TEACHING NEURAL DECODERS WITH LANGUAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain–computer interfaces decode neural activity, yet class-label supervision (CLS) leaves relationships among classes implicit. We propose Language as Supervision (LaS), which aligns neural features with natural-language class descriptions through learned projections and predicts classes using candidate text prototypes. We evaluate seven encoder architectures with independently trained models for visual, motor, and speech decoding from scalp and intracranial recordings. LaS improves speech-decoding accuracy by up to 17.15 pp over matched CLS baselines and supports subject and session adaptation. Frequency-structured neural–text correspondence further enables conditional decoding of unseen stimulus frequencies in steady-state visual evoked potential (SSVEP) responses. Dual-sided attribution visualization shows contributions from neural signal components and prompt words to the same class-specific matching score. These findings suggest that text descriptions can serve as general-purpose references for learning and interpretation in neural decoding.

1 INTRODUCTION

Brain–computer interfaces (BCIs) use representations of scalp electroencephalography (EEG) or intracranial recordings for communication and control, such as finger-level robotic hand control (Ding et al., 2025) and speech neuroprostheses (Willett et al., 2023).

Classical decoding combines engineered features with classifiers or template matching (Ramoser et al., 2000; Tangermann et al., 2012; Nakanishi et al., 2018). Neural decoders learn signal representations jointly with task-specific classifiers (Lawhern et al., 2018; Schirrmeister et al., 2017). Under *class-label supervision* (CLS), one-hot targets specify class identities without explicitly encoding class attributes or semantic relationships.

Natural-language class descriptions provide semantic information that pretrained text models can encode. EmotionCLIP uses templated descriptions for emotion recognition (Yan et al., 2026), while EEG-PRIME combines class prototypes with task instructions and multilevel conditioning (Zhang

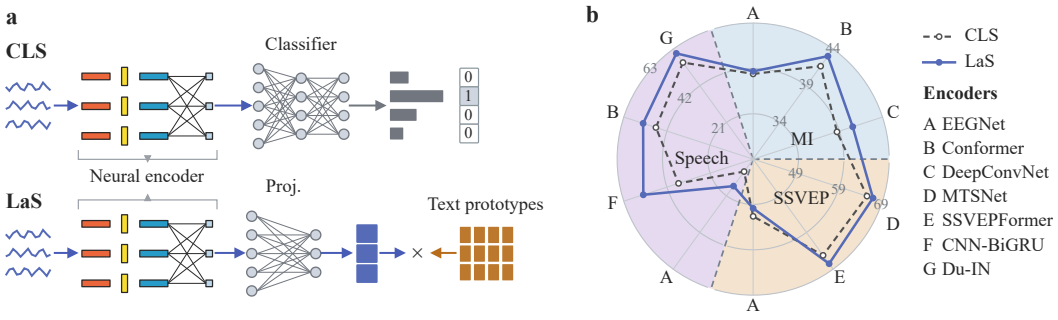


Figure 1: **Language supervision across neural decoders. (a)** CLS uses a task-specific classifier; LaS matches projected neural features to text prototypes. **(b)** Mean accuracy across ten task–encoder settings in MI, SSVEP, and speech decoding. Task sectors use different accuracy scales; radii and polygon areas are not comparable across tasks. Conformer denotes EEG-Conformer. Full comparisons appear in Table 1.

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.

et al., 2026). We study class descriptions as a supervision and prediction interface through neural–text alignment, using paired CLS comparisons that retain established encoder architectures. We ask how this alignment affects decoding performance relative to CLS, whether it supports unseen classes under compatible neural–text structure, and which neural components and words support a specified match.

We introduce *Language as Supervision* (LaS; Figure 1) with three contributions. (1) LaS uses class descriptions as references for learning and candidate prediction, aligning neural features with text embeddings through learnable projections. (2) Paired evaluations across seven encoders show decoding gains in motor imagery (MI), steady-state visual evoked potential (SSVEP), and speech decoding from scalp EEG and intracranial stereo-EEG (sEEG). LaS supports candidate configuration through class descriptions, subject/session adaptation, and conditional unseen-frequency prediction. (3) Theory and controlled language experiments characterize alignment behavior. We propose a dual-sided attribution visualization scheme that shows neural signal components and prompt words contributing to the same class-matching score.

2 Related Work

2.1 Semantic Representations for Classification

Semantic class representations link supervision to candidate prediction. Semantic output codes use class attributes as learning targets, including unseen-word fMRI decoding (Palatucci et al., 2009). DeViSE links visual features to text-derived class-name representations for unseen-class prediction (Frome et al., 2013); Ba et al. (2015) generate classifier weights from descriptions for classes without training images. CLIP combines large-scale contrastive image–text learning with prediction by candidate-description similarity (Radford et al., 2021), demonstrating competitive zero-shot recognition, frozen-feature transfer through linear probes, and improved robustness over ImageNet supervision on the natural shifts studied. Its dataset transfer and instance-level pairing differ from our class holdouts and repeated descriptions. BLIP-2 connects a frozen image encoder and frozen large language model through a Q-Former (Li et al., 2023); CLAP extends contrastive alignment and candidate matching to audio for zero-shot classification (Elizalde et al., 2023).

2.2 Language-Aligned Neural Decoding

Neural–text alignment methods differ in their sources of textual supervision, the units paired during training, and how text is used at prediction. BrainCLIP maps fMRI representations into a visual–language space using visual and text supervision (Ma et al., 2025). Clinical EEG studies use paired reports: Gijsen & Ritter (2025) combine EEG crops, report segments, and multiple-instance learning for label-efficient phenotyping. Their shuffled-report control changes patient–report pairings, whereas ours reassigns fixed class descriptions. EEG-CLIP learns from EEG–report alignment, uses class text prototypes for zero-shot prediction, and visualizes gradients of EEG–text similarity with respect to frequency-domain inputs (Ndir et al., 2025). Class descriptions also support direct EEG recognition: EmotionCLIP inserts emotion names into 16 templates and classifies by EEG–text matching (Yan et al., 2026). The EEG-PRIME preprint combines class-name prototypes, task instructions, and multilevel conditioning across EEG tasks, and compares text encoders (Zhang et al., 2026); some downstream fine-tuning settings use task-specific classification heads. LaS builds on prototype classification and neural–text alignment, examining class descriptions as supervision and candidate references across independently trained decoders.

3 Methods

3.1 Problem Setting and Class-Label Supervision

For training classes $\mathcal{S} = \{1, \ldots, C\}$, $C \geq 2$, trial X_i labeled $y_i \in \mathcal{S}$ yields $N_f(X_i) = \text{neural_encoder}_\theta(X_i) \in \mathbb{R}^{1 \times d_f}$. Class-label supervision (CLS) jointly trains the encoder and its original, possibly nonlinear head by cross-entropy: $Z_{\text{CLS}} = g_\psi(N_f(X_i))$, $p_i^{\text{CLS}} = \text{softmax}(Z_{\text{CLS}})$, $\ell_{\text{CLS}} = -\log p_{i,y_i}^{\text{CLS}}$. LaS pairs labeled trials with class descriptions.

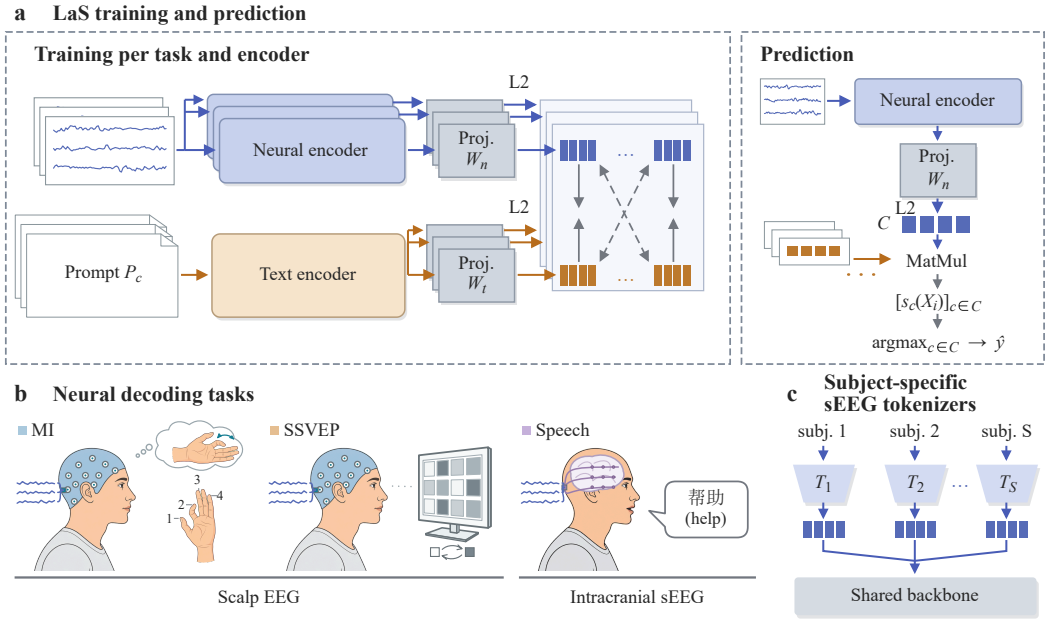


Figure 2: **The overall architecture of LaS. (a)** Prompt P_c is the complete class description supplied to the text encoder. Each task–encoder model trains its neural encoder and projections W_n and W_t for neural–text alignment, then predicts the highest-scoring candidate. **(b)** The evaluated tasks comprise finger MI and SSVEP decoding from scalp EEG, and overt Chinese word reading from intracranial sEEG. Speech class descriptions use English translations. **(c)** Neural input representations follow the selected encoder. Subject-specific sEEG tokenizers feed a backbone shared within the same model.

3.2 Language as Supervision

Fixed prompts P_c (e.g., "A neural SSVEP signal at 10.6 Hz.") yield embeddings $T_{f,c} = \text{text_encoder}(P_c) \in \mathbb{R}^{1\times d_t}$ through a frozen, deterministic text encoder and its prescribed readout. Bias-free projections $W_n \in \mathbb{R}^{d_f\times d_{\text{proj}}}$ and $W_t \in \mathbb{R}^{d_t\times d_{\text{proj}}}$ give

$$z^{(i)}_{\text{n,proj}} = \nu(N_f(X_i)W_n), \qquad z^{(c)}_{\text{t,proj}} = \nu(T_{f,c}W_t),$$
$$\nu(v) = \frac{v}{\|v\|_2}, \qquad s_c(X_i) = z^{(i)}_{\text{n,proj}}\big(z^{(c)}_{\text{t,proj}}\big)^\top. \tag{1}$$

$z^{(c)}_{\text{t,proj}}$ is the text *prototype* and s_c the unscaled cosine score. With the text encoder frozen, training jointly updates W_t and the neural branch, so text prototypes evolve. LaS can thus be viewed as tying classifier-weight geometry to a semantic subspace anchored by text embeddings through the learned projection W_t. A learned inverse temperature $\tau^{-1} = e^t > 0$ gives $a_{ic} = s_c(X_i)/\tau$. Stack normalized vectors for B pairs (X_i, P_{y_i}) into $Z_{\text{n,proj}}, Z_{\text{t,proj}}$, with $L = Z_{\text{n,proj}}Z^\top_{\text{t,proj}}/\tau \in \mathbb{R}^{B\times B}$ and $L_{ij} = a_{i,y_j}$. B counts pairs in the loss matrix, irrespective of gradient accumulation. We train with bidirectional InfoNCE (van den Oord et al., 2018; Radford et al., 2021):

$$\mathcal{L}_{\text{N2T}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{L_{ii}}}{\sum_{j=1}^{B}e^{L_{ij}}}, \qquad \mathcal{L}_{\text{T2N}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{L_{ii}}}{\sum_{j=1}^{B}e^{L_{ji}}}, \tag{2}$$
$$\mathcal{L}_{\text{SA}} = \tfrac{1}{2}(\mathcal{L}_{\text{N2T}} + \mathcal{L}_{\text{T2N}}).$$

N2T queries texts with neural signals; T2N queries trials with text. Both retain all B entries and diagonal targets, introducing same-class off-diagonal competitors.

Each task–encoder model learns its own θ, W_n, W_t, t and alignment space, sharing its backbone only within that model (Figure 2c). Text embeddings can be precomputed; caching final prototypes after training avoids repeated text encoding at inference. For candidate set $\mathcal{C}$, predict $\widehat{y}_i = \arg\max_{c\in\mathcal{C}} s_c(X_i)$.

3.3 Theoretical Properties

Directional asymmetry. For fixed prompts and deterministic text forwards, let $\mathcal{G}_c = \{i : y_i = c\}$, $n_c = |\mathcal{G}_c|$, $\mathcal{S}_b = \{c : n_c > 0\}$ and $\bar{a}_c = n_c^{-1} \sum_{i \in \mathcal{G}_c} a_{ic}$. With m identical texts per class, the diagonal target has probability at most $1/m$, giving N2T a $\log m$ loss floor. The N2T identity assumes all C classes occur m times; T2N allows unequal counts:

$$\begin{aligned} \ell_i^{\mathrm{N2T}} &= \log m - a_{i,y_i} + \log \sum_{c=1}^{C} e^{a_{ic}}, \qquad B = mC, \\ \bar{\ell}_c^{\mathrm{T2N}} &= -\bar{a}_c + \log \sum_{j=1}^{B} e^{a_{jc}}, \qquad \mathcal{L}_{\mathrm{T2N}} = \frac{1}{B} \sum_{c \in \mathcal{S}_b} n_c \bar{\ell}_c^{\mathrm{T2N}}. \end{aligned} \tag{3}$$

Balanced N2T is class cross-entropy plus a constant; unequal counts reweight present classes. For $n_c > 1$, T2N provides implicit regularization toward within-class aggregation of matching responses along the corresponding text-prototype direction. With the mean within-class score and all other-class scores against that prototype fixed, equal within-class matching scores minimize its class loss (Appendix C.1). Other feature directions may still differ.

Repeated-prompt loss floors. Diagonal and Class-aware SigLIP ablations (Zhai et al., 2023) use shared learned bias $b_{\mathrm{sig}} \in \mathbb{R}$, $\xi_{ic} = a_{ic} + b_{\mathrm{sig}}$, $\ell_{\pm}(\xi) = \log(1 + e^{\mp \xi})$ and binary labels $M \in \{0,1\}^{B \times B}$:

$$\begin{aligned} \mathcal{L}_{\mathrm{Sig}}(M) &= \frac{1}{B} \sum_{i,j=1}^{B} \left[M_{ij} \ell_+(\xi_{i,y_j}) + (1 - M_{ij}) \ell_-(\xi_{i,y_j}) \right], \\ M_{ij}^{\mathrm{diag}} &= \mathbf{1}[i = j], \qquad M_{ij}^{\mathrm{class}} = \mathbf{1}[y_i = y_j]. \end{aligned} \tag{4}$$

Both sum B^2 terms and divide by B. Deterministic class-text copies imply these bounds; for $B = mC$, the last row lists balanced values for InfoNCE, Diagonal and Class-aware SigLIP, respectively:

$$\begin{aligned} \mathcal{L}_{\mathrm{SA}} &\geq \frac{1}{B} \sum_{c \in \mathcal{S}_b} n_c \log n_c, \qquad \mathcal{L}_{\mathrm{Sig}}^{\mathrm{class}} \geq 0, \\ \mathcal{L}_{\mathrm{Sig}}^{\mathrm{diag}} &\geq \frac{1}{B} \sum_{c \in \mathcal{S}_b} n_c \left[n_c \log n_c - (n_c - 1) \log(n_c - 1) \right], \\ n_c = m: \quad & \log m, \qquad m \log m - (m-1) \log(m-1), \qquad 0. \end{aligned} \tag{5}$$

Logs are natural, with $0 \log 0 = 0$; attainability depends on separation and realizable logits (Appendix C.2).

4 Experimental Evaluation

4.1 Evaluation Setup

We train separate models on each dataset's pooled training subjects. **MI** contains 21 subjects imagining thumb, index, middle, or pinky movements (Ding et al., 2025). Five within-subject physical-trial folds keep each trial's windows together; decisions aggregate 1-s inputs over 5-s trials. Following the released implementation of Ding et al. (2025), checkpoints maximize validation-window accuracy. In our five-fold evaluation, the same held-out fold serves checkpoint selection and reporting of trial-level accuracy. **SSVEP** contains 35 subjects and 40 stimulus classes (Wang et al., 2017), evaluated in seven folds of 25/5/5 training/validation/test subjects; checkpoints are selected on validation subjects. **Speech** contains 12 subjects overtly reading 61 Chinese words with English prompt translations (Zheng et al., 2024), 3-s sEEG inputs, and a shared physical-trial split of approximately 80%/10%/10% for training/validation/testing.

We evaluate EEGNet (Lawhern et al., 2018), DeepConvNet (Schirrmeister et al., 2017), EEG-Conformer (Song et al., 2023), SSVEPFormer (Chen et al., 2023), MTSNet (Lan et al., 2025), CNN-BiGRU (Moses et al., 2021), and Du-IN (Zheng et al., 2024) as neural encoders, with FBCSP-LDA (Ang et al., 2008) as an MI reference.

Table 1: **Decoding performance across tasks and neural encoders.** CLS–LaS comparisons; FBCSP-LDA is a separate MI reference. Performance: subject-mean accuracy (%) ± sample SD; mean MI κ in parentheses. Δ = LaS − CLS in percentage points. Blue/gray: highest/second-highest means within each MI or Speech metric across CLS and LaS. [1]SSVEP: RW training → IW evaluation. Stars: Holm-corrected paired t-tests within encoder (Section 4.1; Appendix B.6).

Task	Method/encoder	Params (M)		Duration (s)	Performance		Δ Acc. (pp)
		CLS	LaS		CLS	LaS	
MI	FBCSP-LDA	—	—	5.0	32.47±5.94 (κ=0.100)		—
	EEGNet	0.002	0.30		38.39±4.76 (κ=0.179)	38.68±7.08 (κ=0.182)	+0.28
	DeepConvNet	0.29	0.79		38.72±6.05 (κ=0.183)	40.54±8.65 (κ=0.207)	+1.82*
	Conformer	0.29	0.48		41.67±8.77 (κ=0.222)	43.02±9.29 (κ=0.240)	+1.36**
SSVEP RW→IW[1]	EEGNet	0.04	0.32	0.6	26.76±11.55	26.98±10.91	+0.21
		0.05	0.37	0.8	41.07±16.34	40.42±15.82	-0.65
		0.05	0.39	1.0	51.62±18.32	49.86±16.43	-1.76
	MTSNet	9.92	7.37	0.6	34.32±15.23	36.56±15.27	+2.24**
		10.87	8.12	0.8	53.27±20.52	55.17±19.93	+1.89
		11.89	8.95	1.0	65.39±19.49	66.79±19.83	+1.39
	SSVEPFormer	3.09	1.11	0.6	33.46±15.19	35.94±15.35	+2.48*
		3.09	1.11	0.8	52.30±19.89	54.51±19.17	+2.21*
		3.09	1.11	1.0	65.23±20.73	67.50±19.10	+2.27*
Speech	EEGNet	0.02	0.13	3.0	7.13±2.51	15.46±6.39	+8.33***
	CNN-BiGRU	0.62	0.84		36.33±9.26	53.47±13.43	+17.15***
	Conformer	4.18	3.96		47.34±11.73	53.56±15.04	+6.21***
	Du-IN	5.05	4.94		55.27±13.17	60.49±14.30	+5.22***

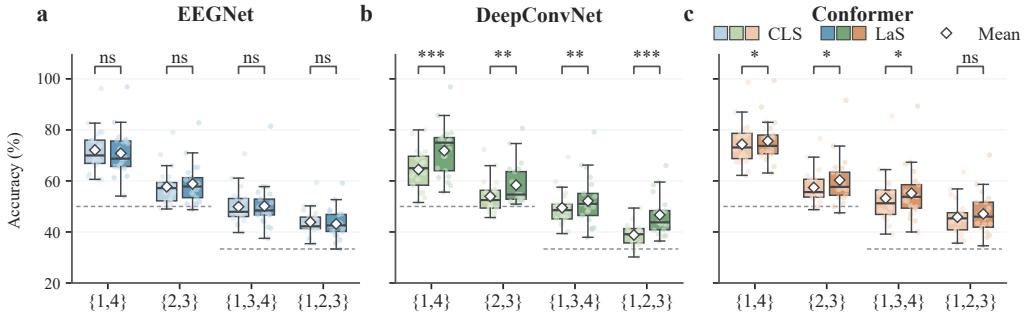


Figure 3: **Candidate-subset reuse in motor imagery.** CLS trains per subtask; LaS restricts candidates in each fold's four-class checkpoint without further optimization. Class IDs 1, 2, 3, 4 denote Thumb, Index, Middle, Pinky. Boxes: medians and interquartile ranges (IQRs); whiskers: 1.5×IQR rule; circles: subject accuracies (n = 21); diamonds: means; dashed lines: chance (50%/33.3%). Stars: Holm-corrected paired CLS–LaS comparisons (Section 4.1).

CLS–LaS pairs share encoder architectures and splits, training to convergence with AdamW and their respective output modules and settings. Training uses PyTorch on a single NVIDIA RTX 4090 GPU. For Table 1, LaS uses frozen EmbeddingGemma (300M) (Schechter Vera et al., 2025), one fixed correct prompt per class (P0), and bidirectional InfoNCE. Parameter counts include all trainable modules and both projections, excluding the frozen text encoder; Duration includes the decision's full signal span and aggregation (Appendix A).

