$$\mathcal{L}_{\text{OWT}}(\theta) = -\frac{1}{|\mathcal{I}|}\sum_{t \in \mathcal{I}} \log p_\theta(x_t \mid c_t), \qquad \text{PPL}_{\text{OWT}}(\theta) = \exp\big(\mathcal{L}_{\text{OWT}}(\theta)\big). \tag{38}$$

Both checkpoints must use the same tokens, tokenizer, context length, and scoring mask, without EEG conditioning. We define

$$\text{OWT } \Delta\text{PPL} = \frac{\text{PPL}_{\text{OWT}}(\theta)}{\text{PPL}_{\text{OWT}}(\theta_0)} - 1 = \exp\big(\mathcal{L}_{\text{OWT}}(\theta) - \mathcal{L}_{\text{OWT}}(\theta_0)\big) - 1. \tag{39}$$

Values are reported as decimal ratios: 0.01 denotes a 1% increase in perplexity. Lower is better, and zero denotes unchanged perplexity. This metric evaluates adaptation-induced changes within the OWT reference domain.

C.3 Reporting Convention

For the classification comparisons in Tables 1 and 2, we run five seeds and report mean $\pm$ standard deviation. Each cell in Table 1 corresponds to a single metric on a single dataset; for Table 2, checkpoint selection uses the highest ACC-B on the held-out validation split. The pairs (ROC AUC / Cohen's Kappa) and (PR AUC / Weighted F1) used in Table 2 reflect the binary-vs-multiclass distinction stated in the table footnote.

For Figure 3 and Table 3, we report means and sample standard deviations across three independent training runs with different seeds. Dataset-macro is first computed within each checkpoint as the unweighted mean of the three dataset-level scores and then summarized across runs. All reported metrics, including OWT ΔPPL, follow this across-run convention; every checkpoint is evaluated on the same OWT token stream, with a common tokenizer, context length, and scoring mask.

C.4 Online Language-Reference Projected Adaptation

This subsection details the update in Section 3.4. The language reference is evaluated at the current decoder and refreshed during EEG adaptation. The analysis concerns local sensitivity to this reference; it does not assume a fixed language-preserving subspace.

Current-model language reference. Write the trainable decoder parameters as $\theta = (\theta_{\text{core}}, \theta_{\text{res}})$, where $\theta_{\text{core}} = \{\mathbf{c}_j : j \in \mathcal{J}\}$ contains the row-direction coefficients in Eq. 46, and $\theta_{\text{res}} = \{A_j, B_j : j \in \mathcal{J}\}$ contains the rank-16 residual factors in Eq. 9. All pretrained base weights remain fixed. For an OpenWebText (OWT) sequence w with scored token positions $\mathcal{I}(w)$, define

$$\mathcal{R}_{\text{OWT}}(\theta) = \mathbb{E}_{w \sim \mathcal{D}_{\text{OWT}}} \left[-\frac{1}{|\mathcal{I}(w)|} \sum_{n \in \mathcal{I}(w)} \log p_\theta(w_n \mid w_{<n}) \right]. \tag{40}$$

The reference uses text without EEG conditioning. At step t, K OWT micro-batches yield core gradients $\mathbf{g}^{(k)}_{\text{lang},t} = \nabla_{\theta_{\text{core}}} \mathcal{L}^{(k)}_{\text{lang}}(\theta_t)$. They are computed with the current core and residual parameters, rather than cached at the pretrained model. Stacking these gradients gives $\mathbf{G}_t = [\mathbf{g}^{(1)}_{\text{lang},t}, \ldots, \mathbf{g}^{(K)}_{\text{lang},t}]$. OWT supplies a reproducible language reference; the resulting constraint concerns that reference domain.

Core attenuation and residual adaptation. Let $\mathbf{g}_t = \nabla_{\theta_{\text{core}}} \mathcal{L}_{\text{task}}(\theta_t)$, using the combined objective in Eq. 10. The gradient passed to the core optimizer is

$$\mathbf{u}_t = \mathbf{P}_t \mathbf{g}_t, \qquad \mathbf{P}_t = \mathbf{I} - \mathbf{G}_t(\mathbf{G}_t^\top \mathbf{G}_t + \alpha \mathbf{I})^{-1} \mathbf{G}_t^\top, \qquad \alpha > 0. \tag{41}$$

The residual factors follow the task objective without this projection. For a nonzero singular value σ_j of $\mathbf{G}_t$, $\mathbf{P}_t$ scales the corresponding reference direction by $\alpha/(\sigma_j^2 + \alpha)$ and leaves its orthogonal complement unchanged. Thus ridge regularization attenuates reference-aligned components rather than removing them exactly. Normalized angular deviation also depends on the remaining gradient components and the selected reference direction. In particular,

$$\mathbf{G}_t^\top \mathbf{u}_t = \alpha(\mathbf{G}_t^\top \mathbf{G}_t + \alpha \mathbf{I})^{-1} \mathbf{G}_t^\top \mathbf{g}_t. \tag{42}$$

As $\alpha \to 0$, this approaches Euclidean projection onto the complement of the sampled gradient span; a pseudoinverse defines the zero-ridge case when the Gram matrix is singular.

First-order interpretation. Consider an isolated core step $\theta_{\mathrm{core},t+1} = \theta_{\mathrm{core},t} - \eta_t \mathbf{u}_t$ with the residual fixed. A first-order expansion of each sampled text loss gives

$$\mathcal{L}^{(k)}_{\mathrm{lang}}(\theta_{t+1}) - \mathcal{L}^{(k)}_{\mathrm{lang}}(\theta_t) = -\eta_t (\mathbf{g}^{(k)}_{\mathrm{lang},t})^\top \mathbf{u}_t + O(\eta_t^2). \tag{43}$$

Equation 42 characterizes the remaining linear term for these sampled losses. Finite sampling does not ensure the same property for every OWT sequence. Moreover, $\mathbf{g}_t^\top \mathbf{P}_t \mathbf{g}_t \geq \|\mathbf{u}_t\|_2^2$: for a γ-smooth task loss, this isolated core step decreases the loss whenever $\mathbf{u}_t \neq 0$ and $0 < \eta_t < 2/\gamma$. These statements describe the supplied gradient and an ideal gradient step.

Finite optimizer steps. Let $\mathbf{s}_t = \theta_{t+1} - \theta_t$ be the actual displacement, including both core and residual updates. If $\mathcal{R}_{\mathrm{OWT}}$ is β-smooth along the step, then

$$\begin{aligned}\mathcal{R}_{\mathrm{OWT}}(\theta_{t+1}) - \mathcal{R}_{\mathrm{OWT}}(\theta_t) \leq{}& \nabla_{\theta_{\mathrm{core}}} \mathcal{R}_{\mathrm{OWT}}(\theta_t)^\top \mathbf{s}_{\mathrm{core},t} \\ &+ \nabla_{\theta_{\mathrm{res}}} \mathcal{R}_{\mathrm{OWT}}(\theta_t)^\top \mathbf{s}_{\mathrm{res},t} + \frac{\beta}{2}\|\mathbf{s}_t\|_2^2. \end{aligned} \tag{44}$$

Momentum, adaptive preconditioning, and weight decay can make $\mathbf{s}_{\mathrm{core},t}$ differ from $-\eta_t \mathbf{u}_t$. The unprojected residual and the finite-step term also remain. Summing Eq. 44 gives a cumulative bound, but does not establish that its terms are small. The update therefore controls local reference sensitivity of the core gradient, without guaranteeing unchanged language behavior after optimization.

Angular diagnostic. For nonzero update and reference vectors $\mathbf{d}$ and $\mathbf{r}$, the absolute deviation from orthogonality is

$$\mathrm{dev}(\mathbf{d}, \mathbf{r}) = \arcsin \frac{|\langle \mathbf{d}, \mathbf{r} \rangle|}{\|\mathbf{d}\|_2 \|\mathbf{r}\|_2}. \tag{45}$$

A zero angle indicates orthogonality to the chosen reference direction. Figure 4 shows per-step angular deviations and their training-wide averages. This diagnostic complements measurements of adaptation-induced changes; an angle alone does not measure retained generation quality or perplexity.

Implementation specification: constants and parameter scope. We use $K = 4$, $\lambda_{\mathrm{cov}} = 0.1$, and a fixed dimensionless ridge coefficient $\rho = 0.5$ throughout training. Let $\mathcal{J}$ contain the following tensors in every text-decoder layer $\ell = 0, \ldots, L-1$ that contains the indicated module. Layer membership is read directly from the backbone's configured `layer_types`; no layer index is selected manually. Paths are relative to the decoder layer and name complete stored matrices, including the gate rows packed in `q_proj`. The names follow the Qwen3.5 Transformers implementation.

Layer component	Weight tensors carrying the structured core and residual adapter
Full attention	`self_attn.{q_proj,k_proj,v_proj,o_proj}.weight`
Linear attention	`linear_attn.{in_proj_qkv,in_proj_z,` `in_proj_a,in_proj_b,out_proj}.weight`
Feed-forward	`mlp.{gate_proj,up_proj,down_proj}.weight`

For a target $W_j^{(0)} \in \mathbb{R}^{m_j \times n_j}$, define the fixed row norms $r_{j,i} = \|W^{(0)}_{j,i:}\|_2$ and $\overline{W}^{(0)}_{j,i:} = W^{(0)}_{j,i:}/r_{j,i}$. For a zero pretrained row, set its normalized direction and core coefficient to zero, omit that coefficient from optimization, and define its forward contribution as zero. The dimensions below retain all rows; the trainable-parameter count assumes nonzero pretrained rows. The structured correction is

$$\Delta W_{\mathrm{core},j}(\mathbf{c}_j) = \mathrm{Diag}(\mathbf{c}_j)\overline{W}_j^{(0)}, \qquad \mathbf{c}_j \in \mathbb{R}^{m_j}. \tag{46}$$

For nonzero pretrained rows, it therefore contains exactly m_j, rather than $m_j n_j$, trainable parameters; no dense correction tensor is allocated. An implementation may evaluate the core without storing $\overline{W}_j^{(0)}$ separately:

$$W_j \mathbf{x} = W_j^{(0)}\mathbf{x} + (\mathbf{c}_j \oslash \mathbf{r}_j) \odot (W_j^{(0)}\mathbf{x}) + B_j(A_j \mathbf{x}), \tag{47}$$

where $\oslash$ and $\odot$ denote element-wise division and multiplication, with $(\mathbf{c}_j \oslash \mathbf{r}_j)_i = 0$ for a zero pretrained row. The nonzero normalized row directions have disjoint matrix-row support and unit

Frobenius norm. Consequently, for coefficient collections $\mathbf{c}$ and $\mathbf{c}'$,

$$\sum_{j\in\mathcal{J}} \left\langle \Delta W_{\mathrm{core},j}(\mathbf{c}_j), \Delta W_{\mathrm{core},j}(\mathbf{c}'_j) \right\rangle_F = \sum_{j\in\mathcal{J}} \mathbf{c}_j^\top \mathbf{c}'_j. \tag{48}$$

Thus the Euclidean geometry used by gradient attenuation corresponds to Frobenius geometry in the represented weight subspace and has no trainable two-factor rescaling ambiguity. For every target, $B_j \in \mathbb{R}^{m_j\times 16}$ and $A_j \in \mathbb{R}^{16\times n_j}$. Initialize $\mathbf{c}_j = \mathbf{0}$, $B_j = \mathbf{0}$, and A_j with Kaiming-uniform initialization. This makes the initial decoder identical to the frozen pretrained decoder while allowing nonzero first-step gradients for the core coefficients and B_j. Core coefficients and residual factors form separate optimizer groups.