Windowing and data augmentation. For SSVEP, initial-window (IW) sampling fixes the start at latency-corrected onset; random-window (RW) augmentation uniformly samples valid starts, expanding the training set to 300,000 windows. Inter-trial recombination (IR) combines segments from same-class training trials. Neural training zero-masks 30% of input time (ratio 0.3) (Ding et al., 2024) for Table 1 (both CLS and LaS), Table B.1, and the seen classes in Figure 4c. SSVEP augmentation details appear in Appendix D.2. All four Speech encoders use Du-IN's downstream augmentation (Zheng et al., 2024) during CLS/LaS training: random left/right shifts of 0–0.3 s with zero padding preserve the 3-s input span. Validation and testing use unshifted inputs.

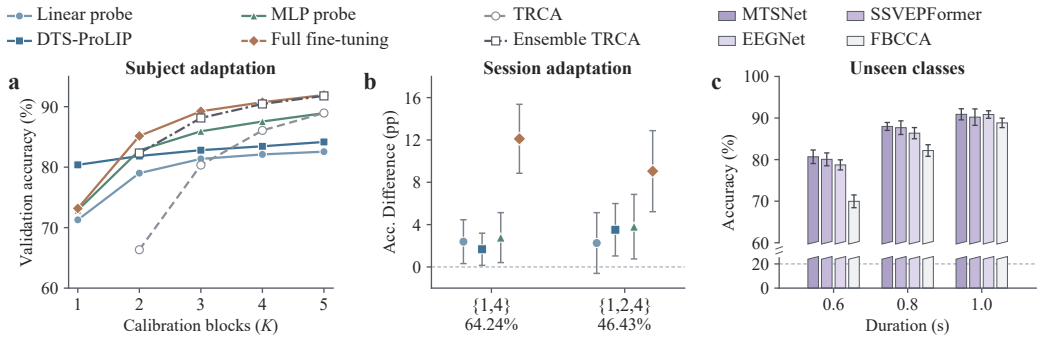


Figure 4: **Subject/session adaptation and conditional unseen-class decoding. (a)** SSVEP subject adaptation: mean validation accuracy with K target-subject calibration blocks (1-s inputs; six leave-one-block-out rounds). **(b)** MI session adaptation: adapted-model accuracy on eight later ("Fine") test runs minus direct-inference accuracy on eight Base runs per subject (mean paired differences across 21 subjects; 95% t intervals). **(c)** SSVEP decoding restricted to five held-out frequency classes, with the same 35 subjects used for training and evaluation (mean $\pm$ SD across eight class splits; 20% chance). The broken axis omits 25–60%; its two segments use different scales.

Table 1 averages subjects equally after pooling MI's five validation-fold predictions or averaging Speech's five seeds per subject. Table 1 and Figures 3, 5, and 8 use two-sided paired t-tests with Holm correction. Appendix B.6 defines the comparison families and reports paired Cohen's d_z for Table 1; $*$/$**$/$***$ denote $p < .05$/.01/.001; ns means nonsignificant.

4.2 Decoding Performance

Motor imagery. LaS (EEG-Conformer) leads evaluated four-class accuracy at $43.02\% \pm 9.29\%$, with mean $\kappa = 0.240$ and a 1.36 pp CLS-relative gain ($p < .01$; Table 1). DeepConvNet gains 1.82 pp ($p < .05$); EEGNet differs nonsignificantly.

By changing only the candidate prompt set, LaS reuses its four-class decoders without further optimization and significantly outperforms subtask-trained CLS on all four seen-class subsets with DeepConvNet and on three with EEG-Conformer (Figure 3). Differences are nonsignificant for all four EEGNet subsets and EEG-Conformer's $\{\mathrm{Thumb}, \mathrm{Index}, \mathrm{Middle}\}$ subset.

SSVEP. With RW training and IW validation and testing, LaS significantly improves MTSNet at 0.6 s and SSVEPFormer at all three durations after Holm correction; EEGNet differences are nonsignificant and negative at 0.8/1.0 s (Table 1). Separately, Table B.1 uses IW training for neural baselines and IW+IR for LaS (MTSNet), all evaluated on IW: LaS leads in accuracy at 0.7–1.0 s and estimated ITR at all five durations; SSVEPFormer leads in accuracy at 0.6 s.

Speech. All four encoders show significant paired LaS gains ($p < .001$), exceeding the modest MI/SSVEP improvements. With its neural encoder trained from scratch, LaS (Du-IN) leads at $60.49\% \pm 14.30\%$ (+5.22 pp over CLS). This accuracy exceeds the 59.18% reported for the multi-subject pretrained Du-IN (poms) variant (Zheng et al., 2024). CNN-BiGRU achieves the largest gain over CLS (17.15 pp). Gains occur with more or fewer trainable parameters (Table 1).

4.3 Adaptation and Conditional Unseen-Class Decoding

The preceding reuse results concern configuration of seen candidate sets; we next examine adaptation to changes in neural inputs and extension beyond training classes. Linear and MLP probes fit heads to frozen features; full fine-tuning updates the neural branch. Based on ProLIP (Fahes et al., 2026), DTS-ProLIP restricts neural projection updates to the span of the current candidates' centered text prototypes, keeping the encoder and source text prototypes fixed (Appendix A.6).

MI session adaptation. Figure 4b replays a subsequent session using the pooled four-class source model. Direct inference uses each subject's eight Base runs. Models adapted on these runs (80/20 physical-trial training/validation split) are evaluated on eight later Fine runs. Full fine-tuning gives

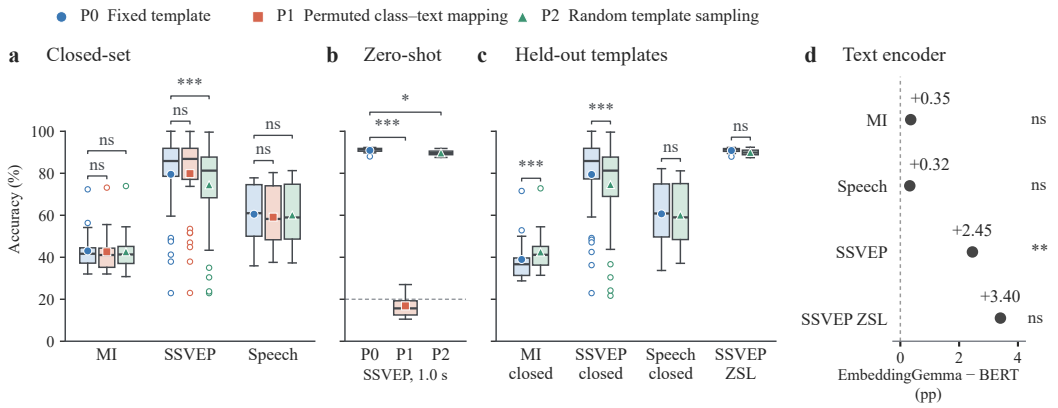


Figure 5: **Training prompt configurations and inference settings. (a)** Closed-set decoding. **(b)** Five-class SSVEP unseen-only decoding with one 1-s initial window; dashed line: 20% chance. Panels (a,b) use fixed prompts at inference. **(c)** P0 and P2 are evaluated with an ensemble of five templates held out from both training configurations. Units in (a–c): subjects (21 MI, 35 SSVEP, 12 Speech) or eight ZSL class splits. Boxes: medians and IQRs; whiskers follow the $1.5 \times$ IQR rule; hollow circles: outliers; colored symbols: means. Brackets: paired comparisons with P0. **(d)** EmbeddingGemma−BERT accuracy differences (pp); best configurations are selected independently of the Frozen–LoRA pairs (Figure B.1). Holm-corrected paired tests (Section 4.1).

the largest mean gains on Thumb/Pinky and Thumb/Index/Pinky, from direct-inference accuracies of $64.24\%/46.43\%$. The three-class linear probe's 95% paired-difference t interval crosses zero.

SSVEP subject adaptation. Figure 4a compares LaS (MTSNet) adaptations with TRCA and Ensemble TRCA (Nakanishi et al., 2018) at 1 s and matched calibration budgets. Six leave-one-block-out rounds adapt on K remaining blocks, each containing one trial per class; TRCA variants start at $K = 2$. Reported validation accuracy also selects MLP/full-fine-tuning epochs. DTS-ProLIP leads the displayed means at $K = 1$, full fine-tuning at $K = 2$–5; Ensemble TRCA approaches the latter with increasing K.

Conditional unseen-class decoding. Conditional zero-shot learning (ZSL; Figure 4c) uses the same 35 subjects for training and evaluation across eight class splits, each reserving all EEG from five unseen frequency classes for testing; 35 seen classes supply five training blocks and one validation block per subject, with the latter selecting checkpoints. After RW training, IW evaluation scores all 40 candidates in seen-class validation and only five unseen candidates in testing (20% chance). All three LaS encoders outperform FBCCA numerically at 0.6/0.8/1.0 s, especially shorter durations. Stimulus frequency supplies structured neural–text correspondence in this setting.

5 Analysis and Discussion

5.1 Class Correspondence and Language Configuration

Figure 5a–c compares training prompt configurations: fixed correct prompts (P0), a fixed derangement of seen-class descriptions (P1; $\pi(c) \neq c$ for every seen class), and random sampling of five correct training templates (P2). Panels (a,b) use one fixed prompt per class at inference; P2 uses P0 prompts. P1 maps closed-set predictions through the inverse permutation and uses true unseen descriptions for unseen-class prediction. P1 averages five derangement environments within subject or class split; Speech P2 averages three training seeds per subject.

Closed-set P0/P1 differences are nonsignificant, whereas P1 reduces unseen-only SSVEP accuracy to near chance ($p < .001$). P1 reassigns an unchanged raw text embedding set, testing class–description correspondence while preserving embedding relationships. With full-row-rank seen text embeddings and defined prototypes, a text projection can restore all seen prototypes while the neural branch remains fixed (Appendix C.1).

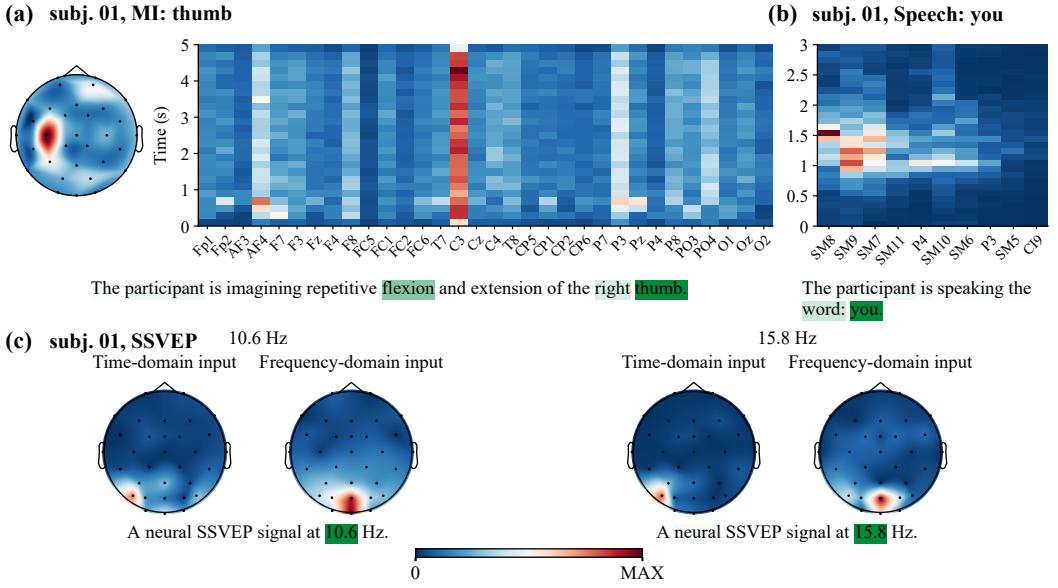


Figure 6: **Dual-sided attribution visualization with Integrated Gradients.** Training-trial averages: **(a)** MI, **(b)** Speech, and **(c)** SSVEP. Blue denotes low nonnegative neural display values, not negative attribution; green denotes positive mean word contributions. Separate panel scales preclude absolute comparisons.

With fixed P0 prompts at inference, P2 worsens SSVEP closed-set ($p < .001$) and unseen-only decoding ($p < .05$), with nonsignificant MI/Speech differences. Panel (c) compares P0 and P2 using an ensemble of five held-out templates, none used in either training configuration. P2 improves MI and worsens closed-set SSVEP (both $p < .001$); Speech and unseen-only SSVEP differences are nonsignificant.

Figure 5d compares independently selected best configurations of BERT (Devlin et al., 2019) and EmbeddingGemma. EmbeddingGemma improves closed-set SSVEP by 2.45 pp ($p = .00485$); its 3.40 pp unseen-only gain is nonsignificant ($p = .0838$), as are the MI/Speech differences. Text-encoder freezing versus LoRA adaptation (Hu et al., 2022) has task- and encoder-dependent effects (Figure B.1; Appendix B.3).

5.2 Interpreting Neural–Text Matches

We propose a dual-sided attribution visualization scheme for LaS class matching (Figure 6). Each example averages attributions over all training trials of one class from the first subject in its dataset: 120 right-thumb trials for MI, 38 "you" trials for Speech, and six original recordings for each SSVEP class (10.6 or 15.8 Hz). Grouped Integrated Gradients (Sundararajan et al., 2017) quantify signed contributions of input groups (sets of input components) to changes in the same unscaled class-specific cosine score along specified baseline-to-input paths. Neural attribution fixes class text; conditional word attribution fixes the neural representation and other words (Appendices B.4 and C.3). For display, ReLU-rectified neural attributions are divided by each complete input's maximum positive attribution before aggregation. This input is one MI window, one Speech trial, or both SSVEP branches jointly. Word attributions retain their signs during averaging; positive means are then divided by their maximum within each prompt for display.

Right-thumb MI shows a hotspot near C3 on the left central scalp over the sensorimotor region, with sustained C3 contributions and emphasis on "thumb" and "flexion." An early AF4 contribution may reflect ocular or muscle artifacts. In Speech, earlier contributions on SM9 and SM7 precede a stronger peak on SM8, followed by weaker contributions toward the trial's end; "you" is the strongest word contribution, followed by "word." SSVEP time-domain attributions emphasize the left occipital scalp, and frequency-domain attributions the occipital midline; word attributions emphasize "10.6" and "15.8."

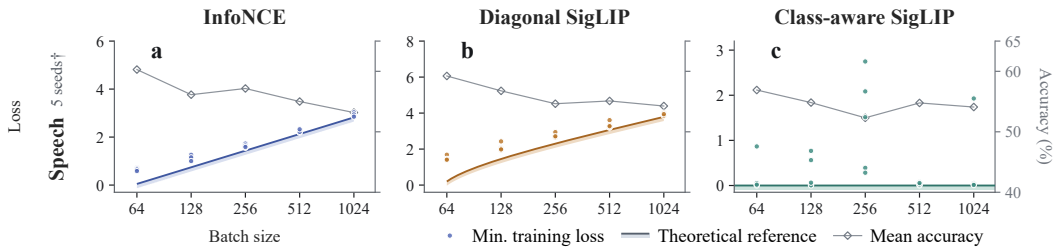


Figure 7: **Speech training objectives across batch sizes.** Du-IN with **(a)** InfoNCE, **(b)** Diagonal SigLIP, and **(c)** Class-aware SigLIP. Colored dots: each seed's minimum epoch-average full training loss (left axis); gray diamonds: mean accuracy (right axis). Lines show the balanced-class lower bounds in Equation (5), extended continuously to $m = B/61$. Pale bands and upward point offsets are display aids, not uncertainty. †At $B = 1024$ in (c), four loss observations are shown; accuracy includes all five seeds. The full three-task comparison appears in Figure B.6.

5.3 Training Objectives and Directional Effects

For each Speech condition, accuracy is averaged over five seeds per subject and then across the 12 subjects (batch grids: Appendix B.5). All observed minimum training losses in Figure 7 lie above the corresponding curves derived from Equation (5). Across the three tasks (Figure B.6), larger batches generally raise InfoNCE and Diagonal SigLIP minima without corresponding accuracy gains. Identical same-class descriptions create competition in InfoNCE and label conflicts in Diagonal SigLIP. Class-aware labels resolve these conflicts and increase positive-pair weights, but eliminating the positive loss floor does not consistently improve decoding accuracy (Section 3.3; Appendix C.2).

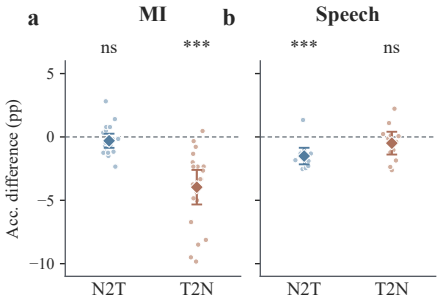


Figure 8: **Directional InfoNCE ablation.** Single-direction minus bidirectional accuracy (pp): **(a)** MI, $n = 21$; **(b)** Speech, $n = 12$. Dots: subject differences; diamonds: means; bars: unadjusted 95% paired-difference t intervals. Stars use Holm-corrected tests (Section 4.1).

Figure 8 compares MI (EEG-Conformer) and Speech (Du-IN), with five seeds per Speech condition. N2T promotes class discrimination, while T2N encourages within-class aggregation of matching responses along the corresponding text-prototype direction under the conditions in Section 3.3. Bidirectional InfoNCE outperforms T2N-only in MI and N2T-only in Speech (both $p < .001$); other contrasts are nonsignificant.

5.4 Scope and Limitations

Models are trained independently for each dataset and encoder; transfer across datasets or devices remains untested. Unseen-class evidence is limited to SSVEP's frequency-structured responses and descriptions. Initial nonstationarity in SSVEP responses limits transfer from RW training to IW evaluation in Table 1; Appendix D examines this effect in detail. Online interaction and long-term use remain untested.

6 Conclusion

LaS yields larger Speech gains and modest MI/SSVEP benefits, while allowing candidate sets to be configured through class descriptions. LaS also supports subject and session adaptation and conditional unseen-frequency decoding. The analysis establishes loss bounds for repeated class descriptions; deranging class–description assignments reduces unseen-frequency accuracy to near chance. Dual-sided attribution visualization shows contributions from neural signal components and prompt words to the same class-specific matching score.

Reproducibility Statement. The main text and appendices describe the data preprocessing, model architectures, training settings, and evaluation protocols. We will release the implementation code and detailed training configurations upon acceptance.

AI USE STATEMENT

We used DeepSeek and ChatGPT (OpenAI) to identify relevant literature, draft selected parts of the manuscript, and assist with writing and language editing. We reviewed the AI-assisted content and take responsibility for the final manuscript.

REFERENCES

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 2391–2398, 2008. doi: 10.1109/IJCNN.2008.4634130. URL https://doi.org/10.1109/IJCNN.2008.4634130.

Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4247–4255, 2015. doi: 10.1109/ICCV.2015.483. URL https://arxiv.org/abs/1506.00511.

Jianbo Chen, Yangsong Zhang, Yudong Pan, Peng Xu, and Cuntai Guan. A transformer-based deep neural network model for SSVEP classification. *Neural Networks*, 164:521–534, 2023. doi: 10.1016/j.neunet.2023.04.045. URL https://doi.org/10.1016/j.neunet.2023.04.045.

Xiaogang Chen, Yijun Wang, Shangkai Gao, Tzyy-Ping Jung, and Xiaorong Gao. Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface. *Journal of Neural Engineering*, 12(4):046008, 2015. doi: 10.1088/1741-2560/12/4/046008. URL https://doi.org/10.1088/1741-2560/12/4/046008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Wenlong Ding, Aiping Liu, Ling Guan, and Xun Chen. A novel data augmentation approach using mask encoding for deep learning-based asynchronous SSVEP-BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:875–886, 2024. doi: 10.1109/TNSRE.2024.3366930. URL https://doi.org/10.1109/TNSRE.2024.3366930.

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, 2025. doi: 10.1038/s41467-025-61064-x. URL https://www.nature.com/articles/s41467-025-61064-x.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095889. URL https://arxiv.org/abs/2206.04769.

Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. CLIP's visual embedding projector is a few-shot cornucopia. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3254–3264, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Fahes_CLIPs_Visual_Embedding_Projector_is_a_Few-shot_Cornucopia_WACV_2026_paper.html.

Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://papers.nips.cc/paper_files/paper/2013/hash/7cce53cf90577442771720a370c3c723-Abstract.html.

Sam Gijsen and Kerstin Ritter. EEG-language pretraining for highly label-efficient clinical phenotyping. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19480–19504. PMLR, 2025. URL https://proceedings.mlr.press/v267/gijsen25a.html.

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. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. LP++: A surprisingly strong linear probe for few-shot CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23773–23782, 2024. doi: 10.1109/CVPR52733.2024.02244. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_LP_A_Surprisingly_Strong_Linear_Probe_for_Few-Shot_CLIP_CVPR_2024_paper.html.

Zhen Lan, Zixing Li, Chao Yan, Xiaojia Xiang, Dengqing Tang, Min Wu, and Zhenghua Chen. MTSNet: Convolution-based transformer network with multi-scale temporal-spectral feature fusion for SSVEP signal decoding. *IEEE Journal of Biomedical and Health Informatics*, 29(11): 8034–8047, 2025. doi: 10.1109/JBHI.2025.3573410. URL https://doi.org/10.1109/JBHI.2025.3573410.

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. doi: 10.1088/1741-2552/aace8c. URL https://doi.org/10.1088/1741-2552/aace8c.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Fabien Lotte. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces. *Proceedings of the IEEE*, 103(6):871–890, 2015. doi: 10.1109/JPROC.2015.2404941. URL https://doi.org/10.1109/JPROC.2015.2404941.

Yongqiang Ma, Yulong Liu, Liangjun Chen, Guibo Zhu, Badong Chen, and Nanning Zheng. Brain-CLIP: Brain representation via CLIP for generic natural visual stimulus decoding. *IEEE Transactions on Medical Imaging*, 44(10):3962–3972, 2025. doi: 10.1109/TMI.2025.3537287. URL https://doi.org/10.1109/TMI.2025.3537287.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

Masaki Nakanishi, Yijun Wang, Xiaogang Chen, Yu-Te Wang, Xiaorong Gao, and Tzyy-Ping Jung. Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis. *IEEE Transactions on Biomedical Engineering*, 65(1):104–112, 2018. doi: 10.1109/TBME.2017.2694818. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5783827/.

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

doi: 10.3389/frobt.2025.1625731. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2025.1625731/full.

Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://papers.nips.cc/paper_files/paper/2009/hash/1543843a4723ed2ab08e18053ae6dc5b-Abstract.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Herbert Ramoser, Johannes Müller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 2000. doi: 10.1109/86.895946. URL https://ieeexplore.ieee.org/document/895946/.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. Embeddinggemma: Powerful and lightweight text representations, 2025. URL https://arxiv.org/abs/2509.20354.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730. URL https://doi.org/10.1002/hbm.23730.

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, 2023. doi: 10.1109/TNSRE.2022.3230250. URL https://doi.org/10.1109/TNSRE.2022.3230250.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI Competition IV. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2012.00055/full.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

François-Benoît Vialatte, Monique Maurice, Justin Dauwels, and Andrzej Cichocki. Steady-state visually evoked potentials: focus on essential paradigms and future perspectives. *Progress in Neurobiology*, 90(4):418–438, 2010. doi: 10.1016/j.pneurobio.2009.11.005. URL https://doi.org/10.1016/j.pneurobio.2009.11.005.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for SSVEP-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2017. doi: 10.1109/TNSRE.2016.2627556. URL https://ieeexplore.ieee.org/document/7740878/.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. *Nature*, 620 (7976):1031–1036, 2023. doi: 10.1038/s41586-023-06377-x. URL https://www.nature.com/articles/s41586-023-06377-x.

Rui Yan, Yibo Li, Han Ding, and Fei Wang. Cross-domain EEG-based emotion recognition with contrastive learning. In *ICASSP 2026 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7887–7891, 2026. doi: 10.1109/ICASSP55912.2026.11460520. URL https://doi.org/10.1109/ICASSP55912.2026.11460520.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Shuailei Zhang, Muyun Jiang, Wei Zhang, Jinbo Chen, Zhiwei Guo, Yong Li, Yi Ding, and Cuntai Guan. EEG-PRIME: Prototype-aligned representation learning with multi-level conditioning for EEG decoding, 2026. URL https://arxiv.org/abs/2608.13072.

Hui Zheng, Hai-Teng Wang, Wei-Bang Jiang, Zhong-Tao Chen, Li He, Pei-Yang Lin, Peng-Hu Wei, Guo-Guang Zhao, and Yun-Zhe Liu. Du-IN: Discrete units-guided mask modeling for decoding speech from Intracranial Neural signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79996–80033. Curran Associates, Inc., 2024. doi: 10.52202/079017-2542. URL https://papers.nips.cc/paper_files/paper/2024/hash/92559987ee79e42a2b01d534a54682ee-Abstract-Conference.html.

A Experimental Details

This appendix specifies the data processing, encoder interfaces, and evaluation measures used in the main performance comparison. Section A.6 describes the adaptation algorithms separately. MI denotes finger motor imagery, SSVEP denotes visual target decoding from steady-state visual evoked potentials, and Speech denotes overt word reading. MI and SSVEP use scalp EEG; Speech uses intracranial stereoelectroencephalography (sEEG). Within each task, CLS and LaS use the same training and evaluation splits. Each neural encoder is trained separately and learns its own alignment space.

A.1 Shared Conventions

Table A.1 defines the task names, window conditions, and stages of text representation used throughout the paper.

Table A.1: Terminology and naming conventions.

Preferred term	Meaning and scope
Language as Supervision (LaS)	Language-supervised decoding by aligning neural representations with natural-language class descriptions.
Class-label supervision (CLS)	Neural-classification comparator trained with discrete labels and a task-specific head; LaS also uses class labels to establish pairs.
One-hot targets; cross-entropy (CE)	Target encoding and loss function, respectively.
Motor imagery (MI)	Finger motor-imagery decoding task.
SSVEP	Steady-state visual evoked potential decoding of visual stimulus targets.
Speech	Speech-decoding task using overt word reading.
Scalp EEG; sEEG	Non-invasive scalp recording and invasive stereotactic intracranial recording, respectively; MI and SSVEP use the former, Speech the latter.
Random-window (RW)	Contiguous window of length L with a random valid start within a training trial.
Initial-window (IW)	First L seconds from the defined latency-corrected valid onset; used for IW training or IW evaluation.
Inter-trial recombination (IR)	Segment recombination among same-class training trials; implementation is defined in Appendix D.2.
IW $\rightarrow$ IW; IW + IR $\rightarrow$ IW; RW $\rightarrow$ IW	Training-window condition is left of the arrow and evaluation condition right of it; IR denotes additional recombination.
Prompt template	Sentence template containing a class variable.
Prompt, P_c	Complete text passed to the text encoder.
Text embedding, $T_{f,c}$	Text-encoder vector before the task-specific text projection.
Text prototype, $z_{\mathrm{t,proj}}^{(c)}$	Class reference after the current model's text projection and ℓ_2 normalization.
Held-out templates; unseen classes	Templates absent from training and classes absent from neural-model training, respectively.
Similarity score; logit	$s_c(X_i)$ is cosine similarity; $a_{ic} = s_c(X_i)/\tau$ is the temperature-scaled logit.
Batch size, B	Number of neural–text pairs in one pairing matrix.
Class count within a batch, n_c	Number of pairs of class c in the current batch; $B = \sum_{c=1}^{C} n_c$.
Balanced class multiplicity, m	Number of pairs per class in a balanced batch; $B = mC$.
Frozen/LoRA	Frozen or low-rank-adapted text encoder; the neural and projection modules are specified separately.

The datasets are the finger motor-imagery recordings of Ding et al., the SSVEP Benchmark dataset of Wang et al., and the Chinese word-reading sEEG dataset of Zheng et al. Du-IN names an encoder. The Speech targets are Chinese words, whereas their prompts use the corresponding English translations. Prompt text shown in the figures follows the actual input wording.

Supervision and text representations. CLS trains a task-specific classification head with class labels. One-hot encoding specifies a target representation, and cross-entropy specifies a loss. LaS also uses class labels to pair neural trials with text, then predicts through matching between neural representations and class text prototypes.

Let P_c be the complete prompt supplied to the text encoder, and let $T_{f,c}$ be the resulting text embedding before the model-specific projection. Text projection followed by ℓ_2 normalization produces a text prototype. A prompt template becomes a prompt after its class variable is filled. The main performance comparison uses frozen EmbeddingGemma (300M), checkpoint `google/embeddinggemma-300m` (Schechter Vera et al., 2025), with one fixed, semantically matched P0 prompt per class during training and evaluation. The model input has one preceding `task: sentence similarity | query:` prefix. Attention-mask mean pooling, the model's native `2_Dense` and `3_Dense` layers, and ℓ_2 normalization yield a 768-dimensional embedding. Frozen text features can be precomputed; the neural encoder and both projections remain trainable in LaS.

Prompt conditions. P0 uses template 0 for each task. P2 samples from templates 0–4 during training while preserving the class meaning; its regular validation and evaluation use template 0. Held-out templates S0–S4 are absent from template training. For their ensemble, the five sets of candidate-class logits are averaged classwise before prediction. The exact training and held-out templates appear in the respective task sections.

Reading the architecture tables. The following tables omit the batch dimension and use C for the number of classes in the current task. For each encoder, the CLS and LaS paths begin at a common network boundary and list the operations performed thereafter. Neural and text projections are bias-free, and their outputs are ℓ_2-normalized. Feature readouts differ between paths for some encoders.

A.2 Motor Imagery (MI)

Data and main comparison. The finger motor-imagery data of Ding et al. (2025) contain 21 participants and four classes: Thumb, Index, Middle, and Pinky. The original 128-channel recordings were sampled at 1024 Hz. Offline trials are located by `Target` and the corresponding `TrialEnd` event, retaining at most 5.0 s after `Target`. We first apply common-average referencing across the 128 channels, select 32 channels, and resample to 250 Hz by the polyphase method. Each segment receives 250 zero samples at both ends before fourth-order Butterworth 4–40 Hz band-pass filtering with `lfilter`; the padding is then removed. The selected channels, in order, are Fp1, AF3, F7, F3, FC1, FC5, T7, C3, CP1, CP5, P7, P3, Pz, PO3, O1, Oz, O2, PO4, P4, P8, CP6, CP2, C4, T8, FC6, FC2, F4, F8, AF4, Fp2, Fz, and Cz.

The models receive 1.0-s windows at a 0.2-s stride; a complete 5.0-s trial yields 21 windows. Each window is standardized channelwise using its own statistics. Windows must lie entirely within a valid trial, and non-finite inputs are excluded. Physical trials are stratified by class into five folds within each participant. In each fold, the training trials from four folds are pooled across the 21 participants, and the remaining fold supplies held-out predictions. All windows from a trial inherit the same fold. Metrics are computed after combining held-out predictions within each participant, followed by an equal-weight participant average.

Table A.2: MI protocol for the main performance comparison.

Item	Setting
Trial split	Class-stratified five-fold split of physical trials within each of 21 participants; four folds train and one fold is held out. All windows of a trial stay in one fold.
Model input	32 channels at 250 Hz; 1.0-s window, 0.2-s stride, and 5.0-s complete trial.
Candidate classes	Four finger motor-imagery classes.
Aggregation	Combine held-out trial predictions within each participant, then average participant metrics equally.

Training and checkpoint selection. We train neural models with AdamW and cosine learning-rate decay, using weight decay and dropout for regularization. Optimization settings vary by encoder and supervision scheme. The checkpoint with the highest window-level accuracy on the held-out fold is selected for trial-level evaluation on that same fold.

Encoders and CLS/LaS interfaces. The MI encoders are EEGNet (Lawhern et al., 2018), a Deep-ConvNet1s variant of DeepConvNet (Schirrmeister et al., 2017), and EEG-Conformer (Song et al., 2023). Tables A.3 and A.4 specify the classification and projection paths used here.

Table A.3: MI encoders: CLS readout and classification paths.

Module	Sub-module	Parameter	Value
EEGNet	Common boundary	Position; shape	After separable-convolution ELU; 16×25
	CLS readout	Average pool	Kernel/stride $(1, 8)/(1, 8)$
		Dropout	$p = 0.5$
		Flatten	$16 \times 3 \to 48$
	CLS head	Linear	$48 \to C$
DeepConvNet 1s	Common boundary	Position; shape	After fourth convolutional block; $200 \times 1 \times 6$
	CLS head	Classification convolution	Channels $200 \to C$; kernel $(1, 6)$; stride 1; padding 0
EEG-Conformer	Common boundary	Position; shape	Transformer output; 11×40 tokens
	CLS readout	Flatten	440
	CLS hidden 1	Linear, ELU, dropout	$440 \to 256$; $p = 0.5$
	CLS hidden 2	Linear, ELU, dropout	$256 \to 32$; $p = 0.3$
	CLS output	Linear	$32 \to C$

Table A.4: MI encoders: LaS readout and projection paths.

Module	Sub-module	Parameter	Value
EEGNet	Common boundary	Position; shape	After separable-convolution ELU; 16×25
	LaS readout	Flatten	$16 \times 25 \to 400$
	Neural projection	Linear	$400 \to 256$
	Text projection	Linear	$768 \to 256$
DeepConvNet 1s	Common boundary	Position; shape	After fourth convolutional block; $200 \times 1 \times 6$
	LaS readout	Flatten	$200 \times 1 \times 6 \to 1200$
	Neural projection	Linear	$1200 \to 256$
	Text projection	Linear	$768 \to 256$
EEG-Conformer	Common boundary	Position; shape	Transformer output; 11×40 tokens
	LaS readout	Flatten tokens	$11 \times 40 \to 440$
	Neural projection	Linear	$440 \to 256$
	Text projection	Linear	$768 \to 256$
Shared LaS setting	Both projections	Bias; output normalization	None; separate ℓ_2 normalization
	Numerical path	Precision	FP32 projection and normalization

FBCSP-LDA (Ang et al., 2008) is a separate conventional MI baseline fitted on the corresponding training folds. EEGNet's LaS feature is read before its second pooling layer, giving that pair distinct readout paths.

Class prompts. The variable {finger} takes thumb, index finger, middle finger, or pinky. Table A.5 gives the English templates before the EmbeddingGemma prefix is added.

Table A.5: MI P0/P2 training and held-out templates.

ID	Template text
0 (P0)	The participant is imagining repetitive flexion and extension of the right {finger}.
1	The person is performing motor imagery of moving the right {finger}.
2	This EEG corresponds to imagined flexion and extension of the right {finger}.
3	The participant is mentally rehearsing repetitive movement of the right {finger}.
4	The intended action is flexion and extension of the right {finger}, imagined without physical movement.
S0	an EEG of imagined right {finger} motion.
S1	a motor-imagery signal for the right {finger}.
S2	a neural pattern for right {finger} imagery.
S3	brain activity from an imagined right {finger} action.
S4	an imagined right {finger} gesture.

Table A.6: MI parameter counts and classification forward FLOPs.

Encoder	Duration (s)	Params (M)		Forward GFLOPs	
		CLS	LaS	CLS	LaS
EEGNet	5.0	0.002	0.30	0.22	0.23
DeepConvNet1s	5.0	0.29	0.79	0.55	0.56
EEG-Conformer	5.0	0.29	0.48	0.85	0.85

The complete set of template-0 (T0) candidate prompts is encoded once, with a text-encoding cost of 53.1880 GFLOPs.

A.3 Steady-State Visual Evoked Potentials (SSVEP)

Data and main comparison. We use the SSVEP Benchmark dataset of Wang et al. (2017). It contains 35 participants and 40 stimulus classes. Each participant completed six blocks, each with one trial per class. Neural models in the main performance comparison use nine channels (Pz, PO5, PO3, POz, PO4, PO6, O1, Oz, O2) at 250 Hz. For each trial and channel, we subtract the mean over the 0.5-s pre-stimulus interval. A sixth-order Butterworth 6–64 Hz band-pass filter is applied with filtfilt to the complete recorded segment before window extraction. This processing path applies no further resampling or referencing. Each extracted window is standardized channelwise using its own statistics.

Stimulus onset is $t = 0$. An initial window (IW) of length L occupies $[0.14, 0.14 + L)$ s. For random-window (RW) training, a valid sample-aligned start a is drawn uniformly with $0.14 \leq a \leq 5.00 - L$; all channels are extracted over the same interval $[a, a + L)$. The matched neural-encoder comparison uses RW training and IW evaluation at 0.6, 0.8, and 1.0 s. Approximately 300,000 RW windows are generated from each fold's training partition. The relevant time-domain inputs receive contiguous temporal masking at ratio 0.3 under the same condition for paired CLS and LaS encoders. Appendix D.2 defines the augmentation operation.

Participants are divided into seven groups of five. In each fold, one group is evaluated, another is used for validation, and the remaining five groups (25 participants) are used for training. Each group serves as the evaluation group in turn. This subject split defines the Table 1 encoder comparison.

Table A.7: SSVEP protocol for the paired encoder comparison in Table 1.

Item	Setting
Participant split	Seven folds; each has 25 training, five validation, and five evaluation participants.
Training input	RW of 0.6, 0.8, or 1.0 s; temporal masking at ratio 0.3 for the applicable time-domain input.
Evaluation input	IW of the corresponding length, starting 0.14 s after stimulus onset.
Candidate classes	All 40 stimulus classes.

Training and checkpoint selection. Within each cross-subject fold, we train neural models with AdamW and cosine learning-rate decay, using weight decay and dropout for regularization. Optimization settings and training duration vary by encoder, input-window length, and supervision scheme (CLS or LaS). The checkpoint with the highest classification accuracy on the validation subjects is selected for evaluation on the held-out test subjects.

Encoders and CLS/LaS interfaces. The SSVEP encoders are EEGNet (Lawhern et al., 2018), SSVEPFormer (Chen et al., 2023), and MTSNet (Lan et al., 2025). MTSNet's frequency branch zero-pads the observed short window to 1250 samples and concatenates the real and imaginary parts of its complex spectrum for $8 \leq f < 64$ Hz. In Tables A.8 and A.9, T is the number of input samples. Durations of 0.6, 0.8, and 1.0 s correspond to $T = 150, 200, 250$; slash-separated dimensions follow this order.

Table A.8: SSVEP encoders: CLS readout and classification paths.

Module	**Sub-module**	**Parameter**	**Value**
EEGNet	Common boundary	Position	After final dropout
	CLS readout	Flattened width d	384/576/672
	CLS head	Linear	$d \to C$
SSVEPFormer	Common boundary	Position	After frequency-domain encoder
	CLS readout	Flatten	10080
	CLS hidden	Linear	$10080 \to 240$
		LayerNorm, GELU, dropout	Width 240; $p = 0$
	CLS output	Linear	$240 \to C$
MTSNet	Common boundary	Position	After time-frequency fusion
	CLS readout	Flatten, dropout	$d = 18(T + 560)$; $p = 0.5$
	CLS hidden	Linear	$d \to 240$
		LayerNorm, GELU, dropout	Width 240; $p = 0.5$
	CLS output	Linear	$240 \to C$

Table A.9: SSVEP encoders: LaS readout and projection paths.

Module	**Sub-module**	**Parameter**	**Value**
EEGNet	Common boundary	Position	After final dropout
	LaS readout	Flattened width d	384/576/672
	Neural projection	Linear	$d \to 128$
	Text projection	Linear	$768 \to 128$
SSVEPFormer	Common boundary	Position	After frequency-domain encoder
	LaS readout	Avg + max pooling	Elementwise sum; kernel/stride 10/10
		Flattened width	1008
	Neural projection	Linear	$1008 \to 256$

Table A.9 (continued)

Module	Sub-module	Parameter	Value
	Text projection	Linear	$768 \to 256$
MTSNet	Common boundary	Position	After time-frequency fusion
	LaS readout	Avg + max pooling	Elementwise sum; kernel/stride 10/10
		Flattened width d	1278/1368/1458
	Neural projection	Linear	$d \to 256$
	Text projection	Linear	$768 \to 256$
Shared LaS setting	Both projections	Bias; output normalization	None; separate ℓ_2 normalization

EEGNet projects both modalities to 128 dimensions; SSVEPFormer and MTSNet use 256 dimensions. Each encoder learns its own projections.

Class prompts. The variable {`frequency`} is filled with the stimulus frequency in hertz. Table A.10 reproduces the template text before the EmbeddingGemma prefix.

Table A.10: SSVEP P0/P2 training and held-out templates.

ID	Template text
0 (P0)	`A neural SSVEP signal at {frequency} Hz.`
1	`A {frequency}-Hz SSVEP.`
2	`SSVEP signal, frequency {frequency} Hz.`
3	`Steady-state visual evoked potential at {frequency} Hz.`
4	`The neural response SSVEP corresponds to a {frequency}-Hz rhythm`
S0	`An EEG response at {frequency} Hz.`
S1	`An SSVEP at {frequency} Hz.`
S2	`A response to {frequency} Hz flicker.`
S3	`A response to {frequency} Hz stimulation.`
S4	`A steady-state response at {frequency} Hz.`

Table A.11: SSVEP parameter counts and classification forward FLOPs.

Encoder	Duration (s)	Params (M)		Forward GFLOPs	
		CLS	LaS	CLS	LaS
EEGNet	0.6	0.04	0.32	0.04	0.06
	0.8	0.05	0.37	0.05	0.07
	1.0	0.05	0.39	0.07	0.08
MTSNet	0.6	9.92	7.37	0.33	0.34
	0.8	10.87	8.12	0.36	0.37
	1.0	11.89	8.95	0.40	0.41
SSVEPFormer	0.6	3.09	1.11	0.05	0.06
	0.8	3.09	1.11	0.05	0.06
	1.0	3.09	1.11	0.05	0.06

The complete set of template-0 (T0) candidate prompts is encoded once, with a text-encoding cost of 1087.5419 GFLOPs.

A.4 Overt Word Reading (Speech)

Data and main comparison. We use the Chinese word-reading sEEG data released by Zheng et al. (2024), comprising 12 participants and 61 word classes. The published 3.0-s physical trials

are used without additional event realignment, with ten fixed channels per participant. Acquisition, upstream filtering, and bipolar referencing follow the released data. After loading these segments, we select the channels and sampling rate required by each encoder and apply normalization. Du-IN receives 3000 samples at 1000 Hz; EEGNet and EEG-Conformer receive 750 samples at 250 Hz; CNN-BiGRU receives 90 samples at 30 Hz. All inputs span the same 3.0-s trial.

Within each participant's run, physical trials are shuffled and approximately 80% of each class enters training. Remaining trials are pooled and split approximately equally between validation and evaluation, yielding an overall split near 80%/10%/10%. Training portions from the participants are pooled, and CLS and LaS share the split. At each input sampling rate, training, validation, and evaluation sets are standardized separately using their own unlabeled per-participant, per-channel statistics. During training, all four encoders use the random temporal shift from the Du-IN downstream protocol: each 3.0-s segment is shifted left or right by up to 0.3 s and zero-padded. Validation and evaluation use no such shift (Zheng et al., 2024). Each trial receives one prediction.