All original backbone parameters remain frozen: the selected bases $W_j^{(0)}$, unselected weights, embeddings, language head, normalization parameters, biases, linear-attention convolution kernels, `A_log`, and `dt_bias`. The row norms are fixed buffers rather than parameters. Any vision branch is inactive and frozen. The pretrained EEG tokenizer encoder, projection, and codebook are frozen. The trainable EEG-side group contains the covariate encoder/projector, descriptor–neural alignment heads, EEG-to-decoder interface, feature modulation, and token-geometry parameters; it receives unprojected task gradients. A classification head, when used, is also in this unprojected group.

Loss reduction, gradient aggregation, and solve. Each OWT micro-batch uses mean next-token loss over valid target tokens; padding is excluded. For each of the K micro-batches, sum token losses and token counts across distributed workers before forming its gradient. Accumulated task micro-batches use the same token weighting for $\mathcal{L}_{\mathrm{gen}}$; $\mathcal{L}_{\mathrm{cov}}$ is the mean squared Euclidean alignment error over valid paired windows. Unscale gradients before attenuation when loss scaling is used. Concatenate the coefficient gradients $\nabla_{\mathbf{c}_j}\mathcal{L}$ for all $j \in \mathcal{J}$ in a fixed layer–module–row order, and use exactly the same order for the task gradient and every OWT gradient. Reduce these coefficient gradients across workers before forming $\mathbf{G}_t$, retain all K OWT columns, and do not normalize the columns or replace them by their mean. All gradients use the same pre-update model with both the structured core and residual adapters active. OWT has no EEG conditioning and supplies reference gradients only, not an additional optimizer update.

$$\mathbf{H}_t = \mathbf{G}_t^\top \mathbf{G}_t, \qquad s_t = \frac{\mathrm{tr}(\mathbf{H}_t)}{K}, \qquad \alpha_t = \rho s_t. \tag{49}$$

The coefficient ρ is fixed, while α_t scales with the reference-gradient magnitude. For $s_t > 0$, use a Cholesky solve in at least FP32:

$$(\mathbf{H}_t + \alpha_t \mathbf{I})\mathbf{a}_t = \mathbf{G}_t^\top \mathbf{g}_t, \qquad \mathbf{u}_t = \mathbf{g}_t - \mathbf{G}_t \mathbf{a}_t. \tag{50}$$

No explicit inverse or parameter-sized projector is formed. If $s_t = 0$, set $\mathbf{u}_t = \mathbf{g}_t$. Retry failed factorization in FP64; a persistent failure aborts the step and is logged. Map the attenuated vector back to the coefficient gradients $\{\nabla_{\mathbf{c}_j}\mathcal{L}\}_{j\in\mathcal{J}}$, combine it with the unprojected residual-factor and EEG-side task gradients, apply global norm clipping at 1.0, and take one AdamW step under Appendix N. Refresh the K-column reference at every optimizer step. This filters a local core gradient; it does not strictly preserve language behavior under finite optimizer and residual updates.

D DATASETS

We evaluate Ewen on eight publicly available EEG benchmarks that cover seven downstream directions: abnormal EEG screening, epileptiform event classification, mental-workload recognition, motor-imagery decoding, depression detection, imagined-speech decoding, and cross-subject emotion recognition. Unless otherwise stated, all signals are resampled to a common 200 Hz, band-pass filtered to $[0.1, 75]$ Hz with a notch filter at the local power-line frequency, robustly z-scored per channel, and segmented into nonoverlapping windows whose length matches the original protocol. We use the dataset-specific partitions described below. The TUAB and TUEV partitions follow the NeuroLM pipeline (Jiang et al., 2025c); Mental Arithmetic and SHU–MI use the subject partitions in the CBraMod benchmark (Wang et al., 2025a). SEED-V uses the cross-subject partition of ECHO (Liu et al., 2026a).

TUAB (Clinical Normality Detection). Derived from the Temple University Hospital EEG corpus, TUAB contains 23-channel scalp recordings labeled as *normal* or *abnormal* by board-certified neurologists. We use the official train/test partition shipped with the corpus; 20% of the training set is held out by patient as a validation pool. Each recording is segmented into 10-second windows sampled at 200 Hz.

TUEV (EEG Event Classification). Also part of the TUH corpus, TUEV provides 23-channel recordings annotated with six event categories: spike-and-slow-wave (SPSW), generalized periodic epileptiform discharge (GPED), periodic lateralized epileptiform discharge (PLED), eye movement (EYEM), artifact (ARTF), and background (BCKG). We use the official patient-level split. Windows are 5 s, 200 Hz.

Mental Arithmetic (Cognitive Workload). Recorded with 19 scalp electrodes (international 10–20) at a native rate of 500 Hz, this dataset contains paired *rest* and *mental-arithmetic* segments from 36 subjects. We resample to 200 Hz, band-pass to $[0.5, 45]$ Hz, and segment into 5-second windows. The subject-disjoint split uses 28 subjects for training, 4 for validation, and 4 for testing: original file IDs Subject00–Subject27, Subject28–Subject31, and Subject32–Subject35, respectively (Wang et al., 2025a).

SHU–MI (Motor-Imagery). Released by the Shanghai University BCI group, SHU–MI contains 32-channel scalp EEG (originally 250 Hz, resampled to 200 Hz) with two classes corresponding to imagined left- and right-hand movements. We use a subject-disjoint benchmark split: subjects 1–15 for training, 16–20 for validation, and 21–25 for testing, keeping all five sessions of each subject in the same partition (Wang et al., 2025a).

Mumtaz 2016 (Depression Detection). A 19-channel EEG dataset of 64 subjects (34 with major depressive disorder, 30 healthy controls) recorded during eyes-open and eyes-closed resting conditions. The native sampling rate is 256 Hz; we resample to a unified 200 Hz to match the rest of the pipeline. Windows are 5 s. The split is patient-level, with 11 subjects in the full test set.

BCIC IV 2a (Motor Imagery). Nine subjects performed cued motor imagery of left hand, right hand, foot, and tongue. The dataset is recorded at 250 Hz with 22 EEG channels and is resampled to 200 Hz; windows are 4 s. We adopt the canonical evaluation protocol that uses session 1 for training and session 2 for testing, with 20% of training trials held out per subject for validation.

BCIC 2020–T3 (Imagined-Speech). Released as Track 3 of the 2020 IEEE BCI competition, this dataset contains 64-channel imagined-speech recordings over 5 semantic categories. We resample to 200 Hz and band-pass to $[0.5, 45]$ Hz. The official competition partitions trials from the same 15 subjects: 60 training, 10 validation, and 10 test trials per class per subject, giving 4,500/750/750 trials across the five classes.

SEED-V (Emotion Recognition). The Shanghai Jiao Tong University Emotion EEG Dataset V records 62-channel EEG at a native 1,000 Hz, downsampled here to 200 Hz, from 16 subjects exposed to film clips eliciting happy, sad, neutral, fear, and disgust states. Following the ECHO dataset protocol (Liu et al., 2026a), original subject 7 is excluded due to data corruption, leaving 15 usable subjects. We use its strict *cross-subject* partition with five retained subjects each for training, validation, and testing. This subject-disjoint design directly probes inter-individual generalization, which is the most challenging axis of variation for emotion EEG.

A summary of the eight benchmarks is provided in Table 4.

E BASELINES

We compare Ewen with EEG classification baselines and EEG–language models. Classification comparisons follow the preprocessing and splits in Appendix D; external systems in the comparison are task-aligned on the same narratives. The descriptions below specify each baseline and its use in our setup.

Table 4: Summary of the eight downstream datasets used to evaluate Ewen.

Dataset	Task family	Classes	Subjects	Channels	Modalities	Native rate	Window	Total windows
TUAB	Clinical normality	2	2,329	23	EEG, EOG, ECG, EMG	250 Hz	10 s	409,455
TUEV	Clinical event	6	370	23	EEG, EOG, ECG, EMG	250 Hz	5 s	113,353
Mental Arithmetic	Cognitive workload	2	36	19	EEG, ECG	500 Hz	5 s	1,707
SHU–MI	Motor imagery (binary)	2	25	32	EEG	250 Hz	4 s	11,988
Mumtaz 2016	Depression detection	2	64	19	EEG	256 Hz	5 s	7,083
BCIC IV 2a	Motor imagery (4-class)	4	9	22	EEG, EOG	250 Hz	4 s	5,088
BCIC 2020–T3	Imagined speech	5	15	64	EEG	256 Hz	3 s	6,000
SEED-V	Emotion recognition	5	15	62	EEG	1,000 Hz	1 s	117,744

EEGNet (Lawhern et al., 2018). A compact convolutional neural network specifically designed for EEG-based brain–computer interfaces. EEGNet uses depthwise and separable convolutions to factorize the spatial mixing matrix from temporal feature extraction, yielding a strong inductive bias for low-channel-count signals and stable performance under limited training data. We adopt the canonical EEGNet-8,2 configuration and train it from scratch on each downstream dataset.

EEGPT (Wang et al., 2024a). A pretrained Transformer for universal EEG representation learning. EEGPT combines masked autoencoding with a spatio-temporal representation-alignment objective in a hierarchical encoder that decouples spatial and temporal feature extraction, enabling efficient transfer across heterogeneous BCI tasks.

BIOT (Yang et al., 2023). A biosignal Transformer that tokenizes EEG with a learnable channel-aware patchifier and trains across heterogeneous datasets via cross-data contrastive learning. BIOT is robust to mismatched channel sets and sampling rates and is therefore a strong baseline whenever cross-dataset generalization is a concern.

LaBraM (Jiang et al., 2024). A scalable Transformer foundation model pretrained on a diverse collection of EEG corpora using masked-token reconstruction over a learned discrete codebook. LaBraM is one of the strongest publicly available encoder-only EEG foundation models on TUAB-style benchmarks.

CBraMod (Wang et al., 2025a). A criss-cross brain foundation model that decouples temporal and spatial dependencies via two parallel attention branches. CBraMod is reported as the strongest single-task baseline on several motor-imagery and emotion benchmarks in recent EEG foundation-model literature.

NeuroLM (Jiang et al., 2025c). A multi-task EEG foundation model that bridges language and EEG by learning a text-aligned EEG tokenizer and adapting a frozen GPT-style decoder via cross-modal pretraining and instruction tuning. NeuroLM is the most direct predecessor of Ewen: both treat EEG as a sequence of language-aligned tokens and both rely on a pretrained language backbone for downstream prediction. We use the publicly released NeuroLM checkpoints under the same downstream protocol as our model.

ECHO (Liu et al., 2026a). A decoder-centric large EEG model that reformulates EEG modeling as sequence-to-sequence learning and incorporates discrete support samples to construct contextual cues, equipping the model with in-context learning. ECHO matters as a baseline because it is the strongest published LLM-aligned EEG model that, like Ewen, foregrounds the language-side decoder; it provides a relevant comparison with a decoder-centered system, while the within-Ewen ablations examine individual component contributions.

CELM (Pradeepkumar et al., 2026a). A clinical EEG–language model that connects pretrained EEG representations to a language decoder through sequence-aware alignment. In Table 3, CELM denotes our window-adapted implementation based on its official code. We freeze the pretrained CBraMod encoder and Qwen3.5-0.8B backbone, and train the sequence projector, EEG boundary embeddings, and waveform-derived auxiliary token modules on the same narrative supervision. The projector is initialized for this task, and evaluation uses generated text under the common scoring protocol.

F CONSTRUCTION OF DESCRIPTORIZED PHYSIOLOGICAL COVARIATES

This appendix specifies the descriptor bank used by Eq. (2) and Eq. (4). The bank is curated from a high-quality literature shortlist; the full reference list is shipped as supplementary material. We retain 18 metric types, representing 22 numerical values in total, organized into seven groups: EOG self-state (3), EOG–EEG coupling (2), ECG self-state (3), ECG–EEG coupling (2), EMG self-state (3), EMG–EEG coupling (2), and EEG self-state (3). For every indicator we list its name, the modality family, and the numerical value or values extracted from the recording window. Relative band power is one metric type with five values; the remaining 17 types each contribute one value. Let $X^{(k)} \in \mathbb{R}^{C_k \times N}$ denote the windowed signal of modality k, with N samples at sampling rate f_s and duration $T = N/f_s$ seconds. Let RR_i denote an inter-beat interval in seconds, N_{RR} the number of valid intervals, and $P_{xx}(f)$ the Welch power spectral density.

F.1 EOG SELF-STATE

Blink Rate. Number of detected blinks normalized to events per minute,

$$\mathrm{BlinkRate} \;=\; \frac{N_{\mathrm{blink}}}{T} \cdot 60. \tag{51}$$

Saccade Rate. Number of detected saccades per minute,

$$\mathrm{SaccadeRate} \;=\; \frac{N_{\mathrm{saccade}}}{T} \cdot 60. \tag{52}$$

Median Saccadic Peak Velocity. Median across saccades of the maximum absolute derivative of the EOG channel,

$$v_{\mathrm{peak}} \;=\; \mathrm{median}_{s \in \mathcal{S}} \left(\max_{t \in s} \left| \tfrac{d\,\mathrm{EOG}}{dt}(t) \right| \right). \tag{53}$$

F.2 EOG–EEG COUPLING

Max Lagged Correlation. Maximum cross-correlation between EOG and a representative EEG channel within a ± 200 ms lag window,

$$\rho_{\max} \;=\; \max_{|\tau| \le 200\,\mathrm{ms}} \mathrm{corr}\big(\mathrm{EOG}(t), \mathrm{EEG}(t+\tau)\big). \tag{54}$$

EOG→EEG Regression Gain. Per-window slope of a linear least-squares fit predicting EEG from EOG,

$$\mathrm{EEG}(t) \;=\; \beta\,\mathrm{EOG}(t) + \varepsilon, \qquad \hat{\beta} \;=\; \frac{\langle \mathrm{EEG}, \mathrm{EOG} \rangle}{\|\mathrm{EOG}\|_2^2 + \epsilon}. \tag{55}$$

F.3 ECG SELF-STATE

Mean Heart Rate.

$$\mathrm{HR} \;=\; \frac{60}{\overline{\mathrm{RR}}}. \tag{56}$$

SDNN. Standard deviation of normal-to-normal RR intervals,

$$\mathrm{SDNN} \;=\; \mathrm{std}_i(\mathrm{RR}_i). \tag{57}$$

RMSSD. For $N_{\mathrm{RR}} \ge 2$ consecutive valid inter-beat intervals,

$$\mathrm{RMSSD} \;=\; \sqrt{\tfrac{1}{N_{\mathrm{RR}}-1} \textstyle\sum_{i=1}^{N_{\mathrm{RR}}-1} (\mathrm{RR}_{i+1} - \mathrm{RR}_i)^2}. \tag{58}$$

F.4 ECG–EEG COUPLING

Heartbeat-Evoked Potential (HEP) Amplitude. EEG amplitude averaged over the fixed post-R-peak interval $[t_R + \Delta_1, t_R + \Delta_2]$, with $\Delta_2 > \Delta_1$. Here $\mathcal{R}$ contains R peaks whose full averaging interval is available in the window:

$$\mathrm{HEP} \;=\; \frac{1}{|\mathcal{R}|} \sum_{t_R \in \mathcal{R}} \frac{1}{\Delta_2 - \Delta_1} \int_{\Delta_1}^{\Delta_2} \mathrm{EEG}(t_R + u)\, du. \tag{59}$$

Cardio-Cerebral Phase-Locking Value (PLV). Phase synchronization between EEG and ECG instantaneous phases,

$$\mathrm{PLV} \;=\; \left| \frac{1}{N} \sum_{n=1}^{N} \exp\big(j(\phi_{\mathrm{EEG}}(n) - \phi_{\mathrm{ECG}}(n))\big) \right|. \tag{60}$$

F.5 EMG SELF-STATE

Root Mean Square Amplitude.

$$\mathrm{RMS} \;=\; \sqrt{\tfrac{1}{N} \textstyle\sum_{i=1}^{N} x_i^2}. \tag{61}$$

Median Frequency. Spectral median where the cumulative power crosses one half of the total in $[0, f_s/2]$,

$$\int_0^{\mathrm{MDF}} P_{xx}(f)\, df \;=\; \tfrac{1}{2} \int_0^{f_s/2} P_{xx}(f)\, df. \tag{62}$$

Spectral Entropy.

$$H \;=\; -\sum_i p_i \log p_i, \qquad p_i \;=\; \frac{P_{xx}(f_i)}{\sum_j P_{xx}(f_j)}. \tag{63}$$

F.6 EMG–EEG COUPLING

Corticomuscular Coherence (CMC). The scalar mean of magnitude-squared EEG–EMG coherence over the β band, with $f_\beta^- = 13\,\mathrm{Hz}$ and $f_\beta^+ = 30\,\mathrm{Hz}$:

$$\mathrm{CMC} \;=\; \frac{1}{f_\beta^+ - f_\beta^-} \int_{f_\beta^-}^{f_\beta^+} \frac{|S_{\mathrm{EEG,EMG}}(f)|^2}{S_{\mathrm{EEG,EEG}}(f)\, S_{\mathrm{EMG,EMG}}(f)}\, df. \tag{64}$$

Coherence Phase-Slope Delay. A scalar delay from the ordinary least-squares slope b_ϕ of unwrapped EEG–EMG cross-spectral phase over valid frequency bins in $[13, 30]$ Hz. Define $S_{xy} = X\overline{Y}$ with EEG as x and EMG as y:

$$(b_0, b_\phi) = \arg\min_{a,b} \sum_{f_j \in [13,30]} [\phi(f_j) - a - b f_j]^2, \qquad \tau = -\frac{b_\phi}{2\pi}. \tag{65}$$

Positive τ then denotes EEG lagging EMG under this cross-spectrum convention. At least two distinct valid frequency bins are required.

F.7 EEG SELF-STATE

Relative Band Power. This is one metric type contributing five numerical dimensions. For each canonical band $b \in \{\delta, \theta, \alpha, \beta, \gamma\}$,

$$\mathrm{RP}_b \;=\; \frac{\int_{f_l^b}^{f_h^b} P_{xx}(f)\, df}{\int_{f_{\min}}^{f_{\max}} P_{xx}(f)\, df}. \tag{66}$$

Mean Instantaneous α Frequency. Time-averaged instantaneous frequency of the α-filtered analytic signal,

$$f_\alpha(t) \;=\; \frac{1}{2\pi}\frac{d\phi_\alpha(t)}{dt}, \qquad \bar{f}_\alpha \;=\; \frac{1}{T}\int_0^T f_\alpha(t)\,dt. \tag{67}$$

Aperiodic Spectral Exponent. Slope of the $1/f$ background fit on the log–log spectrum,

$$\log P(f) \;=\; b - \chi \log f \quad \text{(fixed)}, \qquad \log P(f) \;=\; b - \log(k + f^{\chi}) \quad \text{(knee)}. \tag{68}$$

Stacking into the covariate vector. The 18 metric types contribute 22 numerical dimensions: EOG self-state (3), EOG–EEG coupling (2), ECG self-state (3), ECG–EEG coupling (2), EMG self-state (3), EMG–EEG coupling (2), and EEG self-state (7). The EEG block contains five relative band-power values, mean instantaneous α frequency, and the aperiodic spectral exponent. Following Section 3.1, we concatenate the EEG block first, then the EOG, ECG, and EMG blocks (self-state followed by coupling within each auxiliary modality), then the three binary availability indicators $(\delta_{\mathrm{EOG}}, \delta_{\mathrm{ECG}}, \delta_{\mathrm{EMG}})$. Thus $\mathbf{m} \in \mathbb{R}^{25}$: 22 numerical values plus 3 indicators. Numerical dimensions are z-scored using training-set statistics from observed values; missing auxiliary blocks are set to zero after standardization. The indicators remain binary and are not z-scored. Missing modalities therefore retain the same input layout. During downstream training and inference, the numerical vector is encoded as $\hat{\mathbf{c}} = E_\phi(\mathbf{m})$ by a two-layer MLP; descriptor text is not an additional encoder input. The resulting backbone-sized representation is used by the feature-wise affine modulation in Eq. (2) and the prefix-token pathway in Section 3.3.

G Generation Prompt and Answer Templates

Ewen is supervised on four field-aligned sub-tasks: *label*, *relation*, *band*, and *summary*. To mitigate prompt overfitting we register multiple paraphrases per task and randomize both the prompt template and the answer prefix during training. Tables 5–8 provide representative selections from the full template bank; the complete set of templates and their dataset-specific instantiations is included in the released code. Template fields such as `{dataset_name}` are filled per window from the dataset metadata. The field grammar, ground-truth fields, and description-bank references are used only by the post-hoc evaluator after generation, and are never provided in the prompt, retrieval context, or decoding constraints. Here, *field summary* denotes the summary sub-task evaluated by field accuracy. Table 9 gives five illustrative paraphrases of the combined *Summary* instruction: first a Direct summary, then one natural-language answer addressing withheld comparisons within the same EEG window.

Table 5: Representative label prompt templates.

Sub-task	Prompt template (question)	Answer-style hint
Label	In the `{dataset_name}` dataset, what category does this window belong to? For the `{dataset_name}` dataset, which class best matches this window? In `{dataset_name}`, identify the class of this EEG window. What is the clinical or task label for this `{dataset_name}` window? Classify the following `{dataset_name}` EEG window into one of the defined categories. Based on the EEG signal, which `{dataset_name}` label applies to this window? In the context of `{dataset_name}`, determine the class for this segment. Which predefined `{dataset_name}` category does this window most likely represent? $\cdots$ *additional label-question paraphrases are sampled during training.*	*Class label only.*

Table 6: Representative band prompt templates.

Sub-task	Prompt template (question)	Answer-style hint
Band	In the `{dataset_name}` dataset, which EEG band has the dominant power in this window? For this `{dataset_name}` window, what is the dominant EEG frequency band? In `{dataset_name}`, which EEG band is most prominent here? Identify the dominant spectral band in this `{dataset_name}` EEG window. What spectral band dominates this `{dataset_name}` recording segment? Describe the dominant EEG rhythm observed in this `{dataset_name}` window. $\cdots$ *additional band-question paraphrases are sampled during training.*	*One sentence naming the band.*

Table 7: Representative relation prompt templates.

Sub-task	Prompt template (question)	Answer-style hint
Relation	In the `{dataset_name}` dataset, describe the EEG-related physiological relationships in this window. For this `{dataset_name}` window, summarize the EEG-related coupling and regime patterns. In `{dataset_name}`, explain the EEG-linked physiological evidence in this window. What cross-modal physiological relationships are evident in this `{dataset_name}` window? Describe any EOG, ECG, or EMG interactions with the EEG in this `{dataset_name}` segment. Summarize the physiological coupling state observed in this `{dataset_name}` window. What physiological covariate patterns accompany the EEG signal in this `{dataset_name}` window? $\cdots$ *additional relation-question paraphrases are sampled during training.*	*Short coupling description.*

Table 8: Representative field-summary prompt templates.