Table A.12: Three-second Speech trial inputs for the main performance comparison.

Encoder	Sampling rate	Samples per trial and evaluation unit
Du-IN	1000 Hz	3000; one prediction per trial
EEGNet	250 Hz	750; one prediction per trial
EEG-Conformer	250 Hz	750; one prediction per trial
CNN-BiGRU	30 Hz	90; one prediction per trial

Training and checkpoint selection. Speech training follows the downstream setup of Du-IN (Zheng et al., 2024). We use AdamW with weight decay and learning-rate warmup followed by cosine decay. Encoder-specific settings, including batch size and training duration, are set separately for CLS and LaS. For both schemes, the checkpoint with the highest validation accuracy is selected for evaluation on the test split.

Encoders and CLS/LaS interfaces. All four neural backbones are trained from scratch. Tables A.13 and A.14 specify their readout and output paths.

Table A.13: Speech encoders: CLS readout and classification paths.

Module	**Sub-module**	**Parameter**	**Value**
EEGNet	Common boundary	Position	Before final pooling
	Backbone end	Average pool	Kernel/stride $(1, 8)/(1, 8)$
		Dropout	$p = 0.5$
	CLS readout	Flatten	144
	CLS head	Linear; output	$144 \to C$; raw logits
CNN-BiGRU	Common boundary	Position	Bidirectional GRU output
	CLS readout	Last sequence position	256
	CLS hidden	Linear	$256 \to 128$
	CLS output	Linear, sigmoid	$128 \to C$
EEG-Conformer	Common boundary	Position; shape	Transformer output; 50×128 tokens
	CLS readout	Flatten	6400
	CLS hidden	Linear	$6400 \to 128$
	CLS output	Linear, sigmoid	$128 \to C$
Du-IN	Common boundary	Position; shape	Transformer output; 30×160 tokens
	CLS readout	Flatten	4800
	CLS hidden	Linear, ReLU	$4800 \to 128$
	CLS output	Linear, sigmoid	$128 \to C$

Table A.14: Speech encoders: LaS readout and projection paths.

Module	Sub-module	Parameter	Value
EEGNet	Common boundary	Position	Before final pooling
	Backbone end	Average pool	Kernel/stride $(1, 8)/(1, 8)$
		Dropout	$p = 0.1$
	LaS readout	Flatten	144
	Neural projection	Linear	$144 \to 128$
	Text projection	Linear	$768 \to 128$
CNN-BiGRU	Common boundary	Position	Bidirectional GRU output
	LaS readout	Concatenated final hidden states	256
	Neural projection	Linear	$256 \to 256$
	Text projection	Linear	$768 \to 256$
EEG-Conformer	Common boundary	Position; shape	Transformer output; 50×128 tokens
	LaS readout 1	Tokenwise bias-free linear	$128 \to 32$
	LaS readout 2	Dropout, flatten	$p = 0.1$; 1600
	Neural projection	Linear	$1600 \to 256$
	Text projection	Linear	$768 \to 256$
Du-IN	Common boundary	Position; shape	Transformer output; 30×160 tokens
	LaS readout 1	Tokenwise bias-free linear	$160 \to 40$
	LaS readout 2	Dropout, flatten	$p = 0.1$; 1200
	Neural projection	Linear	$1200 \to 256$
	Text projection	Linear	$768 \to 256$
Shared LaS setting	Both projections	Bias; output normalization	None; separate ℓ_2 normalization

In the Speech EEG-Conformer and Du-IN Transformer backbones, the attention and first feed-forward dropout ratios are 0.2 for CLS and 0.1 for LaS; the second feed-forward dropout is zero in both paths. The token outputs in the two tables identify the corresponding layer, while their values depend on the branch configuration.

Speech EEGNet applies multiclass cross-entropy to raw logits and class indices. The Du-IN, CNN-BiGRU, and EEG-Conformer CLS implementations call `F.cross_entropy` on sigmoid outputs with one-hot targets, following the public Du-IN Speech benchmark implementation.

Subject-specific input frontends. Figure 2c depicts routing by participant identity into a subject-specific input frontend, followed by the shared part of that encoder instance. Table A.15 specifies the layers before the shared boundary for each architecture. The frontend outputs are restored to the original batch order. CLS and LaS are trained separately for each encoder. Shapes omit the batch dimension; convolutional channel counts are written input-to-output.

Table A.15: Subject-specific input frontends and shared-backbone boundaries for Speech decoding.

Module	Sub-module	Parameter	Value
Du-IN	Input	Shape	10×3000 (channels $\times$ samples)
	Segmentation	Output	30 segments of 10×100
	Temporal Conv1d 1	Channels; kernel; stride; padding	$10 \to 128$; 19; 10; 9
	Normalization 1	BatchNorm1d	128 channels; $\epsilon = 10^{-5}$; momentum 0.1

Table A.15 (continued)

Module	Sub-module	Parameter	Value
	Temporal Conv1d 2	Channels; kernel; stride; padding	$128 \to 128$; 3; 1; 1
	Normalization 2	BatchNorm1d	128 channels; $\epsilon = 10^{-5}$; momentum 0.1
	Temporal Conv1d 3	Channels; kernel; stride; padding	$128 \to 16$; 3; 1; 1
	Normalization 3	BatchNorm1d	16 channels; $\epsilon = 10^{-5}$; momentum 0.1
	Frontend readout	Reshape; output	Flatten 16×10 per segment; 30 tokens of width 160
	Shared boundary	First shared module	Temporal embedding, then Transformer
EEGNet	Input	Shape	10×750; Conv2d input $1 \times 10 \times 750$
	Temporal Conv2d	Channels; kernel; stride; padding	$1 \to 8$; $(1, 80)$; 1; same
	Normalization 1	BatchNorm2d	8 channels; $\epsilon = 10^{-3}$; momentum 0.01
	Spatial Conv2d	Channels; kernel; stride; groups	$8 \to 16$; $(10, 1)$; 1; 8
	Normalization 2	BatchNorm2d; output	16 channels; $\epsilon = 10^{-3}$; momentum 0.01; $16 \times 1 \times 750$
	Shared boundary	First shared module	ELU, then pooling, separable convolution, and readout
CNN-BiGRU	Input	Shape	10×90 (channel-first)
	Temporal Conv1d	Channels; kernel; stride; padding	$10 \to 128$; 5; 3; 2
	Normalization	BatchNorm1d; output	128 channels; $\epsilon = 10^{-5}$; momentum 0.1; transpose to 30 tokens of width 128
	Shared boundary	First shared module	Bidirectional GRU, then branch-specific readout
EEG-Conformer	Input	Shape	10×750; Conv2d input $1 \times 10 \times 750$
	Temporal Conv2d	Channels; kernel; stride; padding	$1 \to 128$; $(1, 25)$; $(1, 1)$; $(0, 12)$
	Spatial Conv2d	Channels; kernel; stride; padding	$128 \to 128$; $(10, 1)$; 1; 0
	Normalization	BatchNorm2d; output	128 channels; $\epsilon = 10^{-5}$; momentum 0.1; $128 \times 1 \times 750$
	Shared boundary	First shared module	ELU, then pooling, token projection, and Transformer

The Du-IN, CNN-BiGRU, and EEG-Conformer frontend convolutions use $\mathrm{groups} = 1$ and bias. EEGNet uses a bias-free temporal convolution and a grouped spatial convolution. Activations and pooling follow these frontends in the shared backbone. Attribution groups are defined in Appendix C.3.

Class prompts. The source targets are 61 Chinese words. Their English translations, listed in Table 4 of Zheng et al. (2024), fill `{label}`. Table A.16 reproduces the templates before the EmbeddingGemma prefix.

Table A.16: Speech P0/P2 training and held-out templates.

ID	Template text
0 (P0)	`The participant is speaking the word: {label}.`
1	`This sEEG recording corresponds to the spoken word: {label}.`
2	`The intended speech content is: {label}.`

Table A.16 (continued)

ID	Template text
3	`The neural activity was recorded while the participant said: {label}.`
4	`The target speech label for this intracranial neural signal is: {label}.`
S0	`An sEEG response for spoken {label}.`
S1	`A neural signal for saying {label}.`
S2	`Brain activity from uttering {label}.`
S3	`An intracranial response to the word {label}.`
S4	`A speech-related neural pattern for {label}.`

Table A.17: Speech parameter counts and classification forward FLOPs.

Encoder	Duration (s)	Params (M)		Forward GFLOPs	
		CLS	LaS	CLS	LaS
EEGNet	3.0	0.02	0.13	0.01	0.02
CNN-BiGRU	3.0	0.62	0.84	0.03	0.05
EEG-Conformer	3.0	4.18	3.96	0.45	0.47
Du-IN	3.0	5.05	4.94	0.27	0.29

The complete set of template-0 (T0) candidate prompts is encoded once, with a text-encoding cost of 811.1166 GFLOPs.

A.5 Metrics and Aggregation

For N evaluation trials within a participant, accuracy is

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\hat{y}_i = y_i]. \tag{1}$$

For MI, we also report Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is that participant's trial accuracy and $p_e = \sum_c p_{\text{true}}(c)p_{\text{pred}}(c)$ is determined by the marginal proportions of true and predicted classes. SSVEP estimated information transfer rate (ITR) uses 40 candidate classes, accuracy p, and decision time $T = L + 0.5$ s:

$$\text{ITR} = \frac{60}{T}\left[\log_2 C + p\log_2 p + (1-p)\log_2\frac{1-p}{C-1}\right], \qquad C = 40. \tag{2}$$

Here L is the window duration, and $0\log_2 0$ takes its continuous limit of zero. ITR is computed within each participant before equal-weight averaging across participants.

For MI trial prediction, LaS averages window logits; CLS and FBCSP-LDA use majority voting over window classes. A voting tie is resolved by the tied classes' mean scores, then by a fixed class index if necessary. Accuracy and κ are computed from each participant's combined five-fold held-out trials and averaged equally across the 21 participants. SSVEP accuracy is computed for each of the 35 evaluation participants and then averaged equally. The Speech main comparison uses five independent training runs on the same physical-trial split: run results are first averaged within each participant, then averaged equally across 12 participants.

A.6 Adaptation Algorithms for MI and SSVEP

The methods compared in Figures 4a and 4b update different parts of a source decoder. Let E_0 be the source neural encoder, $h = E_0(X)$ its representation, and W_0 the source neural projection. For the current candidate set $\mathcal{C}$, t_c denotes a fixed unit-length text prototype, and ν denotes ℓ_2 normalization. Methods using candidate-class scores optimize multiclass cross-entropy over $\mathcal{C}$. Tables A.4 and A.9 locate the encoder readouts; projected features are read after that model's neural projection.

Probe heads. A linear probe freezes the source feature extractor, reads an encoder or projected feature $r(X)$, and trains a linear classifier $g(r) = Hr + b$. An MLP probe trains a nonlinear head $g_\theta(r)$ on the same frozen features. Encoder-feature and projected-feature variants differ in their readout position. Both heads use cross-entropy on calibration labels, and their output logits are used at inference.

Adaptation using text references. LP++ combines learned class weights with text representations over frozen vision-language features (Huang et al., 2024). Here the source projected neural feature $v = \nu(hW_0)$ and prototypes t_c are fixed, while class weights w_c and coefficients α_c are trained with scores

$$z_c(X) = v^\top(w_c + \alpha_c t_c). \tag{3}$$

ProLIP motivates adapting a projection while penalizing deviation from its source value (Fahes et al., 2026). In this neural-decoding implementation, the neural encoder and text prototypes are fixed, the neural projection W starts from W_0, and $\gamma\langle\nu(hW), t_c\rangle$ provides class logits. Its calibration objective adds $\lambda\|W - W_0\|_F^2$ to cross-entropy.

DTS-ProLIP. We constrain the trainable neural-projection offset to the discriminative subspace of the current candidate prototypes. Let $W_0 \in \mathbb{R}^{d_f \times d}$. After centering $\{t_c\}_{c\in\mathcal{C}}$, let $U \in \mathbb{R}^{d\times r}$ contain an orthonormal basis for their span, with $U^\top U = I_r$. Only $A \in \mathbb{R}^{d_f \times r}$ is trained, starting at zero:

$$W(A) = W_0 + AU^\top, \qquad A_0 = 0. \tag{4}$$

The neural encoder, text prototypes, and source projection remain fixed. The calibration objective is

$$\mathcal{L}(A) = \frac{1}{M}\sum_{i=1}^{M} \mathrm{CE}\big(\{\gamma\,\langle\nu(h_i W(A)), t_c\rangle\}_{c\in\mathcal{C}}, y_i\big) + \lambda\|W(A) - W_0\|_F^2. \tag{5}$$

Because U has orthonormal columns, $\|W(A) - W_0\|_F^2 = \|A\|_F^2$.

Updates beyond a probe. MLP fine-tuning updates an encoder or projection branch along with its MLP readout head. Full fine-tuning starts from the source model and updates the corresponding neural modules in its encoder- or projection-centered variant. The names *Encoder* and *Projection* identify the feature readout and update path. Classification uses the trained head or the corresponding candidate-prototype scores.

B Additional Experimental Results

This appendix reports subject-level results for Table 1 and Figure 3 and aggregate results for the other analyses, using the evaluation units defined for each experiment. Dataset splits and model-specific protocols are specified in Appendix A, attribution methods and loss references in Appendix C, and SSVEP window analyses in Appendix D.

B.1 Main Results, Candidate-Set Reuse, and Adaptation

SSVEP accuracy and estimated ITR under initial-window evaluation. Table B.1 reports accuracy and estimated ITR at five IW durations. Neural CLS baselines use IW training, LaS (MTSNet) uses IW training with inter-trial recombination, and every method is evaluated on the initial Physical window. FBCCA directly classifies the window (Chen et al., 2015). This experiment has a different training protocol from the RW-to-IW paired comparison in Table 1. Stars in Table B.1 denote unadjusted, two-sided paired t-tests, performed separately for accuracy and estimated ITR.

Table B.1: SSVEP subject-mean accuracy (%) and estimated ITR (bits/min) across five initial-window durations. Stars indicate unadjusted paired-test significance: $^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$.

Method	0.6 s		0.7 s		0.8 s		0.9 s		1.0 s	
	Acc.	ITR	Acc.	ITR	Acc.	ITR	Acc.	ITR	Acc.	ITR
FBCCA	28.25***	41.00***	38.20***	60.40***	48.54***	80.00***	56.90***	94.50***	63.75***	104.70***
EEGNet	45.24***	84.60***	52.51***	96.80***	58.38***	105.10***	63.52***	111.20***	65.45***	108.60***
SSVEPFormer	**55.25**	114.40	63.46	129.60	70.33	139.50	75.65	145.50	77.74	141.20*
MTSNet	52.06*	105.70*	62.61	126.10	68.61**	134.70**	73.98*	140.60*	78.01	142.30
LaS (MTSNet)	55.18	**114.80**	**64.46**	**132.40**	**71.93**	**144.80**	**76.42**	**147.60**	**79.50**	**146.40**

Subject-level data for Table 1. The MI table follows the same five out-of-fold trial splits for all methods, including the separately trained FBCSP-LDA reference (Ang et al., 2008). In the following tables, C denotes supervised CLS and L denotes LaS. EEGNet, DeepConvNet1s, and EEG-Conformer follow their respective neural-encoder implementations (Lawhern et al., 2018; Schirrmeister et al., 2017; Song et al., 2023). Values in paired C/L columns are subject-level metrics for the same participant.

Table B.2: MI four-class trial accuracy by subject (%). C and L denote supervised CLS and LaS.

Subject	FBCSP	EEGNet C	EEGNet L	DeepConv C	DeepConv L	Conformer C	Conformer L
S01	32.00	41.17	38.67	39.17	42.17	42.33	42.00
S02	36.83	42.33	37.00	39.83	44.33	44.17	46.83
S03	27.17	35.17	32.67	43.83	44.67	44.00	44.50
S04	28.91	39.84	36.25	38.44	38.59	36.88	35.94
S05	28.91	38.91	40.63	37.97	40.47	40.63	43.91
S06	30.67	36.00	37.00	35.33	34.33	37.00	38.33
S07	37.17	42.67	45.50	47.00	53.00	50.33	56.33
S08	29.50	39.67	41.00	37.00	36.83	39.83	42.67
S09	39.67	44.33	48.50	49.17	56.00	49.83	54.17
S10	30.33	35.17	34.50	38.67	37.67	36.17	34.50
S11	31.33	36.33	40.17	36.17	34.67	39.33	41.67
S12	31.00	33.17	36.50	34.33	31.67	38.50	38.67
S13	31.29	38.39	37.10	38.06	39.52	40.97	41.61
S14	27.34	38.28	33.13	31.72	31.56	32.34	32.03
S15	34.69	38.44	37.34	39.06	42.03	44.06	42.34
S16	28.91	31.88	36.56	30.31	28.75	35.16	37.19
S17	36.25	44.22	44.22	40.94	47.97	47.34	49.69
S18	28.59	38.91	36.09	35.16	36.88	39.84	41.25
S19	53.13	49.53	61.09	55.31	62.81	72.19	72.34
S20	31.72	32.19	32.03	34.22	36.09	32.97	34.84
S21	26.41	29.69	26.25	31.41	31.41	31.09	32.66

Table B.3: MI four-class Cohen's κ by subject.

Subject	FBCSP	EEGNet C	EEGNet L	DeepConv C	DeepConv L	Conformer C	Conformer L
S01	0.09	0.22	0.18	0.19	0.23	0.23	0.23
S02	0.16	0.23	0.16	0.20	0.26	0.26	0.29
S03	0.03	0.14	0.10	0.25	0.26	0.25	0.26
S04	0.05	0.20	0.15	0.18	0.18	0.16	0.15
S05	0.05	0.19	0.21	0.17	0.21	0.21	0.25
S06	0.08	0.15	0.16	0.14	0.12	0.16	0.18
S07	0.16	0.24	0.27	0.29	0.37	0.34	0.42
S08	0.06	0.20	0.21	0.16	0.16	0.20	0.24
S09	0.20	0.26	0.31	0.32	0.41	0.33	0.39
S10	0.07	0.14	0.13	0.18	0.17	0.15	0.13
S11	0.08	0.15	0.20	0.15	0.13	0.19	0.22
S12	0.08	0.11	0.15	0.12	0.09	0.18	0.18
S13	0.08	0.18	0.16	0.17	0.19	0.21	0.22
S14	0.03	0.18	0.11	0.09	0.09	0.10	0.09
S15	0.13	0.18	0.16	0.19	0.23	0.25	0.23
S16	0.05	0.09	0.15	0.07	0.05	0.14	0.16
S17	0.15	0.26	0.26	0.21	0.31	0.30	0.33
S18	0.05	0.19	0.15	0.14	0.16	0.20	0.22
S19	0.38	0.33	0.48	0.40	0.50	0.63	0.63
S20	0.09	0.10	0.09	0.12	0.15	0.11	0.13
S21	0.02	0.06	0.02	0.09	0.09	0.08	0.10

The SSVEP tables give 35 subject-level accuracies for each encoder and duration in the RW-training to IW-evaluation comparison. Neural-model identities are EEGNet, MTSNet, and SSVEP-Former (Lawhern et al., 2018; Lan et al., 2025; Chen et al., 2023). C and L share the same Physical evaluation trials within a row.

Table B.4: SSVEP subject accuracy (%) for EEGNet after RW training and IW evaluation.

Subject	0.6 C	0.6 L	0.8 C	0.8 L	1.0 C	1.0 L
0	30.42	28.33	45.42	41.25	53.33	52.08
1	33.75	28.75	40.83	39.58	50.00	54.17
2	48.75	47.08	58.33	57.92	75.00	72.08
3	36.67	39.58	64.58	55.83	70.00	57.50
4	20.42	19.17	34.58	30.00	43.75	38.75
5	27.50	32.08	51.67	47.50	66.25	59.17
6	19.58	23.75	38.33	37.08	49.17	48.75
7	15.00	19.58	24.17	24.17	36.25	41.67
8	24.58	25.00	42.50	42.92	53.75	54.17
9	26.67	32.50	31.67	39.17	43.75	53.33
10	6.25	6.67	4.17	5.83	6.25	10.42
11	22.50	25.42	37.92	33.75	53.75	52.08
12	21.25	15.42	24.58	19.17	27.92	29.17
13	14.58	17.50	20.42	20.83	28.33	20.42
14	22.92	21.25	38.33	39.58	44.58	45.83
15	16.67	12.92	25.00	27.08	37.92	35.83
16	32.92	37.92	49.58	50.00	59.58	55.83
17	42.50	36.67	49.58	44.58	63.33	61.25
18	9.58	11.25	17.92	15.00	22.08	24.17
19	19.17	20.83	39.17	33.33	52.08	47.50
20	18.33	20.00	36.67	35.83	55.00	47.50
21	43.33	42.08	69.58	64.58	80.83	78.33
22	25.42	27.50	53.75	50.42	62.50	58.75
23	38.75	32.92	57.92	61.25	76.67	75.00
24	28.33	25.00	40.83	41.67	45.83	50.83
25	30.42	37.92	49.58	50.42	55.42	43.75
26	45.00	41.25	70.42	66.25	81.25	79.58
27	38.75	35.00	55.83	61.67	65.83	62.92
28	13.33	12.50	16.67	17.50	25.42	30.42
29	16.25	22.08	38.75	38.75	48.33	40.42
30	37.92	37.08	48.75	55.83	70.00	61.25
31	47.92	47.92	69.17	67.50	78.75	75.42
32	16.25	17.50	18.75	21.25	27.50	29.58
33	34.58	34.17	46.67	49.17	61.25	55.83
34	10.42	9.58	25.42	27.92	35.00	41.25

Table B.5: SSVEP subject accuracy (%) for MTSNet after RW training and IW evaluation.

Subject	0.6 C	0.6 L	0.8 C	0.8 L	1.0 C	1.0 L
0	35.83	35.83	54.58	48.33	62.08	61.67
1	56.67	55.42	72.92	73.75	82.92	85.83
2	65.00	69.58	87.08	87.92	94.17	93.33
3	62.08	63.33	82.92	75.42	87.50	88.75
4	35.83	39.58	70.42	63.75	77.50	77.50
5	31.25	33.75	56.25	63.75	72.08	68.75
6	32.50	37.50	55.00	67.08	71.25	75.83
7	22.50	19.17	34.17	36.67	47.92	43.75
8	31.25	33.75	42.50	45.42	62.92	63.75
9	27.50	32.92	50.83	51.67	67.08	67.08
10	7.92	11.67	8.75	13.33	14.17	16.25
11	26.25	28.33	55.83	52.08	69.17	63.75
12	19.58	20.42	34.17	27.92	55.83	47.92
13	22.92	34.58	29.17	47.50	38.75	59.17
14	27.50	31.25	41.67	45.83	55.42	55.83
15	14.17	18.33	31.67	40.83	51.67	49.17
16	40.83	37.92	53.75	54.58	63.75	61.67
17	42.92	48.75	59.58	60.83	68.33	72.50
18	7.08	9.17	12.08	14.58	24.17	25.42
19	32.08	37.50	53.33	51.25	65.00	68.33
20	17.92	18.75	48.33	48.33	62.92	64.17
21	50.42	54.17	77.92	82.92	90.00	90.42
22	43.75	39.17	70.83	71.67	79.17	85.83
23	39.17	43.75	65.83	69.58	73.75	83.75
24	43.33	39.17	52.08	61.67	70.00	73.33
25	47.08	50.42	67.50	69.17	80.42	77.50
26	46.25	51.67	74.17	74.58	85.83	87.92
27	50.83	47.92	77.92	78.33	85.00	88.33
28	16.25	15.83	24.17	21.25	37.50	35.00
29	29.58	35.42	51.25	60.83	70.42	67.50
30	50.83	49.58	71.25	72.50	82.50	81.25
31	54.58	60.42	83.33	82.08	91.25	95.83
32	17.08	17.92	22.08	22.50	29.58	33.33
33	37.92	40.83	58.33	61.67	70.83	77.50
34	14.58	15.83	32.92	31.25	47.92	49.58

Table B.6: SSVEP subject accuracy (%) for SSVEPFormer after RW training and IW evaluation.

Subject	0.6 C	0.6 L	0.8 C	0.8 L	1.0 C	1.0 L
0	35.42	42.50	44.58	53.75	62.50	66.67
1	51.25	54.17	74.17	75.00	80.83	83.33
2	62.92	69.17	80.83	84.17	91.67	94.17
3	62.50	52.50	75.00	78.75	88.75	87.50
4	37.50	35.00	60.42	50.83	75.83	72.92
5	31.25	34.58	57.50	60.83	67.08	75.42
6	30.83	39.17	61.67	49.58	69.17	72.50
7	17.08	17.08	35.00	30.83	43.33	47.92
8	32.08	34.17	46.25	45.42	67.08	65.42
9	32.08	32.50	46.67	54.58	63.75	72.92
10	9.17	8.75	10.00	9.58	15.42	17.08
11	24.17	27.50	42.92	52.08	63.33	74.17
12	15.00	27.92	27.92	42.50	34.17	54.58
13	25.83	29.17	42.08	39.58	52.08	46.67
14	21.25	26.25	44.17	45.83	54.58	56.25
15	17.08	21.67	36.67	31.67	48.75	55.00
16	37.92	33.75	54.17	56.25	65.83	65.42
17	47.50	48.33	59.58	64.17	71.67	73.33
18	7.92	8.33	12.08	17.08	20.42	25.42
19	31.25	30.83	47.08	54.17	67.08	66.25
20	17.92	20.83	44.58	48.33	65.42	70.42
21	47.92	55.42	77.50	81.67	87.50	91.67
22	35.42	44.58	65.00	73.33	82.92	83.75
23	42.50	37.08	69.58	66.25	83.33	81.67
24	40.00	42.92	57.50	58.75	77.50	71.25
25	45.83	46.25	71.25	67.50	78.75	73.75
26	50.00	52.08	72.92	76.25	87.08	90.00
27	45.83	50.42	77.08	80.83	88.75	88.33
28	11.67	15.83	17.50	29.58	27.50	39.58
29	30.83	30.00	60.00	54.58	67.50	66.25
30	41.67	54.58	67.50	69.17	78.33	81.25
31	61.67	63.33	82.50	83.75	94.58	92.92
32	15.00	16.25	21.67	27.50	30.83	32.08
33	38.33	40.42	57.08	57.92	76.67	75.00

Subject	0.6 C	0.6 L	0.8 C	0.8 L	1.0 C	1.0 L
34	16.67	14.58	30.00	35.83	52.92	51.67

The Speech table gives the 12 subject-level accuracies for the four neural encoders used in Table 1. Du-IN and the other supervised baselines follow the Speech benchmark implementation (Zheng et al., 2024); the LaS values use the corresponding encoder and text-conditioned classifier.

Table B.7: Speech subject accuracy (%) across four neural encoders.

Subject	EEGNet C	EEGNet L	BiGRU C	BiGRU L	Conformer C	Conformer L	Du-IN C	Du-IN L
S001	9.69	24.06	51.56	70.94	60.31	73.75	73.75	77.81
S002	8.49	18.90	32.05	41.92	44.11	46.03	46.58	47.95
S003	6.69	4.46	25.63	38.72	32.03	35.93	38.44	45.68
S004	8.20	19.02	42.62	66.56	59.02	69.84	68.85	76.72
S005	11.48	22.62	50.82	69.18	61.31	73.11	73.11	77.38
S006	3.72	8.31	29.80	46.70	39.26	40.40	47.56	50.72
S007	4.96	11.57	30.03	42.15	37.47	45.18	46.01	50.69
S008	6.28	12.02	36.61	49.18	46.72	55.19	52.46	59.84
S009	6.96	14.48	39.00	53.48	49.03	55.15	51.25	62.12
S010	2.76	9.67	23.48	33.15	27.90	28.45	38.40	35.91
S011	6.99	23.40	43.16	69.60	62.31	67.17	71.12	73.86
S012	9.29	16.94	31.15	60.11	48.63	52.46	55.74	67.21

MI candidate-set reuse in Figure 3. Figure 3 compares a supervised classifier trained for each candidate set with a LaS model whose four-class checkpoint is reused with the corresponding candidates. Classes 1–4 denote thumb, index, middle, and pinky, respectively. Each table retains the 21 paired subject accuracies; C and L again denote supervised CLS and LaS.

Table B.8: MI candidate-set reuse: EEGNet subject accuracy (%).

Subject	1–4 C	1–4 L	2–3 C	2–3 L	1–3–4 C	1–3–4 L	1–2–3 C	1–2–3 L
S01	76.00	76.00	61.67	63.33	54.22	52.89	43.11	40.22
S02	74.00	76.33	66.00	65.33	44.44	46.67	45.33	42.67
S03	70.00	65.00	53.00	50.67	50.22	42.00	40.89	36.89
S04	73.44	69.38	52.19	53.44	47.92	47.08	45.83	42.71
S05	79.06	74.69	55.63	51.88	53.13	51.46	41.25	41.67
S06	69.33	67.33	59.33	56.67	47.11	47.56	42.67	44.00
S07	82.67	83.00	70.33	71.00	61.11	55.56	50.22	48.67
S08	79.33	75.67	54.67	59.00	51.78	54.89	47.33	47.11
S09	76.00	77.33	58.33	59.33	53.33	57.78	45.78	52.67
S10	70.00	66.00	55.33	59.67	47.33	44.67	42.22	40.44
S11	62.67	65.67	49.00	51.00	45.33	49.78	41.33	42.44
S12	67.67	68.67	58.67	56.67	46.67	48.44	41.33	43.56
S13	75.48	72.58	57.42	61.29	49.46	49.89	42.15	42.58
S14	60.63	61.88	57.19	57.81	47.92	41.46	41.46	40.21
S15	61.56	61.25	59.38	60.63	48.13	46.67	48.54	47.92
S16	66.88	66.25	51.88	55.31	42.71	46.88	37.71	38.75
S17	78.44	75.31	61.56	65.31	57.92	57.71	50.21	46.88
S18	66.25	68.75	49.69	54.69	46.04	49.17	42.08	38.54
S19	96.25	96.88	79.06	82.81	70.63	81.46	59.38	59.17
S20	68.13	65.31	51.56	48.75	42.92	43.96	37.08	35.63
S21	60.63	54.06	50.63	51.25	39.79	37.50	35.42	33.33

Table B.9: MI candidate-set reuse: DeepConvNet1s subject accuracy (%).

Subject	1–4 C	1–4 L	2–3 C	2–3 L	1–3–4 C	1–3–4 L	1–2–3 C	1–2–3 L
S01	63.33	76.33	55.67	59.67	50.67	55.11	39.56	47.78
S02	67.33	76.00	66.00	67.00	48.22	51.11	40.89	50.67
S03	67.00	75.00	45.67	54.33	55.33	58.67	39.11	47.56
S04	69.69	75.63	48.13	52.50	48.54	51.46	35.00	40.83
S05	71.56	78.75	52.50	52.81	50.83	53.96	39.58	42.71
S06	55.33	64.00	52.67	54.67	44.44	44.89	37.33	43.56
S07	78.67	85.67	72.33	74.67	59.78	66.22	42.67	58.44
S08	63.00	72.67	56.33	52.00	49.56	49.78	41.33	48.22
S09	65.67	77.33	51.67	63.67	57.56	65.33	45.11	59.56
S10	71.67	77.00	54.00	58.00	48.44	50.22	36.67	43.78
S11	58.33	61.33	52.33	52.00	45.11	46.44	41.33	40.67

Subject	1–4 C	1–4 L	2–3 C	2–3 L	1–3–4 C	1–3–4 L	1–2–3 C	1–2–3 L
S12	55.67	61.67	56.67	54.67	45.78	41.33	40.00	43.11
S13	68.71	75.16	49.35	58.39	49.89	51.40	35.70	48.39
S14	53.44	58.13	49.38	53.75	40.21	41.67	35.42	38.54
S15	61.88	65.00	59.69	64.69	50.42	51.04	41.88	55.83
S16	51.56	55.63	47.81	50.94	43.13	37.92	38.75	39.17
S17	70.63	78.44	56.25	64.69	53.96	61.46	38.75	47.92
S18	65.00	69.69	50.31	52.81	46.25	48.54	32.50	43.13
S19	80.00	96.88	59.38	80.63	69.38	79.17	49.38	66.04
S20	60.63	69.69	49.69	50.94	42.71	47.92	34.38	36.46
S21	55.00	59.38	46.88	52.81	39.38	41.25	30.21	38.33

Table B.10: MI candidate-set reuse: EEG-Conformer subject accuracy (%).

Subject	1–4 C	1–4 L	2–3 C	2–3 L	1–3–4 C	1–3–4 L	1–2–3 C	1–2–3 L
S01	77.00	76.67	60.67	65.33	56.44	52.00	46.00	46.89
S02	73.33	76.67	64.33	67.33	51.56	53.78	45.33	53.11
S03	77.67	77.67	56.67	59.67	58.67	60.00	45.33	46.00
S04	69.38	72.19	54.69	53.44	46.88	48.75	39.17	39.79
S05	78.75	82.81	54.38	56.56	55.83	58.54	44.79	43.13
S06	70.67	71.33	60.67	61.33	44.44	49.33	43.56	43.11
S07	87.00	89.00	69.33	73.67	64.44	66.67	54.67	58.67
S08	81.00	78.00	55.00	57.00	49.33	54.22	49.11	49.33
S09	78.67	83.00	60.67	61.33	64.22	67.33	53.33	55.78
S10	67.33	68.33	53.67	56.67	45.56	45.33	40.67	41.56
S11	67.33	71.67	55.33	50.00	48.44	55.78	44.00	43.33
S12	67.67	70.67	50.00	57.67	49.78	49.78	47.56	47.56
S13	77.42	77.10	56.77	62.90	55.05	56.13	46.45	48.39
S14	63.75	65.63	51.88	54.38	39.17	40.00	36.46	38.75
S15	72.19	70.00	62.19	63.44	55.42	52.29	56.88	51.67
S16	69.69	71.56	49.69	53.75	47.92	50.21	40.83	41.88
S17	80.31	80.63	55.63	65.00	59.38	63.75	47.29	51.67
S18	73.13	73.75	55.94	56.88	51.25	53.96	44.17	45.42
S19	98.75	99.38	76.56	91.56	85.42	89.38	63.54	70.21
S20	68.75	69.38	53.13	47.50	45.83	47.08	35.63	34.58
S21	62.19	63.13	48.75	51.25	42.29	44.38	37.29	37.92

MI cross-session adaptation in Figure 4b. Table B.11 reports subject-mean trial accuracy and between-subject SD for nine methods in the three-class and binary tasks, including LP++ and ProLIP (Huang et al., 2024; Fahes et al., 2026). Direct inference is evaluated on the Base block and adapted models on the Fine block. In Figures 4a and 4b, "MLP probe" uses encoder features, and "Full fine-tuning" denotes the Projection variant described in Appendix A.6.

Table B.11: MI cross-session adaptation: subject-mean trial accuracy ± between-subject SD (%).

Method	3-class	2-class
Direct inference	46.43 ± 9.14	64.24 ± 9.56
Linear Probe	48.69 ± 10.08	66.63 ± 9.80
LP++	48.67 ± 9.81	65.76 ± 11.08
ProLIP	49.13 ± 9.67	65.55 ± 11.48
DTS-ProLIP	49.94 ± 9.36	65.91 ± 11.93
MLP Probe Encoder	50.23 ± 10.39	67.01 ± 11.06
MLP Probe Projection	49.44 ± 10.72	66.33 ± 10.56
Full Encoder	55.34 ± 12.87	75.14 ± 10.83
Full Projection	55.48 ± 13.08	76.35 ± 10.84

SSVEP subject adaptation in Figure 4a. Table B.12 reports leave-one-block validation accuracy for eight neural adaptation methods and three TRCA configurations. K is the number of calibration blocks. TRCA methods require at least two blocks under this protocol, so their K=1 cells are unavailable (Nakanishi et al., 2018). The Figure 4a curve labelled Ensemble TRCA uses the single-band ensemble configuration; Table B.12 also reports the other Ensemble TRCA configuration.

Table B.12: SSVEP K-block adaptation: leave-one-block validation accuracy (%).

Family	Method	K=1	K=2	K=3	K=4	K=5
LaS	DTS-ProLIP	80.38	81.85	82.81	83.45	84.17

Family	Method	K=1	K=2	K=3	K=4	K=5
LaS	ProLIP	79.46	80.14	80.70	80.94	81.25
LaS	Linear Probe	71.29	79.01	81.37	82.11	82.56
LaS	LP++	75.44	77.58	79.01	79.71	80.26
LaS	MLP Probe Encoder	72.89	82.77	85.93	87.55	88.93
LaS	Full Finetune Encoder	62.92	77.26	82.88	85.57	88.07
LaS	MLP Probe Projection	71.05	79.92	83.07	83.89	84.71
LaS	Full Finetune Projection	73.18	85.15	89.24	90.74	91.93
TRCA	Standard TRCA	–	66.35	80.35	86.07	88.95
TRCA	Ensemble TRCA	–	81.92	88.50	90.90	92.27
TRCA	Single-band Ensemble TRCA	–	82.33	88.13	90.45	91.79

B.2 Conditional Unseen-Class Decoding

Figure 4c averages five-class conditional decoding across eight held-out class splits. Tables B.13–B.15 retain the accuracy of each of the 40 stimulus classes when it belongs to its unseen set. Class index i is zero based and sorted by stimulus frequency, $f_i = 8 + 0.2i$ Hz. Each class contributes 210 Physical trials to its evaluation. Neural results use the fixed initial window at offset 0.0 s and the five unseen candidates; FBCCA uses its corresponding signal-processing reference (Wang et al., 2017; Chen et al., 2015).

Table B.13: Five-class unseen-only accuracy (%) by stimulus frequency with 0.6 s inputs.

Class	Hz	MTSNet	SSVEPFormer	EEGNet	FBCCA
0	8.0	72.38	82.86	78.57	85.71
1	8.2	81.90	83.33	83.81	86.19
2	8.4	84.76	79.05	79.05	86.19
3	8.6	85.24	85.24	75.24	85.24
4	8.8	82.86	78.57	81.90	82.38
5	9.0	72.86	72.38	70.48	85.24
6	9.2	84.76	86.67	85.71	86.19
7	9.4	89.52	90.00	89.52	84.76
8	9.6	84.76	76.19	85.71	73.81
9	9.8	83.81	81.43	78.57	69.05
10	10.0	77.62	86.19	79.52	68.10
11	10.2	79.52	77.62	82.86	75.24
12	10.4	90.00	90.00	82.86	83.33
13	10.6	86.67	83.33	75.24	88.10
14	10.8	82.38	78.57	79.05	87.14
15	11.0	79.52	76.67	77.14	84.76
16	11.2	86.19	85.71	78.57	81.43
17	11.4	79.05	80.00	83.33	83.33
18	11.6	87.14	80.00	83.33	85.24
19	11.8	70.48	75.71	68.57	80.00
20	12.0	69.52	67.14	72.86	71.90
21	12.2	85.71	86.67	86.19	65.71
22	12.4	83.81	80.48	85.71	60.48
23	12.6	81.90	79.05	78.10	57.62
24	12.8	83.33	79.05	71.90	64.29
25	13.0	70.00	67.14	65.24	50.00
26	13.2	80.48	83.33	79.05	59.05
27	13.4	86.67	81.90	88.10	64.76
28	13.6	74.29	75.24	73.81	57.62
29	13.8	80.00	75.71	71.90	52.86
30	14.0	65.71	66.19	65.24	48.10
31	14.2	80.48	77.62	77.62	64.76
32	14.4	82.86	78.10	78.57	50.48
33	14.6	77.14	80.00	76.67	46.67
34	14.8	88.10	87.14	84.29	61.90
35	15.0	76.19	79.05	77.62	49.52
36	15.2	80.95	85.71	77.14	54.29
37	15.4	81.43	87.14	83.81	54.29
38	15.6	80.00	77.14	81.43	60.48
39	15.8	77.14	79.05	74.76	62.38

Table B.14: Five-class unseen-only accuracy (%) by stimulus frequency with 0.8 s inputs.

Class	Hz	MTSNet	SSVEPFormer	EEGNet	FBCCA
0	8.0	89.52	92.86	84.29	93.81
1	8.2	90.48	90.95	87.62	94.76
2	8.4	92.86	90.95	88.10	90.48

Class	Hz	MTSNet	SSVEPFormer	EEGNet	FBCCA
3	8.6	93.81	88.57	86.19	91.43
4	8.8	90.95	82.86	89.05	90.95
5	9.0	84.76	86.19	85.24	90.95
6	9.2	93.81	92.38	92.38	88.10
7	9.4	92.86	95.71	94.76	93.33
8	9.6	93.33	89.05	92.38	82.86
9	9.8	90.48	87.14	87.14	82.86
10	10.0	80.00	87.62	89.05	80.95
11	10.2	91.90	90.95	83.81	83.81
12	10.4	94.76	97.14	90.00	90.48
13	10.6	89.05	90.48	87.62	93.33
14	10.8	87.62	88.57	86.19	91.90
15	11.0	85.71	84.29	83.81	90.48
16	11.2	87.62	90.95	90.48	91.43
17	11.4	88.10	88.10	90.00	88.10
18	11.6	90.48	90.48	86.19	93.33
19	11.8	80.00	75.24	78.57	86.19
20	12.0	78.57	76.67	79.52	86.19
21	12.2	89.52	92.86	88.57	79.52
22	12.4	91.43	90.00	90.95	78.10
23	12.6	91.43	90.95	86.19	77.62
24	12.8	88.10	87.62	82.86	77.14
25	13.0	81.90	80.00	70.00	69.52
26	13.2	81.90	86.67	84.76	74.76
27	13.4	91.90	92.38	93.81	76.67
28	13.6	87.14	84.76	84.29	72.86
29	13.8	91.43	84.76	80.48	78.10
30	14.0	73.33	70.00	75.24	67.14
31	14.2	88.10	85.71	89.05	83.33
32	14.4	89.05	89.05	88.10	67.14
33	14.6	92.38	89.52	85.71	65.24
34	14.8	91.43	90.95	90.95	80.95
35	15.0	77.14	86.19	82.38	69.05
36	15.2	84.29	85.24	85.71	73.33
37	15.4	88.10	90.48	91.43	68.10
38	15.6	88.10	86.67	87.62	77.14
39	15.8	85.71	86.19	83.81	75.71

Table B.15: Five-class unseen-only accuracy (%) by stimulus frequency with 1.0 s inputs.