Sub-task	Prompt template (question)	Answer-style hint
Field summary	Summarize this `{dataset_name}` window in a short paragraph. Write a brief natural-language summary for this `{dataset_name}` window. For this `{dataset_name}` window, provide one short paragraph. Give a concise neurophysiological description of this `{dataset_name}` window. Provide a structured narrative for this `{dataset_name}` EEG segment. Describe the key neurophysiological features of this `{dataset_name}` window in one paragraph. Synthesize context for this `{dataset_name}` segment into one brief paragraph. $\cdots$ *additional summary-question paraphrases are sampled during training.*	*One paragraph covering all fields.*

Answer-prefix bank. To avoid the model memorizing a single prefix string, the answer is rendered with a randomized prefix sampled uniformly from

{`Answer:`, `Response:`, `Report:`, `Interpretation:`, `Assessment:`, `Finding:`},

and so on (the full bank contains additional prefix variants registered in the released code).

Table 9: Representative summary prompt templates.

Sub-task	Prompt template (question)	Answer-style hint
Summary	Use the supplied EEG waveform representation and descriptor covariates for `{dataset_name}`. Write the Direct summary first, covering the state, dominant EEG band, physiological relations, and recording modalities. Then give exactly one natural-language answer to the withheld comparisons within the current EEG window.	*Direct summary first; then exactly one natural-language answer to the withheld waveform comparisons.*
	Describe the current EEG window from `{dataset_name}` using its waveform representation and descriptor covariates. Begin with the Direct summary: state, dominant EEG band, physiological relations, and recording modalities. Follow it with exactly one natural-language answer to the withheld waveform comparisons within this window.	
	For the supplied `{dataset_name}` EEG window, first write a Direct summary of the state, dominant EEG band, physiological relations, and recording modalities, using the waveform representation and descriptor covariates. Then answer the withheld comparisons within the window in exactly one natural-language response.	
	Using the EEG waveform representation and descriptor covariates for `{dataset_name}`, start with the Direct summary, including the state, dominant EEG band, physiological relations, and recording modalities. Next, provide exactly one natural-language answer to the withheld comparisons, based on the waveform within the current window.	
	Examine the supplied EEG waveform representation and descriptor covariates for the current `{dataset_name}` window. First summarize the state, dominant EEG band, physiological relations, and recording modalities as the Direct summary. Then give exactly one natural-language answer addressing the withheld waveform comparisons within that same window.	
	$\cdots$ *additional summary-question paraphrases are sampled during training.*	

Answer parsing. For *label* we case-fold the model's generated answer and run a normalized longest-match against the dataset label set. For *band* we match against the alias bank {delta, theta, alpha, beta, gamma} and the modifier suffixes (e.g. "alpha-dominant"). For *relation* we extract sentence-level evidence patterns of the form *"<modality> <level> / <coupling-strength>"*; mismatches at any field invalidate the relation match. For *summary* we re-extract {label, band, relation} sub-fields from the paragraph using the same parsers, then count it correct if every sub-field matches the ground-truth field.

Withheld-attribute parsing. We lowercase the generated summary, normalize whitespace, and match it against a fixed bank of attribute–state clauses. Exactly one matched state yields a prediction; absent, invalid, or conflicting matches are treated as missing. For an observable target, a correct prediction adds a true positive, an incorrect state adds a false positive to the predicted class and a false negative to the true class, and a missing prediction adds only a false negative to the true class. Unavailable targets are excluded. We aggregate these counts according to Eq. (37); auxiliary classifier logits are not used. The same clause bank and matching rules apply to all methods and input conditions.

H Code Access

The anonymous code repository for review will be available at `https://anonymous.4open.science/r/Ewen/`.

I Generation Results and BERTScore Breakdown

We provide dataset-level results supporting Figure 5; the BERTScore decomposition by reference distribution is shown in Figure 6. Three adaptation schemes are compared on three datasets that span clinical (Mumtaz 2016), cognitive (Mental Arithmetic), and motor-imagery (SHU-MI) paradigms. All three schemes share the same backbone, the same training data, and the same number of optimizer steps.

Because the generation stage is trained jointly across the three datasets, the training loss is shared by the whole mixture rather than tied to a single test set. Table 10 therefore reports it once by method. Table 11 reports evaluation loss together with Mix-BERT, and Table 12 reports the four per-field accuracies. Total F-Acc averages these field accuracies; its comparison across adaptation methods is shown in Figure 5. The field grammar, ground-truth fields, and description-bank references are used only by the post-hoc evaluator after generation, and are never provided in the prompt, retrieval context, or decoding constraints. Targets and covariates don't share discrete answer fields: covariates provide latent conditioning, while targets provide text supervision. Ground-truth rules are used only offline for references and evaluation, never in prompts, retrieval, or decoding, preventing template lookup.

Table 10: Joint training loss, reported once by method.

Method	Train Loss ↓
LoRA	0.0300 ± 0.0002
Full FT	$\mathbf{0.0262 \pm 0.0005}$
Ours	0.0275 ± 0.0003

Table 11: Generation: evaluation loss and Mix-BERTScore.

Dataset	Method	Eval Loss ↓	Mix-BERT ↑
Mumtaz	LoRA	0.0795 ± 0.0007	$\mathbf{0.9153 \pm 0.0075}$
	Full FT	$\mathbf{0.0449 \pm 0.0002}$	0.9039 ± 0.0042
	Ours	0.0472 ± 0.0007	0.9125 ± 0.0093
Mental Arithmetic	LoRA	0.1399 ± 0.0014	0.9080 ± 0.0041
	Full FT	0.1157 ± 0.0008	0.9024 ± 0.0036
	Ours	$\mathbf{0.1047 \pm 0.0017}$	$\mathbf{0.9115 \pm 0.0070}$
SHU-MI	LoRA	0.0446 ± 0.0006	$\mathbf{0.9157 \pm 0.0087}$
	Full FT	0.0461 ± 0.0003	0.9047 ± 0.0078
	Ours	$\mathbf{0.0425 \pm 0.0009}$	0.9125 ± 0.0097
Overall	LoRA	0.0880 ± 0.0009	$\mathbf{0.9130 \pm 0.0040}$
	Full FT	0.0689 ± 0.0004	0.9037 ± 0.0052
	Ours	$\mathbf{0.0648 \pm 0.0011}$	0.9122 ± 0.0086

Table 12: Generation: per-field F-Acc on the four sub-tasks.

Dataset	Method	Label ↑	Summary ↑	Relation ↑	Band ↑
Mumtaz	LoRA	0.6134 ± 0.0092	0.8721 ± 0.0157	0.8858 ± 0.0062	$\mathbf{0.9960 \pm 0.0030}$
	Full FT	$\mathbf{0.9054 \pm 0.0091}$	$\mathbf{0.9451 \pm 0.0123}$	$\mathbf{0.9735 \pm 0.0156}$	0.9609 ± 0.0183
	Ours	0.8932 ± 0.0063	0.9368 ± 0.0103	0.9379 ± 0.0131	0.9930 ± 0.0060
Mental Arithmetic	LoRA	0.4917 ± 0.0079	0.5592 ± 0.0106	0.4592 ± 0.0028	$\mathbf{0.9960 \pm 0.0030}$
	Full FT	0.6688 ± 0.0094	0.7192 ± 0.0122	$\mathbf{0.6480 \pm 0.0130}$	0.8542 ± 0.0043
	Ours	$\mathbf{0.7625 \pm 0.0046}$	$\mathbf{0.7201 \pm 0.0065}$	0.6376 ± 0.0077	0.9896 ± 0.0060
SHU-MI	LoRA	0.5953 ± 0.0119	0.8041 ± 0.0040	0.9795 ± 0.0078	0.9702 ± 0.0097
	Full FT	0.5958 ± 0.0071	0.8780 ± 0.0132	$\mathbf{0.9974 \pm 0.0020}$	0.9024 ± 0.0063
	Ours	$\mathbf{0.6021 \pm 0.0048}$	$\mathbf{0.8805 \pm 0.0088}$	0.9955 ± 0.0040	$\mathbf{0.9855 \pm 0.0080}$
Overall	LoRA	0.5668 ± 0.0096	0.7451 ± 0.0052	0.7748 ± 0.0056	0.9874 ± 0.0052
	Full FT	0.7233 ± 0.0085	$\mathbf{0.8474 \pm 0.0125}$	$\mathbf{0.8730 \pm 0.0102}$	0.9058 ± 0.0054
	Ours	$\mathbf{0.7526 \pm 0.0052}$	0.8458 ± 0.0085	0.8570 ± 0.0082	$\mathbf{0.9894 \pm 0.0066}$

We additionally evaluate label generation under a six-dataset joint-training protocol. Labels are recovered from generated responses using the normalized parser in Appendix G. This larger run was performed once; Table 13 reports its single-run label accuracy. Its results and the closed-vocabulary

results in Table 1 use different training and evaluation protocols, so their differences cannot be attributed solely to the output format. Description-bank references are initialized from standardized field schemas and manually refined by domain experts for field-level evaluation.

Table 14 presents representative qualitative outputs from Ewen Full. The descriptions include the predicted states and qualitative spectral, temporal, and spatial comparisons; numerical magnitudes are omitted for readability. These examples illustrate the answer content; aggregate performance is reported in Table 3.

Table 13: Label accuracy under six-dataset joint training on the six primary benchmarks.

Dataset	**Label Accuracy ↑**
TUAB	0.8032
Mental Arithmetic	0.6771
SHU-MI	0.5624
Mumtaz 2016	0.8923
BCIC IV 2a	0.4387
SEED-V	0.2494

Table 14: Representative qualitative examples from *Ewen* Full. Sample IDs identify the illustrated recordings; aggregate test performance is reported separately.

Dataset	**Summary**
Mumtaz MDD-S30 EO Window 42	The model labels this EEG-only window as major depressive disorder. Delta is the strongest band, with theta the next most prominent component. Alpha is weaker than delta, and its mean frequency lies in the middle of the alpha range. The background spectrum has a steep aperiodic slope, with stronger low-frequency content. From the first to the second half, the spectral balance shifts further toward delta as the relative contributions of theta, alpha and beta decline. The spatial picture adds a different distinction: alpha power is greater over the frontal than the posterior region. Together, these observations describe an evolving low-frequency-weighted window with a regional alpha contrast, rather than a uniform change at every electrode.
Arithmetic Subject29 Recording 1 Window 1	Alpha is the predominant EEG band in this EEG-and-ECG recording, followed by delta. The alpha rhythm has a mid-range mean frequency, while the aperiodic slope is moderate. The cardiac descriptors indicate a high heart rate and strong cardio-cerebral coupling. The model identifies the state as background before the mental arithmetic task. From the first to the second half, the relative contribution of delta increases and that of alpha decreases, while theta and beta remain approximately stable. Alpha activity is stronger at posterior than frontal sites, with little left–right difference at the occipital sites. The left and right central sites also show comparable activity across the alpha-to-beta range.
SHU–MI sub-020 Session 02 Window 5	This EEG-only example is assigned to right-hand motor imagery. Its spectrum is strongly dominated by delta, with beta the next strongest band. Alpha is less prominent than the dominant delta component, and its mean frequency is toward the faster end of the alpha range. The aperiodic background has a moderate slope. The spatial comparison shows stronger posterior than frontal alpha activity, while O1 and O2 are approximately equal. C3 and C4 likewise have comparable activity across the alpha-to-beta range. In the later half of the window, the relative delta contribution is lower and the theta, alpha, and beta contributions are higher.

Figure 9 illustrates complementary waveform comparisons in three separate samples. Direct descriptors give a mean heart rate of 97 bpm and alpha power of 0.250 for Arithmetic, delta/alpha powers of 0.652/0.0387 for SHU–MI, and alpha/beta powers of 0.0924/0.128 for Mumtaz. The withheld relations are more local: Arithmetic O1 variability differs between cardiac cycles 4 and 6 (17.10 vs. 6.18 μV; RR intervals 610 vs. 628 ms); SHU–MI shows stronger O1–O2 than F3–F4 correlation (0.944 vs. 0.148); and Mumtaz shows second-to-first-half SD ratios of 4.26 at Fp1 and 0.98 at O1.

I.1 BERTScore Decomposition by Reference Distribution

EEG-conditioned BERTScore alone cannot establish language retention: adaptation may yield high similarity to dataset references while changing general text behavior. We therefore decompose

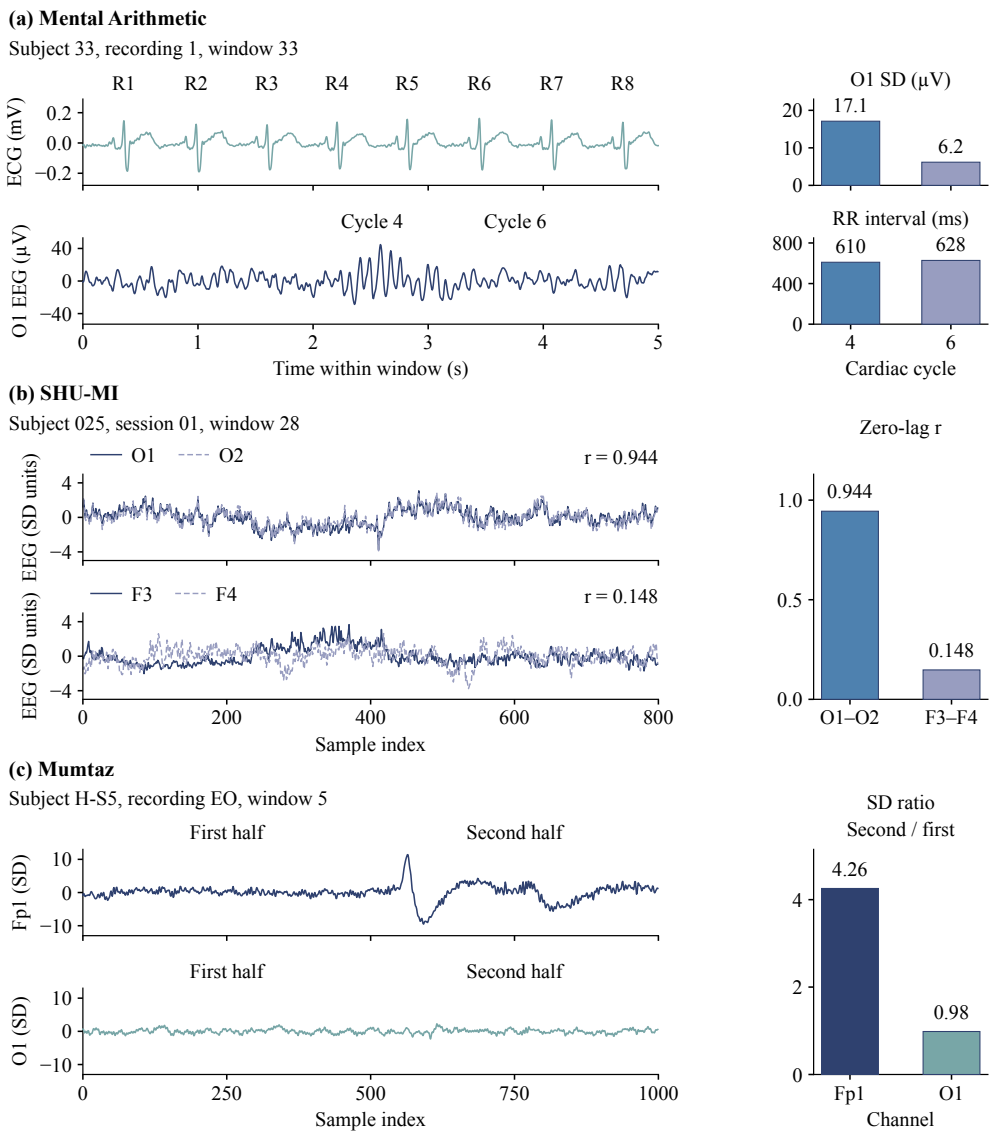

Figure 9: Waveform examples from three datasets: (a) cardiac-cycle variation in Arithmetic, (b) channel-pair correspondence in SHU–MI, and (c) temporal variability in Mumtaz. SHU–MI traces are detrended and individually standardized; Mumtaz traces are detrended on a shared amplitude scale. Panels (b)–(c) use sample indices.

BERTScore by reference distribution, exactly as visualized in Figure 6 of the main paper. *OWT-BERT* is the score against a balanced openwebtext continuation stream sized to each dataset's test split, *Dataset-BERT* is the score against the dataset's own *summary* test references, and *Mix-BERT* is the score against their balanced union.

J Manual Expert Audit for Narration

This section specifies the manual audit that complements the automatic BERTScore decomposition in Figure 6. The scored outputs are judged by EEG-trained human experts using a fixed rubric. For reproducibility and to avoid conflating in-domain template matching with language adaption, the audit covers the three EEG datasets and a held-out OpenWebText control stream.

The five scoring dimensions are: *factual consistency*, *descriptor grounding*, *physiological or discourse coherence*, *specificity*, and *fluency*. For EEG samples, descriptor grounding means that the generated

Table 15: Manual expert-audit sampling plan. Each source contributes 50 windows or continuations, and every output is scored on the same five 1–5 Likert axes.

Source	Sampling unit	Sample count	Stratification / scoring focus
Mumtaz 2016	Test EEG window	50	Balanced over healthy control and major depressive disorder; score label grounding, spectral descriptors, and narration fluency.
Mental Arithmetic	Test EEG window	50	Balanced over baseline and arithmetic-task windows; score EEG–ECG relation grounding and cognitive-state narration.
SHU–MI	Test EEG window	50	Balanced over left- and right-hand motor imagery; score motor-imagery label grounding and rhythm descriptions.
OpenWebText	Held-out continuation	50	Random held-out text continuations; score semantic fidelity, discourse coherence, specificity, and fluency.

band, relation, modality, and summary claims are supported by the extracted window descriptors. For OpenWebText, descriptor grounding is replaced by continuation grounding, i.e., whether the generated continuation preserves the reference semantics without drifting into a different topic. The per-source score is the average over 50 samples, and the final mean is the unweighted average over the five dimensions.

Table 16: Manual-audit reporting grid for the three adaptation methods. Best means and their SD terms are **bold**.

Source	Method	Fact.	Ground.	Coher.	Spec.	Flu.	Mean
Mumtaz	LoRA	3.55 ±0.16	3.72 ±0.15	3.64 ±0.10	3.58 ±0.11	**4.48 ±0.10**	3.79 ±0.12
	Full FT	**4.54 ±0.09**	**4.60 ±0.13**	4.42 ±0.12	4.38 ±0.10	4.02 ±0.15	4.39 ±0.11
	Ours	4.50 ±0.11	4.56 ±0.09	**4.48 ±0.10**	**4.44 ±0.10**	4.30 ±0.11	**4.46 ±0.09**
Mental Arithmetic	LoRA	3.18 ±0.12	3.10 ±0.16	3.20 ±0.09	3.06 ±0.12	**4.43 ±0.14**	3.39 ±0.12
	Full FT	3.70 ±0.16	3.72 ±0.11	3.62 ±0.14	3.55 ±0.10	3.92 ±0.13	3.70 ±0.12
	Ours	**4.05 ±0.11**	**4.00 ±0.12**	**3.94 ±0.15**	**3.88 ±0.12**	4.24 ±0.10	**4.02 ±0.09**
SHU–MI	LoRA	3.70 ±0.10	3.82 ±0.12	3.92 ±0.14	3.75 ±0.11	**4.45 ±0.12**	3.93 ±0.10
	Full FT	3.88 ±0.11	4.10 ±0.09	4.08 ±0.12	3.92 ±0.10	3.94 ±0.11	3.98 ±0.10
	Ours	**4.14 ±0.11**	**4.20 ±0.15**	**4.18 ±0.15**	**4.08 ±0.16**	4.28 ±0.10	**4.18 ±0.09**
OpenWebText	LoRA	**4.42 ±0.15**	4.26 ±0.12	4.28 ±0.13	4.05 ±0.16	**4.58 ±0.11**	4.32 ±0.12
	Full FT	4.08 ±0.11	4.10 ±0.12	4.02 ±0.14	3.86 ±0.13	4.06 ±0.13	4.02 ±0.12
	Ours	4.30 ±0.09	**4.36 ±0.15**	**4.45 ±0.11**	**4.18 ±0.11**	4.38 ±0.10	**4.33 ±0.10**
Overall	LoRA	3.71 ±0.13	3.73 ±0.12	3.76 ±0.11	3.61 ±0.11	**4.49 ±0.11**	3.86 ±0.11
	Full FT	4.05 ±0.11	4.13 ±0.10	4.04 ±0.13	3.93 ±0.10	3.99 ±0.13	4.02 ±0.11
	Ours	**4.25 ±0.10**	**4.28 ±0.11**	**4.26 ±0.12**	**4.15 ±0.12**	4.30 ±0.10	**4.25 ±0.09**

The pilot audit assesses output quality within the sampled domains; it does not establish unchanged general language competence. It follows the same interpretation as Figure 6. On OpenWebText, Ours leads on continuation grounding, discourse coherence, specificity, and the five-axis mean, whereas LoRA remains highest on factual consistency and fluency. Full FT is competitive on dataset-grounded fields but loses fluency after unrestricted task adaptation. *Ewen* provides the best overall manual score across the evaluated sources while adding stronger EEG grounding than LoRA and better-rated continuation quality than Full FT on the sampled outputs. Thus the expert-audit grid and the BERTScore decomposition test the same claim from complementary views: BERTScore quantifies reference-distribution alignment, while manual scoring checks whether that alignment corresponds to readable and physiologically grounded narration.

K Additional Closed-Vocabulary Results

K.1 Experiments on TUEV and BCIC 2020-T3

To complete the eight-benchmark coverage referenced in Section 4.1, Tables 17 and 18 report the two remaining datasets that were omitted from the main comparison for space, namely TUEV (epileptiform event classification) and BCIC 2020–T3 (imagined-speech decoding). The same baselines and the same protocol as Table 1 are used. We use the suffixes **S**, **B**, and **L** to denote our 0.8B, 2B, and 4B variants, consistent with the main text.

On TUEV, *Ewen-L* reaches 0.629 ACC-B and is competitive with NeuroLM, ECHO, BIOT, and EEGNet, but trails CBraMod (the strongest encoder-only baseline) by 3.81 percentage points (0.6671 versus 0.6290). One possible explanation for this gap is a structural mismatch between the TUEV protocol and an LLM-aligned, descriptor-conditioned model. TUEV is a six-way short-event classifier in which each label is determined by the local morphology of a single 5-second window, with categories such as SPSW, GPED, PLED, EYEM, ARTF, and BCKG that are specifically defined as transient waveform shapes; the discriminative information lives at the sample-level transient scale rather than at the window-level descriptor scale that our covariate channel exposes. The auxiliary modalities that motivate our descriptorized covariates – ocular, cardiac, and muscular signals – are also largely uninformative for spike-versus-background discrimination, so the additional context that helps Mental Arithmetic and Mumtaz 2016 contributes little here. On BCIC 2020–T3, the three Ewen variants remain close to ECHO and slightly improve the best ACC-B and Cohen's Kappa, indicating that the descriptor-conditioned interface is more compatible with the imagined-speech setting than with transient epileptiform event detection. Encoder-only specialists like CBraMod, conversely, are trained end-to-end on this exact transient-morphology objective with criss-cross spatio-temporal attention and therefore retain a sharper inductive bias for sample-level epileptiform shape classification.

Table 17: Multi-class classification on the additional TUEV benchmark.