Class	Hz	MTSNet	SSVEPFormer	EEGNet	FBCCA
0	8.0	90.95	92.86	92.38	96.67
1	8.2	96.67	96.67	93.81	98.57
2	8.4	91.43	91.43	91.43	95.24
3	8.6	95.24	93.81	92.86	94.29
4	8.8	92.38	91.90	92.38	94.76
5	9.0	83.33	92.38	89.52	92.86
6	9.2	96.19	96.67	97.14	91.90
7	9.4	96.67	97.14	97.14	94.76
8	9.6	94.76	94.76	95.71	91.43
9	9.8	91.43	90.00	90.48	87.14
10	10.0	88.10	94.29	93.81	89.52
11	10.2	92.86	91.43	93.33	92.38
12	10.4	98.10	96.67	96.19	97.62
13	10.6	94.29	92.86	92.38	95.24
14	10.8	92.86	91.90	89.05	95.71
15	11.0	82.86	86.19	85.71	93.81
16	11.2	93.33	96.19	90.00	93.81
17	11.4	92.86	92.86	93.33	90.95
18	11.6	97.14	90.00	88.10	96.67
19	11.8	88.57	72.86	87.62	93.33
20	12.0	73.81	69.05	78.57	93.33
21	12.2	91.43	90.95	95.71	84.29
22	12.4	95.71	94.76	93.81	84.76
23	12.6	96.67	92.86	89.05	84.76
24	12.8	88.10	89.52	90.00	85.24
25	13.0	80.48	81.43	79.05	80.95
26	13.2	86.67	89.52	89.05	83.81
27	13.4	94.29	93.81	95.71	85.71
28	13.6	89.52	91.43	89.05	84.29
29	13.8	96.19	92.86	86.19	87.14
30	14.0	70.95	70.95	81.43	79.52
31	14.2	91.90	89.52	91.90	88.10
32	14.4	94.29	90.48	90.95	77.14
33	14.6	92.38	90.95	91.90	79.05
34	14.8	91.43	92.86	93.81	89.05

Class	Hz	MTSNet	SSVEPFormer	EEGNet	FBCCA
35	15.0	88.10	89.05	85.71	79.52
36	15.2	86.19	85.71	90.00	79.52
37	15.4	94.29	93.81	94.29	80.48
38	15.6	92.86	89.05	93.81	85.24
39	15.8	90.48	87.14	91.43	84.76

B.3 Language-Side Analyses

Text encoders and adaptation state. Figure B.1 compares paired Frozen/LoRA runs for BERT-base-uncased and EmbeddingGemma-300M (Devlin et al., 2019; Schechter Vera et al., 2025; Hu et al., 2022). A line joins matching MI folds, Speech seeds, SSVEP subject folds, or unseen-class folds. Figure 5d compares independently selected best configurations. In the paired runs, BERT Frozen SSVEP uses random training templates, while the other text-encoder conditions use fixed T0.

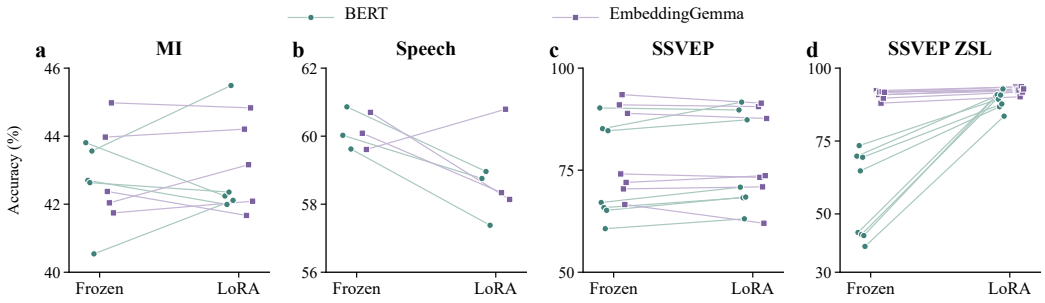


Figure B.1: Paired Frozen and LoRA comparisons for BERT-base-uncased and EmbeddingGemma-300M. Lines connect the matched run units specified in the text.

Held-out short templates. Table B.16 gives the five evaluation templates and their five-template ensemble under P0 and P2. P0 uses fixed template 0 during training; P2 randomly samples from five semantic training templates. For the ensemble, candidate-class logits are averaged across the five held-out templates before prediction. MI and Speech SDs are across subjects, the SSVEP closed-set SD is across seven subject folds, and the conditional five-class ZSL SD is across eight class folds. Speech P0 is seed 3407; for Speech P2, three seeds are first averaged within subject.

Table B.16: Accuracy (%, mean ± SD) under held-out short templates.

Task	Template	P0	P2
MI closed-set	S0	40.20 ± 9.66	42.65 ± 9.59
	S1	38.81 ± 9.91	42.43 ± 9.48
	S2	39.23 ± 9.74	42.54 ± 9.36
	S3	40.19 ± 9.71	42.52 ± 9.45
	S4	35.88 ± 10.06	42.32 ± 9.47
	Five-template ensemble	38.94 ± 9.96	42.51 ± 9.40
SSVEP closed-set	S0	79.27 ± 11.23	74.54 ± 12.48
	S1	79.32 ± 11.54	74.55 ± 12.39
	S2	79.49 ± 11.54	74.43 ± 12.64
	S3	79.39 ± 11.23	74.60 ± 12.15
	S4	79.33 ± 11.36	74.65 ± 12.52
	Five-template ensemble	79.43 ± 11.42	74.63 ± 12.29
SSVEP five-class ZSL	S0	90.88 ± 1.64	89.64 ± 1.55
	S1	90.80 ± 1.42	89.60 ± 1.61
	S2	90.80 ± 1.54	89.88 ± 1.57
	S3	91.06 ± 1.31	90.02 ± 1.34
	S4	90.80 ± 1.60	89.67 ± 1.63
	Five-template ensemble	90.82 ± 1.47	89.92 ± 1.67
Speech closed-set	S0	59.96 ± 15.77	60.08 ± 15.21
	S1	60.29 ± 15.84	60.16 ± 15.09
	S2	59.62 ± 15.77	59.48 ± 15.10
	S3	60.44 ± 15.91	59.82 ± 15.13

Task	Template	P0	P2
	S4	60.19 ± 15.85	59.89 ± 15.03
	Five-template ensemble	60.66 ± 16.12	60.06 ± 15.19

B.4 Additional Signal and Text Attribution Visualizations

The figures in this section use grouped Integrated Gradients for the specified class's unscaled cosine-matching score (Sundararajan et al., 2017). For each neural input, positive group attributions are divided by that input's largest positive group attribution; an input with no positive group remains zero. The normalized attributions are then averaged across the selected training trials without renormalizing the mean. Word attributions retain their signs during trial averaging, and words with positive mean attribution are shaded within each prompt. Each example uses its own relative colour scale.

For MI, each 5 s trial yields 21 overlapping 1 s windows. Each window's attributions are normalized over its 32 channels and five 200 ms time bins, then placed at its trial time; overlapping bins are averaged by their coverage count before 120 trials per class are combined. Speech uses ten channels and 30 bins of 100 ms over a 3 s trial. SSVEP uses 32 time-domain and 32 frequency-domain channel values for each 1 s input; both branches share the normalization denominator for that input. The selected SSVEP records are six original training trials per frequency, without IR synthesis or training masking.

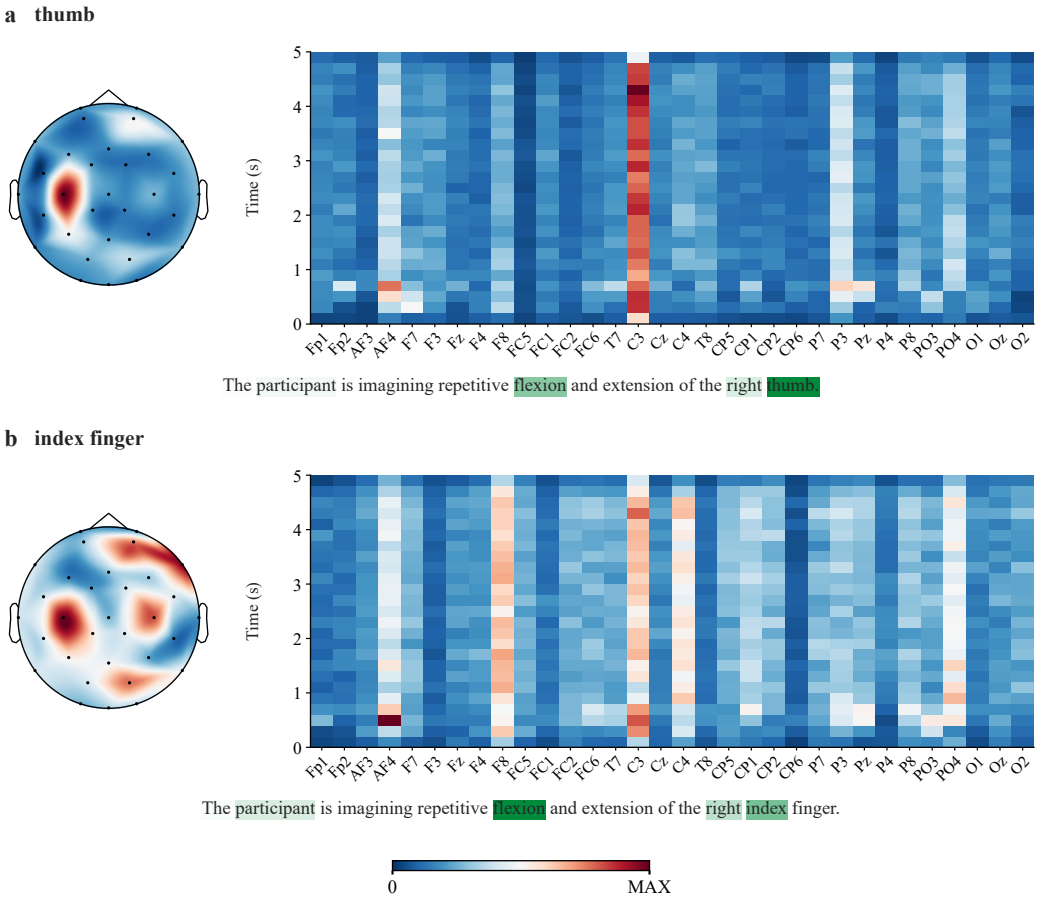


Figure B.2: MI training attribution for thumb and index, subject S01; 120 trials per class.

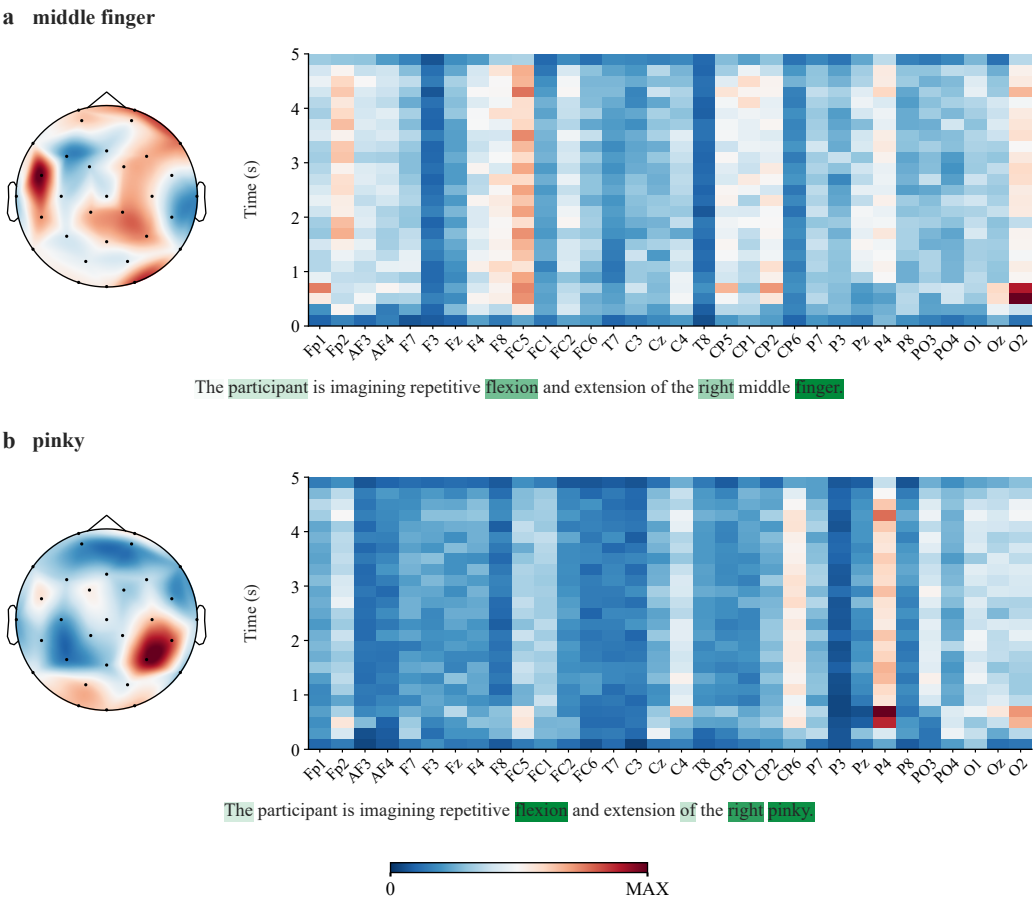


Figure B.3: MI training attribution for middle and pinky, subject S01; 120 trials per class.

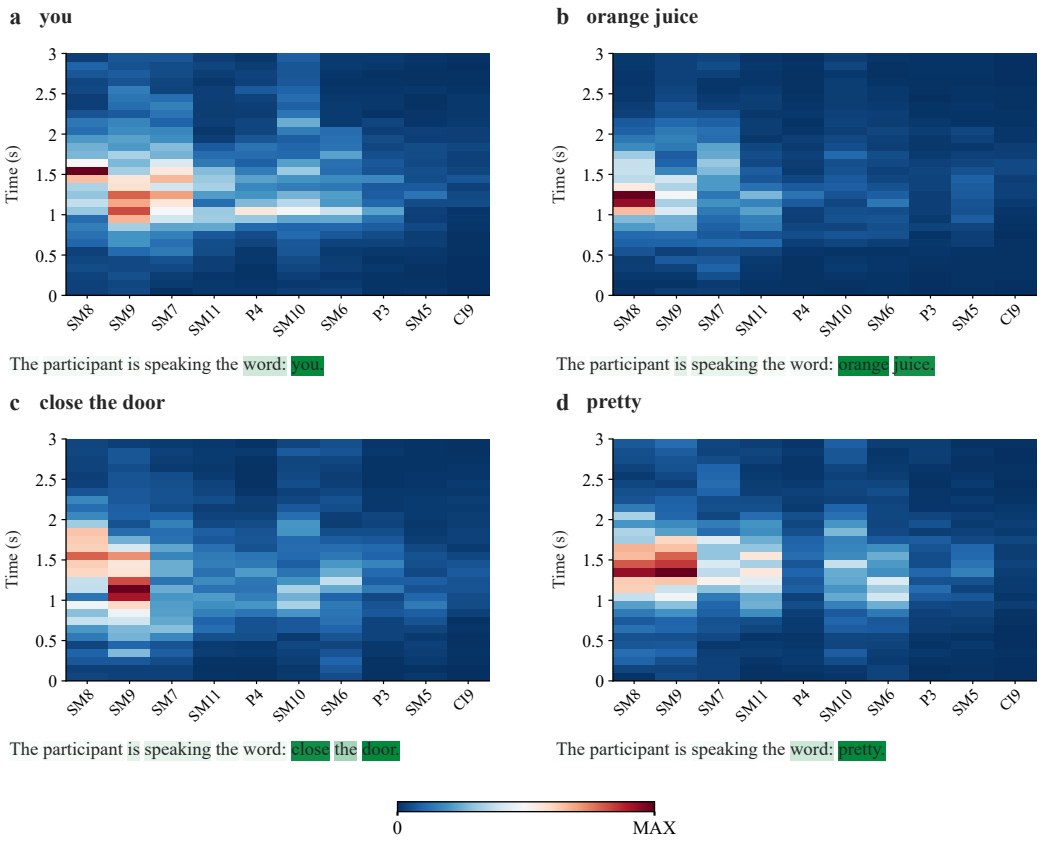


Figure B.4: Speech training attribution for four word classes, subject S001; class supports are 38, 40, 40, and 37.

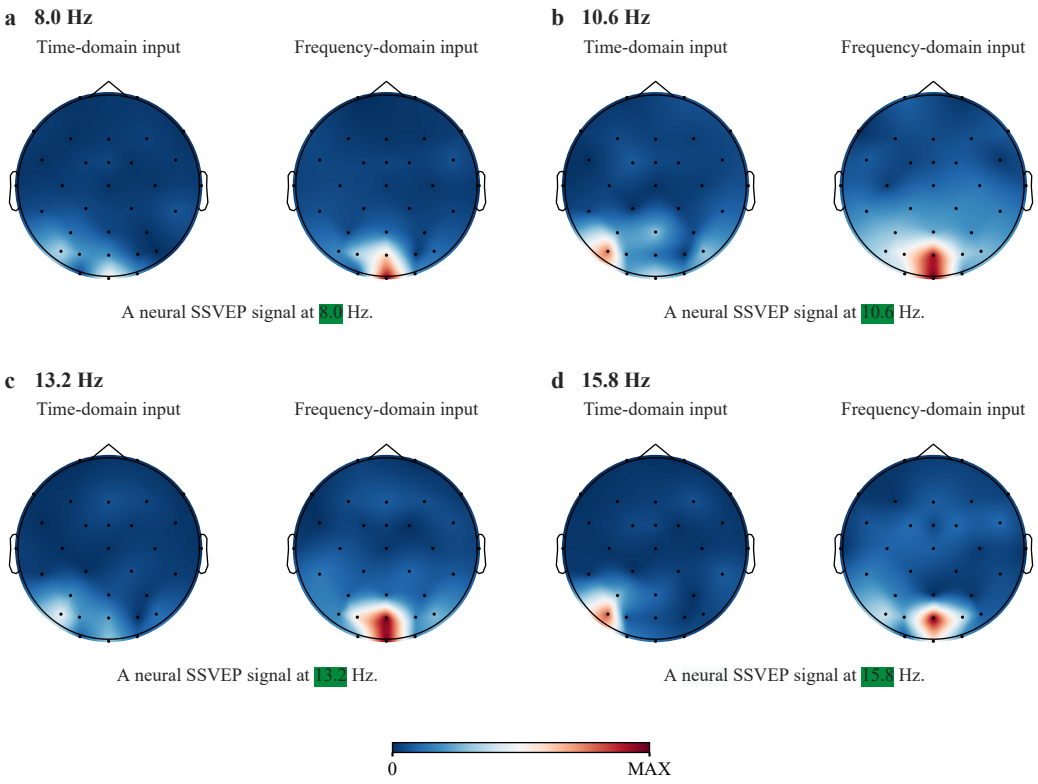


Figure B.5: SSVEP training attribution at 8.0, 10.6, 13.2, and 15.8 Hz, subject S01; six original trials per class.

B.5 Training Objectives, Directional Ablations, and Temporal Aggregation

Training objective and batch-size grid. Figure B.6 compares bidirectional InfoNCE, diagonal SigLIP, and class-aware SigLIP for MI, SSVEP, and Speech (Radford et al., 2021; Zhai et al., 2023). Each training-loss point is the minimum epoch-average loss for one fold or seed; accuracy uses the task-specific aggregation unit. The reference curves use the balanced-class formulas in Appendix C.2.

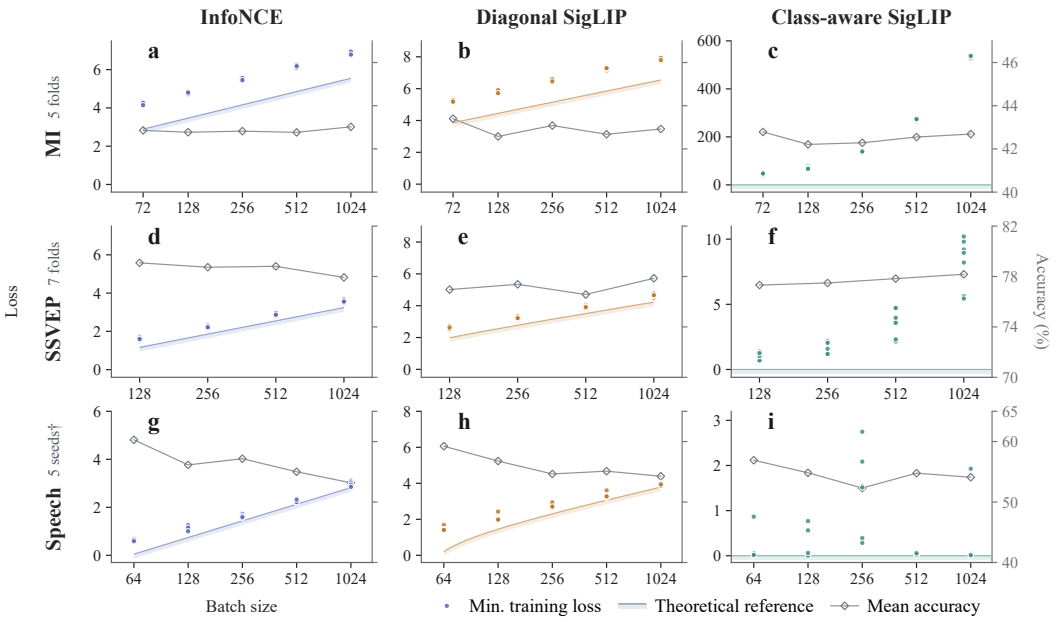


Figure B.6: Full training-objective by batch-size comparison across MI, SSVEP, and Speech. The left and right axes show training loss and classification accuracy, respectively; reference curves are theoretical loss baselines.