Methods	ACC-B	Cohen's Kappa	Weighted F1
EEGNet	0.3975 ± 0.0107	0.3493 ± 0.0113	0.6384 ± 0.0206
EEGPT	0.6232 ± 0.0114	0.6351 ± 0.0134	0.8187 ± 0.0063
BIOT	0.5281 ± 0.0225	0.5273 ± 0.0249	0.7492 ± 0.0082
LaBraM	0.6409 ± 0.0065	0.6637 ± 0.0093	0.8312 ± 0.0050
CBraMod	$\mathbf{0.6671 \pm 0.0107}$	$\mathbf{0.6772 \pm 0.0096}$	$\mathbf{0.8342 \pm 0.0064}$
NeuroLM	0.4679 ± 0.0356	0.4570 ± 0.0498	0.7359 ± 0.0219
ECHO	0.5214 ± 0.0026	0.5085 ± 0.0010	0.7489 ± 0.0028
Ewen-S	0.6147 ± 0.0178	0.5391 ± 0.0084	0.7559 ± 0.0131
Ewen-B	0.6223 ± 0.0068	0.5468 ± 0.0149	0.7635 ± 0.0010
Ewen-L	0.6290 ± 0.0122	0.5515 ± 0.0141	0.7702 ± 0.0171

Table 18: Multi-class classification on the additional BCIC 2020–T3 benchmark.

Methods	ACC-B	Cohen's Kappa	Weighted F1
EEGNet	0.2600 ± 0.0151	0.0750 ± 0.0189	0.2573 ± 0.0163
EEGPT	0.2453 ± 0.0131	0.0567 ± 0.0164	0.2441 ± 0.0105
BIOT	0.4920 ± 0.0086	0.3650 ± 0.0176	0.4917 ± 0.0079
LaBraM	0.5060 ± 0.0155	0.3800 ± 0.0242	0.5054 ± 0.0205
CBraMod	0.5373 ± 0.0108	0.4216 ± 0.0163	0.5383 ± 0.0096
NeuroLM	0.3186 ± 0.0162	0.1483 ± 0.0108	0.3113 ± 0.0145
ECHO	0.5432 ± 0.0022	0.4267 ± 0.0014	0.5410 ± 0.0007
Ewen-S	0.5488 ± 0.0034	0.4317 ± 0.0076	$\mathbf{0.5485 \pm 0.0098}$
Ewen-B	0.5427 ± 0.0172	0.4257 ± 0.0150	0.5416 ± 0.0164
Ewen-L	$\mathbf{0.5494 \pm 0.0078}$	$\mathbf{0.4339 \pm 0.0190}$	0.5461 ± 0.0121

K.2 Covariate-Free Comparison with EEG-Only LLM Baselines

The purpose of this comparison is to disentangle architectural gains from gains due to auxiliary biosignals. Removing Cov. disables both the descriptor pathway and the covariate-aware geometry, yielding an EEG-only Ewen-B that is directly comparable to NeuroLM and ECHO. The covariate-free model remains competitive or superior to these baselines, while the full model in Table 2 further improves performance when physiological context is available. Thus, covariates act as an additive contextual signal rather than the sole source of improvement.

NeuroLM and ECHO are LLM-aligned EEG models that consume EEG tokens alone, with no auxiliary biosignal pathway. To compare with baselines fairly, we run *Ewen-B* with the descriptor pathway and the covariate-aware token geometry both removed; the resulting model receives the same EEG token stream as NeuroLM and ECHO. Table 19 reports the comparison on Mental Arithmetic, SHU–MI, and Mumtaz 2016.

For SHU–MI and Mumtaz 2016 the covariate-free *Ewen-B* numbers are taken directly from the bottom row; for Mental Arithmetic, which is not part of that ablation grid, we re-trained *Ewen-B* under the same covariate-free configuration. The NeuroLM and ECHO entries are copied verbatim from Table 1. Across the three datasets, *Ewen-B* (no Cov.) wins 7 of 9 metric cells: it leads every ACC-B and PR AUC entry, and surpasses NeuroLM on every cell. ECHO leads on ROC AUC for Mental Arithmetic (by 0.0307) and for Mumtaz 2016 (by 0.0015); the latter is within one standard deviation. This pattern indicates that the gain over the EEG-only baselines in the main table is not solely produced by access to auxiliary biosignals.

Table 19: Covariate-free comparison of *Ewen-B* against NeuroLM and ECHO.

Dataset	Method	ACC-B ↑	ROC AUC ↑	PR AUC ↑
Mental Arithmetic	NeuroLM	0.6345 ± 0.0442	0.6130 ± 0.0764	0.5889 ± 0.0423
	ECHO	0.6851 ± 0.0032	$\mathbf{0.7500 \pm 0.0062}$	0.7530 ± 0.0015
	Ewen-B	$\mathbf{0.7184 \pm 0.0181}$	0.7193 ± 0.0146	$\mathbf{0.8462 \pm 0.0193}$
SHU–MI	NeuroLM	0.5416 ± 0.0106	0.5503 ± 0.0137	0.5599 ± 0.0099
	ECHO	0.5406 ± 0.0009	0.5537 ± 0.0021	0.5508 ± 0.0013
	Ewen-B	$\mathbf{0.5820 \pm 0.0171}$	$\mathbf{0.6259 \pm 0.0074}$	$\mathbf{0.6370 \pm 0.0139}$
Mumtaz 2016	NeuroLM	0.8935 ± 0.0048	0.9657 ± 0.0063	0.9704 ± 0.0051
	ECHO	0.9056 ± 0.0012	$\mathbf{0.9745 \pm 0.0023}$	0.9748 ± 0.0015
	Ewen-B	$\mathbf{0.9078 \pm 0.0108}$	0.9730 ± 0.0046	$\mathbf{0.9758 \pm 0.0152}$

L Sensitivity to EEG Input Mask Ratio

Figure 10 visualises the Ewen performance under varying EEG input mask ratios on SHU-MI, Mumtaz 2016, and BCIC IV 2a. The nine dataset–metric curves generally perform well at moderate mask ratios, but their optima differ: Mumtaz PR AUC peaks at $r = 0.1$, while SHU–MI balanced accuracy peaks at $r = 0.3$. Some curves also show small rebounds at larger ratios. These results support $r = 0.2$ as a shared operating choice, without establishing a universal optimum or a monotonic decline beyond it.

M Ewen Training and Inference Algorithm

Algorithm 1 summarizes the pipeline in Section 3. EEG descriptors and available auxiliary descriptors form a semantic covariate that conditions the waveform tokens. Task semantics determine the prompt. During training, OWT gradients are refreshed at the current decoder and used to attenuate reference-aligned components of the core task gradient. The residual and other trainable EEG-side parameters receive their task gradients directly. The optimizer operates on these supplied gradients, as discussed in Appendix C.4. The target matrices and parameter settings are specified in Appendix C.4.

Here θ_{core} contains the row-direction coefficient vectors on all configured decoder targets, θ_{res} contains their rank-16 residual factors, and θ_{EEG} denotes the remaining trainable EEG-side parameters,

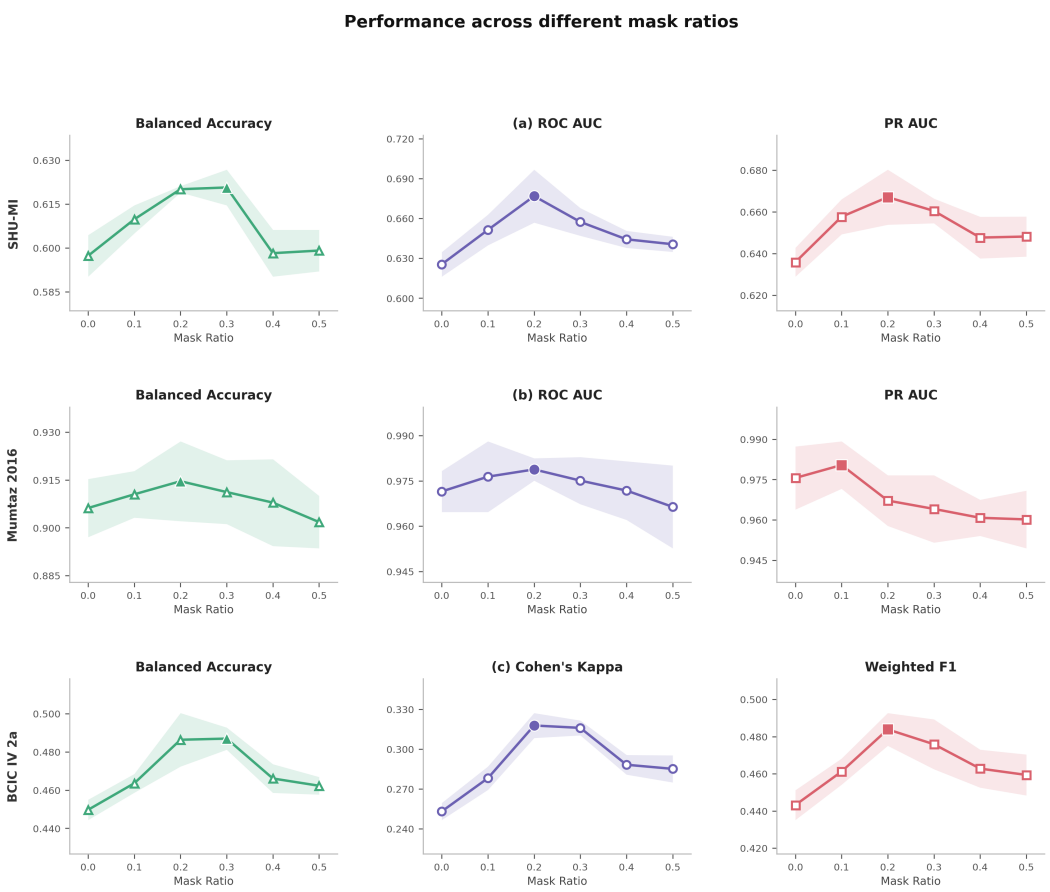

Figure 10: Performance of *Ewen-B* across EEG input mask ratios on three benchmarks. Each row corresponds to one dataset (SHU-MI, Mumtaz 2016, BCIC IV 2a) and contains three metric panels.

including the covariate encoder. All task and OWT gradients in one step are evaluated before the optimizer changes any parameter.

N Training Recipe and Compute Budget

Figures 4–6 report results from a joint training on three datasets: Mumtaz 2016, Mental Arithmetic, and SHU–MI. All runs use PyTorch DDP with `bfloat16` mixed precision and selective gradient checkpointing on the backbone transformer blocks, on $4\times$ NVIDIA H200 (141 GB) GPUs. The optimizer is AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1, and gradient clipping at 1.0; the per-device EEG batch size is 48 and the text batch size is 8. The EEG input mask ratio is 0.2. Training runs for 10 epochs with a cosine learning-rate schedule and a linear warm-up over the first 3% of steps; the LoRA adapters use a peak learning rate of 2×10^{-5} and the orthogonal-update branch uses 2×10^{-5}, both decaying to 10% of the peak. Table 13 additionally reports an extended six-dataset joint run under the same hyperparameters.

For close-vocabulary tasks, each dataset is trained independently for 50 epochs. The optimizer and schedule follow the same AdamW setup described above. Most datasets run on NVIDIA A100 (40 GB) GPUs with a per-device EEG batch size of 8; TUAB, SEED-V, and TUEV run on H200 GPUs with a per-device batch size of 32. The EEG input mask ratio is 0.2 across all datasets. The majority of datasets share a learning rate of 2×10^{-5} and 2 temporal convolution layers in the classification head. The common close-vocabulary recipe uses a learning rate of 2×10^{-5} and 2 temporal convolution layers in the classification head; Mumtaz 2016 uses 2×10^{-5} with 4 layers, and SHU-MI uses 4×10^{-5} with 4 layers.All other optimization choices, including the optimizer, schedule, tokenizer, and covariate pathways, are shared across datasets. Pretraining is conducted on

Algorithm 1 Ewen training and inference

Require: EEG window $X^{(0)}$, available auxiliary signals $\{X^{(k)}\}$ with indicators $\boldsymbol{\delta}$, task semantics ℓ; target $\mathbf{y}$ and OWT stream for training

Condition the EEG tokens

$\mathbf{z}_{1:T_z} \leftarrow \mathcal{Q}(X^{(0)})$

$\mathbf{m} \leftarrow [g_0(X^{(0)});\ \|_{k\in\mathcal{K}}\delta_k g_k(X^{(k)});\ \boldsymbol{\delta}]$

$\hat{\mathbf{c}} \leftarrow E_\phi(\mathbf{m})$

$\tilde{\mathbf{e}}_i \leftarrow \Gamma(\hat{\mathbf{c}}, \mathbf{r}_i) \odot \mathbf{e}_i + \mathrm{B}(\hat{\mathbf{c}}, \mathbf{r}_i)$ for each token i

$\mathbf{h}_i^{(0)} \leftarrow \tilde{\mathbf{e}}_i + \mathbf{p}_i^{\mathrm{loc}} + \mathbf{p}_w^{\mathrm{glob}}$ for each token in window w

$\mathbf{x}^{\mathrm{prompt}} \leftarrow \mathcal{P}(\ell)$

if training **then**

  Compute $\mathcal{L}_{\mathrm{task}} = \mathcal{L}_{\mathrm{gen}} + \lambda_{\mathrm{cov}}\mathcal{L}_{\mathrm{cov}}$ from the conditioned input and $\mathbf{y}$

  $\mathbf{g}_{\mathrm{core},t} \leftarrow \nabla_{\theta_{\mathrm{core}}}\mathcal{L}_{\mathrm{task}}$

  $\mathbf{g}_{\mathrm{res},t} \leftarrow \nabla_{\theta_{\mathrm{res}}}\mathcal{L}_{\mathrm{task}}, \quad \mathbf{g}_{\mathrm{EEG},t} \leftarrow \nabla_{\theta_{\mathrm{EEG}}}\mathcal{L}_{\mathrm{task}}$

  for $k = 1, \ldots, K$ **do**

    Sample an OWT micro-batch and compute its text-only loss at θ_t

    $\mathbf{g}_{\mathrm{lang},t}^{(k)} \leftarrow \nabla_{\theta_{\mathrm{core}}}\mathcal{L}_{\mathrm{lang}}^{(k)}(\theta_t)$

  end for

  $\mathbf{G}_t \leftarrow [\mathbf{g}_{\mathrm{lang},t}^{(1)}, \ldots, \mathbf{g}_{\mathrm{lang},t}^{(K)}]$

  $\alpha_t \leftarrow \rho\,\mathrm{tr}(\mathbf{G}_t^\top \mathbf{G}_t)/K$, with fixed $\rho > 0$

  If $\alpha_t = 0$, set $\mathbf{u}_t \leftarrow \mathbf{g}_{\mathrm{core},t}$; otherwise solve Eq. 50

  Clip the combined supplied gradient to global norm 1.0

  Supply $(\mathbf{u}_t, \mathbf{g}_{\mathrm{res},t}, \mathbf{g}_{\mathrm{EEG},t})$ to the optimizer for $(\theta_{\mathrm{core}}, \theta_{\mathrm{res}}, \theta_{\mathrm{EEG}})$

  Take one optimizer step; keep all pretrained parameters and row-norm buffers fixed

else

  return $\mathrm{Generate}_\theta(\mathbf{h}_{1:T_z}^{(0)}, \mathbf{x}^{\mathrm{prompt}})$

end if

the training splits of eight datasets. We use the accompanying descriptions together with EEG tokens to perform self-supervised next-token prediction. All pretraining runs are carried out on NVIDIA H200 GPUs for 10 epochs.

O Ewen Module Configuration

Ewen is instantiated at three scales by varying only the size of the language backbone, which is Qwen3.5 in every case. The EEG-side tokenizer (Appendix B), the descriptorized covariate projector, the hierarchical position encoder, and the orthogonal-update adapter use the same design across all three scales; matrix shapes follow the backbone dimensions. Per-scale parameter counts are reported in Table 20; the EEG side adds $< 1\%$ extra parameters in every case.

Table 20: *Ewen* variants by language-backbone scale. The EEG side, tokenizer, covariate projector, and adapter are shared.

Variant	Backbone params	Residual LoRA rank
Ewen-S	0.8B	16
Ewen-B	2.0B	16
Ewen-L	4.0B	16

P Limitations and Broader Impacts

Limitations. The current work has several limitations that point toward natural future directions. The descriptor bank is presently scoped to EOG, ECG, EMG, and EEG-intrinsic spectral features; extending it to further physiological modalities such as respiration or skin conductance, or to finer-grained temporal features, is straightforward within the proposed formalism and is left for future work. All experiments instantiate the language backbone with Qwen3.5, which supplies the language backbones at the three evaluated scales; evaluating the framework with other backbone families is an

interesting direction, though the orthogonal update formulation is backbone-agnostic in principle. The evaluation protocol follows segment-level benchmarks that are standard in the field; long-horizon or longitudinal EEG analysis where physiological context evolves over hours is a complementary setting not addressed here.

Broader impacts. The primary positive impact of *Ewen* lies in broadening the interface between EEG recordings and general language models: by grounding neural signals in physiological descriptors and enabling generation, the framework can assist clinicians and neuroscientists in articulating and communicating complex physiological states, and may lower the barrier to EEG-based analysis in resource-limited settings. The general principle of descriptorized covariate conditioning with adaptation designed to control language drift is not EEG-specific and could transfer to other physiological time-series domains. On the risk side, any system that produces natural-language neurophysiological descriptions carries the potential for over-reliance in clinical workflows; *Ewen* is a research prototype and is not intended for standalone diagnostic use without expert oversight. Processing multi-modal physiological signals raises standard data-governance considerations, and deployments should apply appropriate de-identification measures consistent with applicable regulations.

Q Expert Summary Evaluation

Expert summary evaluation separates recovery of predefined waveform comparisons from the factual quality of generated summaries. Withheld F1 evaluates a common set of physiological comparisons; expert precision, reference coverage, and Expert F1 evaluate the claims selected by each summary. Agreement between the two annotators is reported separately from model performance.

Q.1 Withheld F1

Withheld evaluation covers temporal and spatial spectral comparisons whose states are not explicitly supplied by the whole-window, channel-pooled descriptors. Pre-annotated reference measurements and comparison states provide the evaluation targets. Examples include first-to-second-half changes in relative δ, θ, α, and β power using equal halves and the same valid channels, and spatial contrasts in α power between frontal and posterior regions or bilateral occipital channels (O1–O2), together with combined α–β power contrasts between bilateral central channels (C3–C4). The respective bands are $[0.5, 4)$, $[4, 8)$, $[8, 13)$, and $[13, 30]$ Hz, normalized by total $[0.5, 45]$ Hz power. Regional means use available valid frontal $\{Fp1, Fp2, F3, F4, Fz\}$ and posterior $\{P3, P4, Pz, O1, O2\}$ channels, requiring at least two per region. For ordered measurements (u, v), $(v - u)/(u + v)$ defines negative, neutral, and positive states at boundaries -0.05 and 0.05, with equality assigned to neutral. A common reference-derived mask excludes targets with missing required channels, unusable intervals, nonfinite spectra, or nonpositive denominators across all methods and input conditions; missing auxiliary modalities do not exclude an otherwise valid EEG comparison. The common parser maps text to attribute–state pairs; absent or conflicting predictions do not recover a target. Withheld F1 averages three-state macro-F1 over available attributes (Eq. 37); expert review assesses equivalent wording by meaning without rescaling Withheld F1 into Expert F1.

Q.2 Window Sampling and Summary Generation

The audit uses 50 distinct test windows from each of Mumtaz 2016, Mental Arithmetic, and SHU–MI, covering 10, 4, and 5 subjects, respectively. These are audit-subset counts; the corresponding study test folds contain 11, 4, and 5 subjects. Mumtaz and Arithmetic windows last 5 s, and SHU–MI windows last 4 s. A subject's eyes-open/eyes-closed recordings or multiple sessions retain a single subject identifier and remain in the same data split. Here, one EEG window is one input sample; its generated summary is segmented into atomic claims for expert labeling.

Sampling is fixed before inspecting model outputs, stratified by the dataset's class labels, and distributed across eligible subjects and recordings. Arithmetic includes both baseline and arithmetic-task windows; SHU–MI includes left- and right-hand imagery; Mumtaz includes healthy-control and MDD recordings. Eligibility requires adequate signal quality and a nonempty independently verifiable reference fact set. The manifest records subject, recording or trial, start time, duration, class, and exclusion reason. Overlapping and near-duplicate windows are excluded, and one common eligible window set is used for every method.

NeuroLM, CELM, LoRA, Full FT, and Ewen each generate one original summary for every selected window: 250 summaries per dataset and 750 in total. NeuroLM and CELM use the narrative-generation configurations in Appendix E; LoRA, Full FT, and Ewen share the matched backbone and EEG interface. The instruction requests a concise physiological account supported by the available signal inputs, without listing reference facts or requiring a prescribed field-by-field answer. Checkpoints, input availability, decoding settings, and output-length limits are recorded before scoring. Generated outputs, including omissions and unsupported assertions, are preserved without correction.

Q.3 Independent Evidence and Expert Adjudication

Before viewing generated outputs, two EEG-trained experts reconcile a reference fact set $\mathcal{R}_i$ for each window. Each reference identifies a channel or region, time interval, quantity or relationship, and supporting waveform evidence or independent calculation. Its scope covers verifiable spectral, temporal, spatial, and available cross-signal findings. Dataset annotations determine task labels; generic physiological patterns alone do not establish a clinical diagnosis. References and matching rules are frozen before scoring and are unavailable to the generating models.

Each summary from method m is segmented into an atomic claim set $\mathcal{C}_{im}$. Segmentation preserves qualifiers, negation, location, and time scope; repeated paraphrases are deduplicated within that summary only. Claims from different methods remain separate evaluation items, even if they express the same fact. The common inventory is frozen before correctness labeling. Expert inspection checks claim extraction and supporting evidence, identifies omitted claims, and determines every final label.

Blinded ratings and reconciliation. Anonymous outputs are presented in independently randomized order. Each expert assigns every claim one of three labels—*supported*, *contradicted*, or *insufficient evidence*—without access to the other expert's rating or automated verdicts. Initial ratings are preserved. Experts then review disagreements against the waveform and independent calculations; unresolved claims receive *insufficient evidence*. Final metrics use one reconciled label per claim, rather than averaging two expert scores. Agreement statistics use only the initial, pre-discussion labels.

Matching rules. Evidence must support the stated location, interval, quantity, and direction. Explicit numerical claims use absolute tolerances of 0.005 for relative power, 0.05 for correlation or coherence, $0.5\,\mathrm{Hz}$ for frequency, and $2\,\mathrm{bpm}$ for heart rate. Other scalar quantities use 5% of the absolute reference value after unit conversion; a zero reference must match zero at the reference's reporting precision. Qualitative comparisons covered by Withheld F1 use Appendix Q.1. These are scoring tolerances rather than physiological significance thresholds. A claim about an unavailable modality or without reproducible evidence receives *insufficient evidence*. Plausible causal or diagnostic interpretations receive no credit without the evidence required by the annotation scope.