InfoNCE direction. Figure 8 compares neural-to-text (N2T-only), text-to-neural (T2N-only), and bidirectional InfoNCE for MI EEG-Conformer and Speech Du-IN, including their paired differences and 95% intervals.

Cumulative fusion over an MI trial. Figure B.7 follows the four-class LaS EEG-Conformer model as successive 1 s windows become available at a 0.2 s stride. The three rules use the same trial predictions; five held-out folds are pooled within each subject before the 21 subjects are averaged equally. With one window at 1.0 s, logit averaging, probability averaging, and majority voting coincide at 33.78% accuracy. The subject-mean curves generally rise as more overlapping windows are accumulated. At 5.0 s, logit averaging reaches 43.02%, probability averaging 42.68%, and majority voting 42.38%. The logit-averaging endpoint is the LaS EEG-Conformer four-class value in Table 1.

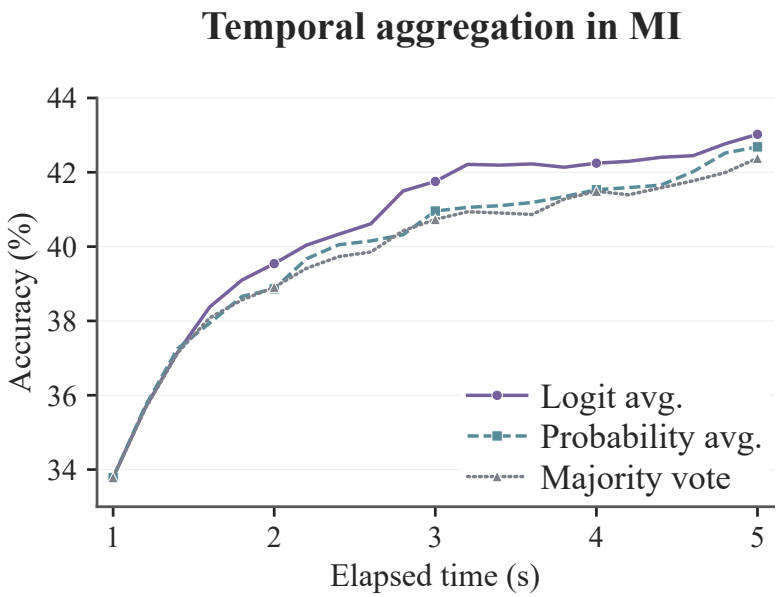


Figure B.7: Cumulative MI trial accuracy with logit averaging, probability averaging, and majority voting. Each time point uses all windows available up to that time.

B.6 Paired Comparisons and Effect Sizes

All comparisons use two-sided paired t-tests. In Table 1, each task–encoder combination forms a Holm comparison family: the three SSVEP window durations are corrected together, while each MI or Speech encoder has one comparison. In Figure 3, each encoder's family contains five comparisons: the four displayed subsets and the four-class result. Figure 5a uses a two-comparison family within each task (P1 versus P0 and P2 versus P0); Figure 5b uses a separate two-comparison family for unseen-class decoding. Figures 5c and 5d each use a four-comparison family across their task conditions. Figure 8 uses a two-comparison family within each task, comparing N2T-only and T2N-only with bidirectional InfoNCE.

Table B.17 reports the Table 1 comparisons. Differences are LaS minus CLS, paired by subject. Paired Cohen's d_z is the mean paired difference divided by its sample standard deviation. MI pairs 21 subjects after pooling each subject's five held-out folds; SSVEP pairs 35 subjects; Speech pairs 12 subjects after averaging the five seeds within each subject.

Table B.17: Paired statistical comparisons for Table 1. n is the number of subjects, Δ is the mean accuracy difference in percentage points, and p_{Holm} is corrected within the comparison families defined above.

Task	Encoder (window)	n	Δ (pp)	$t(df)$	p_{Holm}	d_z
MI	EEGNet	21	+0.28	0.33 (20)	0.747607843	+0.07
MI	DeepConvNet	21	+1.82	2.74 (20)	0.0125586274	+0.60
MI	EEG-Conformer	21	+1.36	3.23 (20)	0.00423557447	+0.70
SSVEP	EEGNet (0.6 s)	35	+0.21	0.35 (34)	0.727554094	+0.06
SSVEP	EEGNet (0.8 s)	35	−0.65	−1.04 (34)	0.613050642	−0.18
SSVEP	EEGNet (1.0 s)	35	−1.76	−2.02 (34)	0.154051070	−0.34
SSVEP	MTSNet (0.6 s)	35	+2.24	3.83 (34)	0.001597153525	+0.65
SSVEP	MTSNet (0.8 s)	35	+1.89	2.05 (34)	0.0953715778	+0.35
SSVEP	MTSNet (1.0 s)	35	+1.39	1.67 (34)	0.103955385	+0.28
SSVEP	SSVEPFormer (0.6 s)	35	+2.48	3.14 (34)	0.0103493646	+0.53
SSVEP	SSVEPFormer (0.8 s)	35	+2.21	2.30 (34)	0.0330535594	+0.39
SSVEP	SSVEPFormer (1.0 s)	35	+2.27	2.52 (34)	0.0330535594	+0.43
Speech	EEGNet	12	+8.33	5.95 (11)	$9.59414678 \times 10^{-5}$	+1.72
Speech	CNN-BiGRU	12	+17.15	9.21 (11)	$1.66659570 \times 10^{-6}$	+2.66
Speech	EEG-Conformer	12	+6.21	5.04 (11)	0.000380624783	+1.45
Speech	Du-IN	12	+5.22	4.54 (11)	0.000837161674	+1.31

C Theory and Attribution Methods

C.1 Neural–Text Alignment: Properties and Correspondence

Supervision paths and notation. Let $\mathcal{S} = \{1, \dots, C\}$ be the training classes, $C \geq 2$, and let (X_i, y_i) denote a neural trial and its class label. We use the row-vector convention of Section 3: $N_f(X_i) \in \mathbb{R}^{1\times d_f}$ is the neural encoder output. A task-specific CLS head g_ψ produces C class logits and minimizes

$$Z_{\mathrm{CLS}}(X_i) = g_\psi(N_f(X_i)), \qquad p_{ic}^{\mathrm{CLS}} = \frac{e^{Z_{\mathrm{CLS}}(X_i)_c}}{\sum_{r=1}^{C} e^{Z_{\mathrm{CLS}}(X_i)_r}}, \qquad \ell_{\mathrm{CLS}}(X_i, y_i) = -\log p_{i,y_i}^{\mathrm{CLS}}. \tag{6}$$

The head can be linear or nonlinear. Both CLS and LaS use class labels; LaS uses them to pair neural trials with class prompts.

For the complete prompt P_c, the text encoder gives $T_{f,c} \in \mathbb{R}^{1\times d_t}$. Bias-free, right-multiplying projections $W_n \in \mathbb{R}^{d_f\times d_{\mathrm{proj}}}$ and $W_t \in \mathbb{R}^{d_t\times d_{\mathrm{proj}}}$ define

$$\begin{aligned} z_{\mathrm{n,proj}}^{(i)} &= \nu\big(N_f(X_i)W_n\big), \quad z_{\mathrm{t,proj}}^{(c)} = \nu(T_{f,c}W_t), \quad \nu(v) = v/\|v\|_2, \\ s_c(X_i) &= z_{\mathrm{n,proj}}^{(i)}\big(z_{\mathrm{t,proj}}^{(c)}\big)^\top, \qquad a_{ic} = s_c(X_i)/\tau, \qquad \tau^{-1} = e^t > 0. \end{aligned} \tag{7}$$

The projected vectors must be finite and nonzero. The normalized text vector is the class prototype, while s_c and a_{ic} denote the unscaled cosine score and temperature-scaled logit, respectively. For a paired batch of B entries, $L_{ij} = a_{i,y_j}$ and

$$\begin{aligned} \mathcal{L}_{\mathrm{N2T}} &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{a_{i,y_i}}}{\sum_{j=1}^{B} e^{a_{i,y_j}}}, \\ \mathcal{L}_{\mathrm{T2N}} &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{a_{i,y_i}}}{\sum_{j=1}^{B} e^{a_{j,y_i}}}, \quad \mathcal{L}_{\mathrm{SA}} = \tfrac{1}{2}(\mathcal{L}_{\mathrm{N2T}} + \mathcal{L}_{\mathrm{T2N}}). \end{aligned} \tag{8}$$

This is the bidirectional InfoNCE objective (Radford et al., 2021). Neural and text encoder parameters are separate from those of CLS; each paired architecture is trained independently.

The repeated-prompt arguments assume one fixed prompt per class and deterministic text forwards, so all copies of a class text have the same embedding and prototype. Same-class neural trials may have different features. In a paired batch, $\mathcal{G}_c = \{i : y_i = c\}$, $n_c = |\mathcal{G}_c|$, $\mathcal{S}_b = \{c : n_c > 0\}$ and $B = \sum_{c\in\mathcal{S}_b} n_c$. Summing over $\mathcal{S}_b$ is equivalent to indexing the represented classes consecutively; it avoids treating absent classes as paired entries. When all C training classes occur equally often, $n_c = m$ and $B = mC$.

All logarithms are natural. InfoNCE averages its two directions; SigLIP sums all B^2 pairwise terms and divides by B. Comparisons of pair repetition or label rules hold the compared representations, temperature and bias fixed. Additional assumptions, including fixed mean scores and full row rank, are stated for the relevant results below.

Effective class weights and linear decisions. Define the effective reading weight $q_c \in \mathbb{R}^{d_f\times 1}$ by

$$q_c = W_n\big(z_{\mathrm{t,proj}}^{(c)}\big)^\top, \qquad \frac{s_c(X_i)}{\tau} = \frac{N_f(X_i)q_c}{\tau\|N_f(X_i)W_n\|_2}. \tag{9}$$

Proposition C.1 (Prototype-generated ordering)**.** *For fixed parameters and a finite candidate set $\mathcal{C}$, on the domain $N_f(X_i)W_n \neq 0$,*

$$\begin{aligned} \arg\max_{c\in\mathcal{C}} s_c(X_i) &= \arg\max_{c\in\mathcal{C}} N_f(X_i)q_c, \\ s_c(X_i) = s_{c'}(X_i) &\iff z_{\mathrm{n,proj}}^{(i)}\big(z_{\mathrm{t,proj}}^{(c)} - z_{\mathrm{t,proj}}^{(c')}\big)^\top = 0 \\ &\iff N_f(X_i)(q_c - q_{c'}) = 0. \end{aligned} \tag{10}$$

Proof. Substituting the normalized neural projection into the matching score yields Equation (9). Its denominator is a positive scalar shared by every candidate for this input, so it preserves all pairwise orderings. The tie relations follow by subtraction.

If $q_c \neq q_{c'}$, their equal-score set is a hyperplane through the origin in neural-feature space, restricted to the valid projection domain. If the weights coincide, the two classes tie throughout that domain. A pairwise tie becomes a prediction boundary only where neither score is below another candidate's. The encoder from the raw signal to N_f may remain nonlinear. The common denominator cancels for rankings, but softmax probabilities and training losses use the complete logits.

Prototype distances and score variation. For fixed neural features, suppose both compared text projections have norm at least $\rho > 0$. With $\|W_t\|_2$ denoting the spectral norm,

$$|s_c(X_i) - s_{c'}(X_i)| \leq \left\| z^{(c)}_{\mathrm{t,proj}} - z^{(c')}_{\mathrm{t,proj}} \right\|_2 \leq \frac{2\|W_t\|_2}{\rho} \|T_{f,c} - T_{f,c'}\|_2. \tag{11}$$

Proof. The first inequality follows from the unit norm of the neural vector and Cauchy–Schwarz. Let $v = T_{f,c}W_t$ and $w = T_{f,c'}W_t$. The triangle and reverse-triangle inequalities give

$$\|\nu(v) - \nu(w)\|_2 \leq \frac{\|v - w\|_2 + |\|w\|_2 - \|v\|_2|}{\|v\|_2} \leq \frac{2\|v - w\|_2}{\rho}. \tag{12}$$

Finally, $\|(T_{f,c} - T_{f,c'})W_t\|_2 \leq \|T_{f,c} - T_{f,c'}\|_2 \|W_t\|_2$. Coincident prototypes therefore give identical scores. Positive rescaling of W_t rescales both ρ and $\|W_t\|_2$, preserving the prototypes and their ratio in the bound.

N2T decomposition by class. Grouping identical class-text copies in the complete B-entry denominator gives, for each neural query,

$$\begin{aligned} \ell_i^{\mathrm{N2T}} &= -a_{i,y_i} + \log \sum_{c \in \mathcal{S}_b} n_c e^{a_{ic}} \\ &= \log n_{y_i} + \log \left[1 + \sum_{\substack{c \in \mathcal{S}_b \\ c \neq y_i}} \frac{n_c}{n_{y_i}} e^{a_{ic} - a_{i,y_i}} \right]. \end{aligned} \tag{13}$$

Thus, multiplicities weight competing classes through n_c/n_{y_i}. For a batch containing all C classes m times, the denominator is $m \sum_{c=1}^{C} e^{a_{ic}}$, so

$$\ell_i^{\mathrm{N2T}} = \underbrace{-a_{i,y_i} + \log \sum_{c=1}^{C} e^{a_{ic}}}_{\text{class cross-entropy on matching logits}} + \log m. \tag{14}$$

For this fixed query, repetition adds a parameter-independent constant and leaves class-score gradients unchanged. Enlarging the batch with new neural trials also introduces new query terms. Class cross-entropy and this decomposition allow same-class trials to have different correct-class scores.

T2N matching balance along the corresponding text-prototype direction. Every copy of class text c has denominator $\sum_{j=1}^{B} e^{a_{jc}}$, while successive diagonal targets select different $i \in \mathcal{G}_c$. Averaging their negative log probabilities yields

$$\bar{a}_c = \frac{1}{n_c} \sum_{i \in \mathcal{G}_c} a_{ic}, \qquad \bar{\ell}_c^{\mathrm{T2N}} = -\bar{a}_c + \log \sum_{j=1}^{B} e^{a_{jc}}, \qquad \mathcal{L}_{\mathrm{T2N}} = \frac{1}{B} \sum_{c \in \mathcal{S}_b} n_c \bar{\ell}_c^{\mathrm{T2N}}. \tag{15}$$

Proposition C.2 (Within-class matching balance)**.** *Fix the mean $\bar{a}_c$ and every other-class score a_{jc}, $j \notin \mathcal{G}_c$. The class loss $\bar{\ell}_c^{\mathrm{T2N}}$ is minimized when all within-class scores equal $\bar{a}_c$.*

The repeated text queries have the same denominator. Under these conditions, their average satisfies

$$\begin{aligned} \bar{\ell}_c^{\mathrm{T2N}} &= \log \left[\sum_{i \in \mathcal{G}_c} e^{a_{ic} - \bar{a}_c} + \sum_{j \notin \mathcal{G}_c} e^{a_{jc} - \bar{a}_c} \right] \\ &\geq \log \left[n_c + \sum_{j \notin \mathcal{G}_c} e^{a_{jc} - \bar{a}_c} \right]. \end{aligned} \tag{16}$$

Proof. The within-class deviations sum to zero. Strict convexity of the exponential gives $\sum_{i\in\mathcal{G}_c} e^{a_{ic}-\bar{a}_c} \geq n_c$, with equality exactly when every deviation is zero. The other-class sum remains fixed under the stated conditions, proving that balanced same-class scores minimize this class loss.

For $n_c > 1$, this balance provides implicit regularization toward within-class aggregation of matching responses along the corresponding text-prototype direction, under the fixed-score conditions above. Other feature directions may retain trial differences; at $n_c = 1$, there are no within-class score differences to balance.

In a two-trial same-class example, one identical text copy targets the first trial and treats the second as an off-diagonal competitor; the other reverses these roles. The shared denominator couples their scores. This class-query coupling and the same-class false negatives arise from the same repeated pairing structure. A fixed prompt permutation preserves a stable common reference within each class, although it changes the assigned descriptions.

Compensating a fixed seen-prompt permutation. Stack the C seen text embeddings in class order. Let $\Pi \in \mathbb{R}^{C\times C}$ be a permutation matrix, fixed throughout training, so the assigned embedding matrix is $\Pi T_f^{\mathcal{S}}$:

$$T_f^{\mathcal{S}} = \begin{bmatrix} T_{f,1} \\ \vdots \\ T_{f,C} \end{bmatrix} \in \mathbb{R}^{C\times d_t}. \tag{17}$$

Proposition C.3 (Seen-prompt permutation compensation)**.** *Assume* $\operatorname{rank}(T_f^{\mathcal{S}}) = C$ *and all original projected prototypes are finite and nonzero. Full row rank requires* $d_t \geq C$*. For any original* W_t*, keeping the neural branch fixed, a projection of the same shape restores all seen prototypes and pairing scores:*

$$W_t^{\Pi} = (\Pi T_f^{\mathcal{S}})^{\dagger} T_f^{\mathcal{S}} W_t, \qquad \Pi T_f^{\mathcal{S}} W_t^{\Pi} = T_f^{\mathcal{S}} W_t, \tag{18}$$

Here † denotes the Moore–Penrose pseudoinverse.

Proof. Multiplication by a permutation preserves row rank, hence $(\Pi T_f^{\mathcal{S}})(\Pi T_f^{\mathcal{S}})^{\dagger} = I_C$. Substituting the proposed projection proves the second identity. Consequently, each assigned seen text has the same projected vector and normalized prototype as before. Keeping the neural branch and temperature unchanged preserves every seen pairing score and the data loss.

Seen correspondence and unseen extension. Unseen classes $\mathcal{U}$ are disjoint from $\mathcal{S}$ and use their true descriptions at prediction. Suppose an unseen embedding lies in the span of the seen embeddings,

$$T_{f,u} = \beta_u^{\top} T_f^{\mathcal{S}}, \qquad \beta_u \in \mathbb{R}^C. \tag{19}$$

Under this span assumption, its projected vector is the same linear combination of the projected seen embeddings. The compensation in Equation (18) restores the *assigned, permuted* seen correspondence, while the neural relationships remain fixed. Extension to the true unseen description also depends on the combination coefficients β_u. Thus, fitting the seen classes and extending to unseen classes impose distinct correspondence requirements.

Successful transfer also requires compatibility between the neural distinctions available in a task and the relations in its class descriptions. In SSVEP, stimulus frequency and the EEG frequency response provide a physical connection to the numerical frequency in a prompt. The model receives frequency text, without an extra hand-designed attribute coordinate. This physical connection motivates the conditional evaluation.

Candidate competition. Unseen-only prediction uses $\mathcal{C} = \mathcal{U}$, whereas adding seen candidates uses $\mathcal{C} = \mathcal{S} \cup \mathcal{U}$. With parameters fixed, both candidate sets preserve each neural representation and class score. If the highest seen score exceeds all unseen scores, the joint prediction selects a seen class even when the true class ranks first within $\mathcal{U}$. Joint prediction therefore also compares the leading unseen score with the seen scores.

Text-representation scope. The preceding score bound concerns text embeddings after the chosen text encoder and readout, followed by the learned projection. A contextual BERT [CLS] vector (Devlin et al., 2019) and an embedding model trained for sentence comparison (Schechter Vera et al., 2025) have different training objectives and readouts. Changing the encoder, readout, or prompt changes $T_{f,c}$ and may change the resulting prototypes; Equation (11) bounds score variation for fixed neural features under its stated norm condition. The permutation construction establishes an available parameter solution for the seen classes.

C.2 Repeated-Prompt Training Objectives

Pair counts. The arguments below use one fixed prompt per class and a deterministic text forward, so same-class text copies have identical embeddings. Diagonal labels mark $i = j$ positive; class-aware labels mark every same-class pair positive. Among off-diagonal entries, the diagonal rule treats $y_i = y_j$ as negative:

$$M_{ij}^{\mathrm{diag}} = \mathbf{1}[i = j], \qquad M_{ij}^{\mathrm{class}} = \mathbf{1}[y_i = y_j], \qquad N_{\mathrm{FN}} = \sum_{c \in \mathcal{S}_b} n_c(n_c - 1). \tag{20}$$

Each neural row contains one diagonal positive, $n_{y_i} - 1$ same-class false negatives and $B - n_{y_i}$ other-class negatives. For $B = mC$, the latter two counts become $m - 1$ and $m(C - 1)$. Class-aware labels turn all same-class pairs positive; the $m = 1$ baseline has no same-class off-diagonal pairs.

InfoNCE class aggregation. Grouping the identical text copies in the full paired denominator gives

$$\begin{aligned} \ell_i^{\mathrm{N2T}} &= -\log \frac{e^{a_{i,y_i}}}{\sum_{j=1}^{B} e^{a_{i,y_j}}} = -a_{i,y_i} + \log \sum_{c \in \mathcal{S}_b} n_c e^{a_{ic}} \\ &= \log n_{y_i} + \log \left[1 + \sum_{\substack{c \in \mathcal{S}_b \\ c \neq y_i}} \frac{n_c}{n_{y_i}} e^{a_{ic} - a_{i,y_i}} \right]. \end{aligned} \tag{21}$$

For $n_c = m$ across all C classes, the full denominator and loss reduce to

$$\begin{aligned} \sum_{j=1}^{B} e^{a_{i,y_j}} &= m \sum_{c=1}^{C} e^{a_{ic}}, \\ \ell_i^{\mathrm{N2T}} &= \underbrace{-a_{i,y_i} + \log \sum_{c=1}^{C} e^{a_{ic}}}_{\text{class cross-entropy on matching logits}} + \log m. \end{aligned} \tag{22}$$

For a fixed neural query, repetition adds a parameter-independent constant and leaves class-score gradients unchanged. New neural samples contribute new queries. In T2N, copies of one class text share a denominator; Equation (15) and the conditional Jensen bound in Equation (16) give their class-averaged response constraint. Bidirectional InfoNCE averages the two full directional losses.