Q.4 Fact-Level Metrics and Statistical Reporting

Let S_{im} count supported generated claims and M_{im} count distinct reference facts covered by supported claims. One-to-one semantic matching preserves channel, time, quantity, and direction and prevents duplicate credit. Define

$$P_{im}^{\mathrm{exp}} = \frac{S_{im}}{|\mathcal{C}_{im}|}, \qquad R_{im}^{\mathrm{ref}} = \frac{M_{im}}{|\mathcal{R}_i|}, \qquad F_{im}^{\mathrm{exp}} = \frac{2P_{im}^{\mathrm{exp}} R_{im}^{\mathrm{ref}}}{P_{im}^{\mathrm{exp}} + R_{im}^{\mathrm{ref}}}. \tag{69}$$

An output with no factual claims receives zero precision, coverage, and F1; a zero harmonic-mean denominator also gives zero. Reference sets are nonempty by eligibility. Contradicted and insufficient-evidence claims remain in the precision denominator. A supported claim absent from the bounded reference set contributes to precision but not reference coverage. Expert F1 therefore measures evidence-supported precision and reference coverage, rather than exhaustive recall over all possible waveform findings.

The beyond-covariate count B_{im} records distinct supported claims requiring evidence not supplied or directly computable from the input covariates. It satisfies $0 \le B_{im} \le S_{im} \le |\mathcal{C}_{im}|$. For 50 windows, its mean lies on the 0.02 grid; the three-dataset mean lies on the $1/150$ grid before display rounding.

For each method and window metric x, Table 21 reports

$$\bar{x}_d = \frac{1}{n_d}\sum_{i=1}^{n_d} x_{di}, \qquad s_d = \sqrt{\frac{1}{n_d - 1}\sum_{i=1}^{n_d}(x_{di} - \bar{x}_d)^2}, \qquad n_d = 50. \tag{70}$$

Expert F1 is computed per window before averaging; it need not equal the harmonic mean of the displayed mean P and R. SD describes variation across windows for fixed checkpoints and reconciled labels. It does not measure variation across experts or training seeds, and two raters do not double the window count. The dataset-macro mean weights the three datasets equally. With 50 windows per dataset, the accompanying pooled sample SD is

$$\bar{x}_{\text{macro}} = \frac{1}{3}\sum_{d=1}^{3}\bar{x}_d, \qquad s_{\text{pool}} = \sqrt{\frac{\sum_{d=1}^{3}[49s_d^2 + 50(\bar{x}_d - \bar{x}_{\text{macro}})^2]}{149}}. \tag{71}$$

Table 21: Expert summary metrics on 50 test windows per dataset, paired across five methods. P is factual precision, R is reference coverage, and Expert F1 is computed per window using Eq. 69. Values are mean ± window SD; beyond-covariate facts are counts per window. Dataset-macro gives the equal-dataset mean and pooled 150-window SD.

Dataset	Method	P	R	Expert F1	Beyond-cov. facts
Mumtaz 2016	NeuroLM	0.61 ± 0.17	0.37 ± 0.11	0.45 ± 0.12	0.74 ± 0.66
	CELM	0.43 ± 0.16	0.28 ± 0.13	0.34 ± 0.14	0.64 ± 0.60
	LoRA	0.68 ± 0.15	0.41 ± 0.14	0.50 ± 0.14	0.84 ± 0.74
	Full FT	0.78 ± 0.19	0.45 ± 0.17	0.56 ± 0.17	0.96 ± 0.83
	Ewen	0.89 ± 0.17	0.57 ± 0.19	0.68 ± 0.18	1.06 ± 0.87
Arithmetic	NeuroLM	0.50 ± 0.12	0.31 ± 0.11	0.38 ± 0.11	0.68 ± 0.65
	CELM	0.33 ± 0.12	0.22 ± 0.11	0.26 ± 0.10	0.48 ± 0.50
	LoRA	0.57 ± 0.15	0.34 ± 0.13	0.42 ± 0.13	0.76 ± 0.72
	Full FT	0.72 ± 0.14	0.44 ± 0.13	0.53 ± 0.12	0.90 ± 0.76
	Ewen	0.84 ± 0.17	0.52 ± 0.18	0.63 ± 0.17	1.06 ± 0.71
SHU–MI	NeuroLM	0.52 ± 0.12	0.32 ± 0.15	0.38 ± 0.14	0.68 ± 0.68
	CELM	0.38 ± 0.16	0.24 ± 0.12	0.29 ± 0.13	0.50 ± 0.58
	LoRA	0.65 ± 0.14	0.42 ± 0.13	0.50 ± 0.12	0.98 ± 0.82
	Full FT	0.75 ± 0.19	0.46 ± 0.17	0.55 ± 0.17	1.06 ± 0.89
	Ewen	0.84 ± 0.18	0.55 ± 0.20	0.65 ± 0.19	1.24 ± 0.92
Dataset-macro	NeuroLM	0.54 ± 0.15	0.34 ± 0.13	0.41 ± 0.13	0.70 ± 0.66
	CELM	0.38 ± 0.15	0.25 ± 0.12	0.29 ± 0.13	0.54 ± 0.56
	LoRA	0.63 ± 0.15	0.39 ± 0.14	0.47 ± 0.14	0.86 ± 0.76
	Full FT	0.75 ± 0.18	0.45 ± 0.16	0.55 ± 0.16	0.97 ± 0.83
	Ewen	0.86 ± 0.17	0.54 ± 0.19	0.65 ± 0.18	1.12 ± 0.83

Independent-rating agreement. For dataset d, let n_{ab} be the number of claims assigned label a by expert A and label b by expert B, pooling all five methods. The paired-claim count and agreement are

$$n_{\text{pair}} = \sum_{a,b} n_{ab}, \quad p_o = \frac{\sum_a n_{aa}}{n_{\text{pair}}}, \quad p_e = \sum_a \frac{n_{a\cdot}}{n_{\text{pair}}}\frac{n_{\cdot a}}{n_{\text{pair}}}, \quad \kappa = \frac{p_o - p_e}{1 - p_e}. \tag{72}$$

Both agreement and disagreement contribute to n_{pair}: it counts claims rated by both experts, not EEG windows, matching labels, or individual rating actions. It is obtained from the claim inventory and is not fixed to a prescribed value. Complete dual review gives $n_{\text{pair}} = \sum_{i,m} |\mathcal{C}_{im}|$; any missing initial rating is excluded from agreement calculations and its count is reported. An empty paired set or $p_e = 1$ makes κ undefined. Table 22 reports raw agreement, chance agreement, and unweighted three-category κ. Agreement includes matching unsupported labels and therefore does not measure model factual accuracy.

Subject-clustered uncertainty. Let S_d denote the number of distinct audit subjects. We use $B = 1{,}000$ subject-cluster bootstrap replicates; B is the number of resampling computations, not

Table 22: Initial agreement between two experts. Windows and Subjects count distinct audit units. Agreement pools the five methods. n_{pair} counts jointly rated claims. The final column is the subject-cluster percentile interval for κ.

Dataset	Windows	Subjects	n_{pair}	p_o	p_e	κ	95% CI
Mumtaz 2016	50	10	756	0.78	0.45	0.60	[0.49, 0.68]
Arithmetic	50	4	775	0.78	0.40	0.64	[0.53, 0.71]
SHU–MI	50	5	763	0.77	0.42	0.60	[0.56, 0.64]

the audit sample size. Each replicate samples S_d subject identifiers with replacement within dataset d. Every draw retains that subject's entire set of windows, claims, and paired outputs from all five methods. A repeated subject contributes repeatedly; consequently, the replicate's window count can differ from 50 when subject clusters have different sizes. Recompute window-weighted metric means using the actual replicate denominator, and average the three dataset means with equal weights. Use the same subject draws for every method and for paired method differences. Initial labels are used to recompute κ; reconciled labels are used for P, R, Expert F1, and beyond-covariate counts.

The interval endpoints are the 2.5th and 97.5th percentiles of the replicate statistics, using linear quantile interpolation. All statistics are computed at full precision and rounded only for display; intervals are not forced to be symmetric. The valid κ replicate count is retained with the evaluation records. Its interval is reported only when κ is defined in all 1,000 replicates; otherwise, the undefined fraction is reported and the interval is left unreported under this protocol. Intervals condition on the fixed checkpoints and reconciled annotations, without resampling experts. They describe subject-sampling uncertainty and are distinct from the window SDs in Table 21.

With only four or five subjects, the bootstrap distribution has limited support and nominal coverage is uncertain. For four subjects sampled four times, there are only $\binom{7}{4} = 35$ distinct multiplicity patterns, with unequal probabilities; 1,000 replicates do not create additional independent subjects. The intervals in Table 23 must therefore be interpreted alongside the subject counts, without treating claims or windows as independent subjects. Fewer than two original subjects does not support a subject-sampling interval.

Table 23: Subject-cluster percentile 95% intervals for the window means in Table 21, from 1,000 paired resampling replicates. Dataset-macro is recomputed with equal dataset weights in every replicate.

Dataset	Method	P	R	Expert F1	Beyond-cov. facts
Mumtaz 2016	NeuroLM	[0.55, 0.68]	[0.34, 0.41]	[0.41, 0.49]	[0.60, 0.90]
	CELM	[0.38, 0.48]	[0.24, 0.33]	[0.29, 0.39]	[0.46, 0.80]
	LoRA	[0.64, 0.73]	[0.36, 0.47]	[0.46, 0.56]	[0.68, 1.02]
	Full FT	[0.71, 0.84]	[0.40, 0.48]	[0.51, 0.60]	[0.78, 1.12]
	Ewen	[0.84, 0.93]	[0.52, 0.62]	[0.63, 0.73]	[0.92, 1.20]
Arithmetic	NeuroLM	[0.46, 0.53]	[0.28, 0.35]	[0.35, 0.40]	[0.56, 0.80]
	CELM	[0.32, 0.35]	[0.20, 0.24]	[0.24, 0.27]	[0.44, 0.52]
	LoRA	[0.52, 0.63]	[0.31, 0.38]	[0.39, 0.45]	[0.68, 0.84]
	Full FT	[0.67, 0.77]	[0.40, 0.49]	[0.50, 0.57]	[0.80, 1.00]
	Ewen	[0.79, 0.90]	[0.49, 0.55]	[0.59, 0.66]	[0.88, 1.24]
SHU–MI	NeuroLM	[0.48, 0.55]	[0.28, 0.35]	[0.36, 0.41]	[0.60, 0.76]
	CELM	[0.33, 0.42]	[0.21, 0.27]	[0.25, 0.32]	[0.38, 0.62]
	LoRA	[0.61, 0.69]	[0.41, 0.44]	[0.49, 0.52]	[0.82, 1.14]
	Full FT	[0.68, 0.81]	[0.44, 0.47]	[0.52, 0.58]	[0.92, 1.20]
	Ewen	[0.81, 0.88]	[0.50, 0.59]	[0.61, 0.69]	[1.08, 1.48]
Dataset-macro	NeuroLM	[0.52, 0.57]	[0.32, 0.35]	[0.39, 0.42]	[0.63, 0.77]
	CELM	[0.36, 0.40]	[0.23, 0.27]	[0.27, 0.31]	[0.47, 0.61]
	LoRA	[0.61, 0.66]	[0.37, 0.42]	[0.45, 0.50]	[0.78, 0.95]
	Full FT	[0.71, 0.78]	[0.42, 0.47]	[0.52, 0.57]	[0.89, 1.05]
	Ewen	[0.83, 0.89]	[0.52, 0.57]	[0.63, 0.68]	[1.02, 1.23]

Scope and verification records. Withheld F1 evaluates predefined waveform comparisons; Expert F1 evaluates generated-claim precision and reference coverage; κ evaluates initial annotator

agreement. These quantities have different denominators and cannot be converted into one another by rescaling. Their interpretation is limited to the sampled test subjects, annotation scope, and fixed checkpoints. The verification record links each window and summary to its references, deduplicated claims, two initial labels, reconciled labels, match decisions, metric contributions, and subject-resampling indices. Access to underlying recordings and annotations follows the applicable dataset and privacy conditions.