SigLIP label conflicts and gradients. With shared bias b_{sig}, define $\xi_{ic} = a_{ic} + b_{\mathrm{sig}}$, $\ell_+(\xi) = \log(1 + e^{-\xi})$, and $\ell_-(\xi) = \log(1 + e^{\xi})$. Both SigLIP variants (Zhai et al., 2023) sum over all B^2 pairs and divide by B:

$$\mathcal{L}_{\mathrm{Sig}}(M) = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} \left[M_{ij} \ell_+(\xi_{i,y_j}) + (1 - M_{ij}) \ell_-(\xi_{i,y_j}) \right]. \tag{23}$$

Grouping the repeated texts gives row losses

$$\begin{aligned} \ell_i^{\mathrm{diag}} &= \ell_+(\xi_{i,y_i}) + (n_{y_i} - 1)\ell_-(\xi_{i,y_i}) + \sum_{\substack{c \in \mathcal{S}_b \\ c \neq y_i}} n_c \ell_-(\xi_{ic}), \\ \ell_i^{\mathrm{class}} &= n_{y_i} \ell_+(\xi_{i,y_i}) + \sum_{\substack{c \in \mathcal{S}_b \\ c \neq y_i}} n_c \ell_-(\xi_{ic}). \end{aligned} \tag{24}$$

The full losses average these B row losses. Since $1 + e^{\xi} = e^{\xi}(1 + e^{-\xi})$, changing a same-class off-diagonal label changes its loss by $\ell_-(\xi) - \ell_+(\xi) = \xi$. Therefore, at the same parameter state,

$$\mathcal{L}_{\mathrm{Sig}}^{\mathrm{diag}} - \mathcal{L}_{\mathrm{Sig}}^{\mathrm{class}} = \frac{1}{B}\sum_{i=1}^{B}(n_{y_i} - 1)\xi_{i,y_i}. \tag{25}$$

Let $\sigma_{\mathrm{sig}}(\xi) = (1 + e^{-\xi})^{-1}$. Using $\ell'_+ = -(1 - \sigma_{\mathrm{sig}})$ and $\ell'_- = \sigma_{\mathrm{sig}}$ gives

$$\frac{\partial \ell_i^{\mathrm{diag}}}{\partial \xi_{i,y_i}} = n_{y_i}\sigma_{\mathrm{sig}}(\xi_{i,y_i}) - 1, \qquad \frac{\partial \ell_i^{\mathrm{class}}}{\partial \xi_{i,y_i}} = -n_{y_i}\big[1 - \sigma_{\mathrm{sig}}(\xi_{i,y_i})\big]. \tag{26}$$

The derivatives of the complete loss include a further factor $1/B$. Diagonal supervision asks one positive term to raise the correct-class score while $n_{y_i} - 1$ identical-score negative terms lower it. Class-aware supervision removes this conflict and increases the number of positive terms to n_{y_i}. The loss difference in Equation (25) can be positive, zero, or negative.

The normalization retains a repetition-dependent scale. If the same representation of each class is copied m times, class-aware pair terms multiply by m^2 but the denominator only by m, so the full loss is m times its one-copy baseline. These pair terms add independently, whereas T2N couples trials through a shared softmax denominator. Removing label conflicts also changes positive-pair weights.

Bidirectional InfoNCE bound. The second logarithm in Equation (13) is nonnegative, hence $\ell_i^{\mathrm{N2T}} \geq \log n_{y_i}$. Equation (16) likewise gives $\ell_c^{\mathrm{T2N}} \geq \log n_c$. Aggregating and averaging directions yields

$$\begin{aligned}
\mathcal{L}_{\mathrm{N2T}} &\geq \frac{1}{B}\sum_{i=1}^{B}\log n_{y_i} = \frac{1}{B}\sum_{c\in\mathcal{S}_b} n_c \log n_c, \\
\mathcal{L}_{\mathrm{T2N}} &\geq \frac{1}{B}\sum_{c\in\mathcal{S}_b} n_c \log n_c, \\
\mathcal{L}_{\mathrm{SA}} &\geq \frac{1}{B}\sum_{c\in\mathcal{S}_b} n_c \log n_c.
\end{aligned} \tag{27}$$

With $n_c = m$ for all C classes, this becomes $\log m$. Approaching the bound requires other-class exponential terms to become negligible relative to same-class terms; T2N additionally requires balanced same-class scores. Identical texts cannot distinguish the m pairing positions even when classes are separated.

Diagonal SigLIP bound. For a class count $n > 1$, the same-class contribution to a row is $f_n(\xi) = \ell_+(\xi) + (n-1)\ell_-(\xi)$. Treating the logit temporarily as a free scalar,

$$\begin{aligned}
f_n'(\xi) &= n\sigma_{\mathrm{sig}}(\xi) - 1 = 0 \quad \Longrightarrow \quad \xi^\star = -\log(n-1), \\
f_n''(\xi) &= n\sigma_{\mathrm{sig}}(\xi)\big[1 - \sigma_{\mathrm{sig}}(\xi)\big] > 0, \\
f_n(\xi^\star) &= \log n + (n-1)\log\frac{n}{n-1} = n\log n - (n-1)\log(n-1).
\end{aligned} \tag{28}$$

The stationary point is therefore the unique minimum. Other-class terms are nonnegative; substituting $n = n_{y_i}$ and averaging the row bounds gives

$$\mathcal{L}_{\mathrm{Sig}}^{\mathrm{diag}} \geq \frac{1}{B}\sum_{c\in\mathcal{S}_b} n_c\left[n_c \log n_c - (n_c - 1)\log(n_c - 1)\right]. \tag{29}$$

For $n_c = m$, this reduces to $m\log m - (m-1)\log(m-1)$. At $n = 1$ the same-class contribution has only a positive term and infimum zero, consistent with $0\log 0 = 0$. Approaching the full bound additionally requires all other-class negative losses to vanish and the same-class row minima to be jointly realizable by the shared neural–text model.

Class-aware SigLIP bound. All pair losses are nonnegative, so $\mathcal{L}_{\text{Sig}}^{\text{class}} \geq 0$. If positive and negative pairs can be simultaneously separated with unbounded logit magnitudes, positive logits can tend to $+\infty$ and negative logits to $-\infty$, making the loss approach zero. Under this separation condition, class-aware labels remove the positive floor caused by conflicting labels on identical copies.

These bounds use the specified objectives and actual paired-batch counts. Their attainability depends on the separation and shared-logit conditions above. Figure 7 continuously extends the balanced formulas to $m = B/C$; fractional m serves as a plotting reference, while literal balanced batches have integer multiplicities. The empirical points are each run's minimum epoch-average full training loss.

C.3 Dual-Sided Attribution for Class Matching

Score and neural-input path. For a trained LaS model and a specified class prompt t, both attribution sides concern the unscaled cosine matching score

$$s(x,t) = \left\langle \frac{P_E(f_E(x))}{\|P_E(f_E(x))\|_2}, \frac{P_T(f_T(t))}{\|P_T(f_T(t))\|_2} \right\rangle. \tag{30}$$

Here x denotes the complete preprocessed model input, including all input branches, and P_E, P_T include the respective projections. This is $s_c(X)$ from Section 3, before division by τ. For neural attribution, the class text is fixed. All neural-input components move together from the zero model-input baseline to x along $x(\alpha) = \alpha x$, $0 \leq \alpha \leq 1$. The zero baseline denotes zero in the preprocessed input space; it has no direct physiological interpretation.

Partition the input into disjoint groups g. For a channel c and a temporal group p, grouped Integrated Gradients (IG) (Sundararajan et al., 2017) gives

$$A_{c,p} = \sum_{u \in p} x_{c,u} \int_0^1 \left. \frac{\partial s(\tilde{x},t)}{\partial \tilde{x}_{c,u}} \right|_{\tilde{x}=\alpha x} d\alpha. \tag{31}$$

The implementation integrates derivatives with respect to a gate shared by the components of each group, using Gauss–Legendre quadrature. If $x(\boldsymbol{\gamma})$ multiplies group g by γ_g, then $\partial s(x(\boldsymbol{\gamma}),t)/\partial \gamma_g$ already contains the factor $x_{c,u}$ in Equation (31); the integrated gate derivative is the group attribution itself. Raw A_g retains its sign. Provided the endpoint scores are defined, the summed groups are checked numerically against

$$\sum_g A_g \simeq s(x,t) - s(0,t). \tag{32}$$

A positive value raises the specified score along this path; a negative value lowers it. These are input-to-score attributions for the trained model.

Input groups and physical coordinates. For MI, each one-second input window is divided into five 200-ms groups for each of 32 channels, giving 32×5 groups. A five-second trial has 21 overlapping windows at 0.2-s steps. Window-group attributions are returned to the trial's physical timeline using each recorded window start. Let $m_{i,b}$ count the windows covering physical 200-ms bin b in trial i. For the signed raw attributions,

$$B_{i,c,b} = \frac{1}{m_{i,b}} \sum_{(w,p) \mapsto b} A_{i,w,c,p}, \qquad H_{c,b} = \frac{1}{N} \sum_{i=1}^{N} B_{i,c,b}. \tag{33}$$

The 32×25 array H is the signed trial-averaged channel-by-time map. MI's scalp values are its means across the 25 physical time bins.

For Speech, each three-second trial has 10 physical channels and 30 successive 100-ms groups per channel. Each resulting 10×30 attribution array is averaged with equal trial weights, without overlapping-window reconstruction. For SSVEP, one input contains 32 time-branch and 32 frequency-branch channel groups. Both branches follow the same α path; trial means are computed separately for the two branches and mapped to two scalp plots. Frequency-branch channel attributions are not assigned a physical time bin. Topomap interpolation changes the display between electrodes, not the channel values passed to it.

Conditional attribution of prompt words. For a displayed word j, set all its token embeddings to zero and restore only that word along a separate interpolation path. Other words, any text prefix, positional information and the attention mask remain fixed. For MI, the neural anchor is the mean of unit-normalized projected window features within each trial and then across the N selected trials:

$$\bar{z}_E = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{W_i}\sum_{w=1}^{W_i}\hat{z}_E(x_{i,w}), \qquad I_j = \int_0^1 \frac{d}{d\alpha}\left[\bar{z}_E\,\hat{z}_T(e_{\neg j}+\alpha e_j)^\top\right] d\alpha. \tag{34}$$

The mean $\bar{z}_E$ is not normalized again. Speech and SSVEP use equal weights for their selected trials in the corresponding neural anchor. The text forward uses the checkpoint's text encoder and projection, including LoRA when that checkpoint contains it. Subject to numerical integration error, I_j is the full-prompt score minus the score obtained by zeroing word j alone. Each word has its own baseline. Word scores are signed and do not sum to a decomposition of the whole prompt.

From raw attributions to Figure 6. Raw signed arrays, endpoint scores and integration residuals are retained before display. In a raw-contribution view, signed values can be aggregated as in Equation (33) and negative values clipped only when applying a positive-contribution color scale. Such a view retains score-change units. The relative positive-contribution view used in Figure 6 applies the following transform to every complete neural model input i *before* temporal or trial aggregation:

$$R_i(g) = \begin{cases} \dfrac{\mathrm{ReLU}(A_i(g))}{\max_h \mathrm{ReLU}(A_i(h))}, & \max_h \mathrm{ReLU}(A_i(h)) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{35}$$

The denominator spans all 160 groups of one MI window, all 300 groups of one Speech trial, or the 64 groups across both SSVEP branches of one trial. An input with no positive attribution remains all zero and remains in the average. For MI, the transformed window values follow the overlap and trial averaging of Equation (33); its topomap averages the resulting 25 time bins. Speech averages transformed trial arrays. SSVEP averages the transformed time and frequency groups separately within each stimulus class. No mean map is normalized again. The resulting dimensionless values show relative positive contributions; Equation (32) applies to the signed raw attributions.

Word highlighting follows a different display rule. Signed conditional word scores are averaged at the relevant window and trial levels; the displayed intensity is $\max(I_j, 0)/M_{\mathrm{word}}$ when the separate prompt-level color limit M_{word} is positive. If all averaged word scores are nonpositive, all words remain white. The neural maps use the `RdBu_r` scale from zero to the indicated MAX: zero and clipped negative values are blue, intermediate positive values approach white, and the largest displayed positive values are red. Raw-contribution plots may share a limit within a specified comparison; relative maps and topomaps may have separate panel limits. Comparisons of color therefore use each panel's MAX, units and processing mode.

D Further Analysis of SSVEP

D.1 SSVEP Responses and Short-Window Evaluation

An SSVEP is an oscillatory visual response whose frequency follows a periodic stimulus (Vialatte et al., 2010). The benchmark used here presents 40 targets distinguished by frequency and phase, with 5 s of stimulation per trial (Wang et al., 2017). Decoding commonly uses a short segment from such a trial. The benchmark study reported a mean visual latency of approximately 137 ms and recommended a shift of approximately 140 ms when extracting analysis windows (Wang et al., 2017). We use $t_0 = 0.14$ s after stimulus onset as a common analysis origin; physiological latency may vary among subjects. An initial window (IW) of length L spans $[t_0, t_0 + L)$; an onset offset δ shifts it to $[t_0 + \delta, t_0 + \delta + L)$. Window length and onset offset are varied separately below.

D.2 Window Augmentation and Closed-Set Results

Let $x_{u,c,r}(t)$ denote all channels of trial r for subject u and class c. The valid stimulus interval ends at $t_1 = 5.00$ s. The following operations apply to training signals; validation and test windows are specified for each result.

Random-window augmentation (RW). RW samples a subject, class, and trial from the training set, then selects one start time within that trial. Every channel contributes the same contiguous segment of length L:

$$\widetilde{x}^{\mathrm{RW}} = x_{u,c,r}[a : a + L], \qquad a \sim \mathcal{U}[t_0, t_1 - L]. \tag{1}$$

The upper bound keeps the window inside the valid interval.

Inter-trial recombination (IR). For one subject and class, IR extracts $[t_0, t_0 + L)$ from each available training trial. It applies common segment boundaries $0 = \tau_0 < \tau_1 < \cdots < \tau_J = L$, chooses a trial r_j for each segment, and concatenates the selected pieces:

$$\widetilde{x}^{\mathrm{IR}} = \bigoplus_{j=1}^{J} x_{u,c,r_j}[t_0 + \tau_{j-1} : t_0 + \tau_j]. \tag{2}$$

The segments retain their temporal positions while their source trials change. This follows the time-domain segmentation and recombination idea of Lotte (Lotte, 2015); the subject, class, and initial-window restrictions above specify our SSVEP implementation.

Temporal masking. Given a window $x \in \mathbb{R}^{H \times T}$ and mask ratio ρ, one contiguous interval of approximately $h = \rho T$ time samples is set to zero in every channel:

$$\widetilde{x}_{j,t} = x_{j,t} M_t, \qquad M_t = \begin{cases} 0, & q \le t < q + h, \\ 1, & \text{otherwise}, \end{cases} \qquad q \sim \mathcal{U}\{0, \ldots, T - h\}. \tag{3}$$

The admissible starts ensure that the masked interval stays inside the window. Mask-based EEG augmentation has a precedent in Ding et al. (2024); our mask is a shared contiguous time interval. In the RW-trained experiments of Table 1, training produces approximately 300,000 windows and uses $\rho = 0.3$ for the neural CLS and LaS models. FBCCA directly classifies the test signal and has no training stage (Chen et al., 2015).

RW-trained closed-set decoding. Table D.1 reproduces the nine neural comparisons from Table 1. All models use RW augmentation and temporal masking during training. Evaluation uses one Physical IW per trial at $\delta = 0$. The LaS results use EmbeddingGemma-300M. Superscripts reproduce the Holm-corrected paired-test marks in Table 1.

Random test windows and IW. Table D.2 reports RW-trained LaS (MTSNet) with EmbeddingGemma-300M. The random-window evaluation samples 10,000 test windows per fold; IW evaluation fixes $\delta = 0$ on Physical trials. The two columns therefore sample different within-trial positions.

Table D.1: Forty-class accuracy (%) after RW training and IW evaluation. Δ is LaS minus CLS in percentage points.

EEG encoder	L	CLS	LaS	Δ
EEGNet	0.6 s	26.76	26.98	+0.21
	0.8 s	41.07	40.42	-0.65
	1.0 s	51.62	49.86	-1.76
MTSNet	0.6 s	34.32	36.56	$+2.24^{**}$
	0.8 s	53.27	55.17	+1.89
	1.0 s	65.39	66.79	+1.39
SSVEPFormer	0.6 s	33.46	35.94	$+2.48^{*}$
	0.8 s	52.30	54.51	$+2.21^{*}$
	1.0 s	65.23	67.50	$+2.27^{*}$

Table D.2: Closed-set accuracy (%) for RW-trained LaS (MTSNet) under random-window and IW evaluation.

Window length L	10,000 random test windows per fold	Physical IW ($\delta = 0$)
0.6 s	48.1171	36.5595
0.8 s	63.4886	55.1667
1.0 s	74.7371	66.7857

IW-trained and IR-augmented closed-set results. Tables D.3 and D.4 reproduce the five-duration results in Appendix B.1. LaS (MTSNet) uses EmbeddingGemma-300M and approximately 300,000 IR training windows. Neural CLS baselines train on IW without IR and retain temporal masking at $\rho = 0.3$; FBCCA uses neither training nor augmentation. Every method is evaluated on IW. Estimated ITR includes a 0.5-s gaze-shift allowance per selection.

Table D.3: Forty-class IW accuracy (%, subject mean) across five window lengths. Values and significance marks reproduce Appendix B.1.

Method	0.6 s	0.7 s	0.8 s	0.9 s	1.0 s
FBCCA	28.25^{***}	38.20^{***}	48.54^{***}	56.90^{***}	63.75^{***}
EEGNet (CLS)	45.24^{***}	52.51^{***}	58.38^{***}	63.52^{***}	65.45^{***}
SSVEPFormer (CLS)	55.25	63.46	70.33	75.65	77.74
MTSNet (CLS)	52.06^{*}	62.61	68.61^{**}	73.98^{*}	78.01
LaS (MTSNet)	55.18	64.46	71.93	76.42	79.50

Table D.4: Estimated ITR (bits/min, subject mean) for the same IW conditions. Values and significance marks reproduce Appendix B.1.

Method	0.6 s	0.7 s	0.8 s	0.9 s	1.0 s
FBCCA	41.0^{***}	60.4^{***}	80.0^{***}	94.5^{***}	104.7^{***}
EEGNet (CLS)	84.6^{***}	96.8^{***}	105.1^{***}	111.2^{***}	108.6^{***}
SSVEPFormer (CLS)	114.4	129.6	139.5	145.5	141.2^{*}
MTSNet (CLS)	105.7^{*}	126.1	134.7^{**}	140.6^{*}	142.3
LaS (MTSNet)	114.8	132.4	144.8	147.6	146.4

The superscripts in these two tables reproduce the paired-test marks reported in Appendix B.1, separately for accuracy and ITR.

IR volume and mask ratio. Tables D.5 and D.6 use MTSNet with BERT-base-uncased at $L = 1.0$ s under IW evaluation. The IR-volume scan holds $\rho = 0.3$; the mask scan holds approximately 300,000 IR training windows. The 300-trial condition at $\rho = 0.3$ is common to both tables.

In the mask scan, the highest recorded accuracy occurs at $\rho = 0.3$ (Table D.6). For the same RW-trained MTSNet LaS model, accuracy is lower on the initial Physical window than on ran-

Table D.5: IR training volume and IW closed-set accuracy (%). The no-IR condition uses approximately 6,000 training windows.

IR trials / approximate training windows	IW accuracy
No IR / 6,000	34.7738
50 / 50,000	71.1667
100 / 100,000	73.3810
200 / 200,000	74.6905
300 / 300,000	76.3571

Table D.6: Temporal mask ratio and IW closed-set accuracy (%) with approximately 300,000 IR windows. The epoch budgets differ among mask conditions.

Mask ratio ρ	IW accuracy
0.0	72.8452
0.1	74.8571
0.2	76.0238
0.3	76.3571
0.4	75.3095
0.5	73.2500

domly sampled test windows at all three durations (Table D.2). In the IR-volume experiment with BERT-base-uncased, initial-window accuracy rises from 34.7738% without IR to 76.3571% with approximately 300,000 recombined windows (Table D.5). At 1.0 s, IR-trained MTSNet LaS with EmbeddingGemma reaches 79.50%, in the range of the supervised MTSNet result of 78.01% obtained without IR (Table D.3). This dependence on IR augmentation motivates examining how the response changes over the first 0.5 s after the latency-corrected onset.

D.3 Early-Window Behaviour and Conditional Unseen-Class Decoding

Window-onset offset. At $L = 1.0$ s, we move the window start by $\delta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ s relative to t_0. Each evaluation uses one window $[0.14 + \delta, 1.14 + \delta)$ per Physical trial, without combining predictions across offsets. Figure D.1 shows 40-class accuracy for FBCCA, RW-trained CLS (MTSNet), and RW-trained LaS (MTSNet with EmbeddingGemma-300M). All three curves use the same 8,400 Physical trials at each offset.

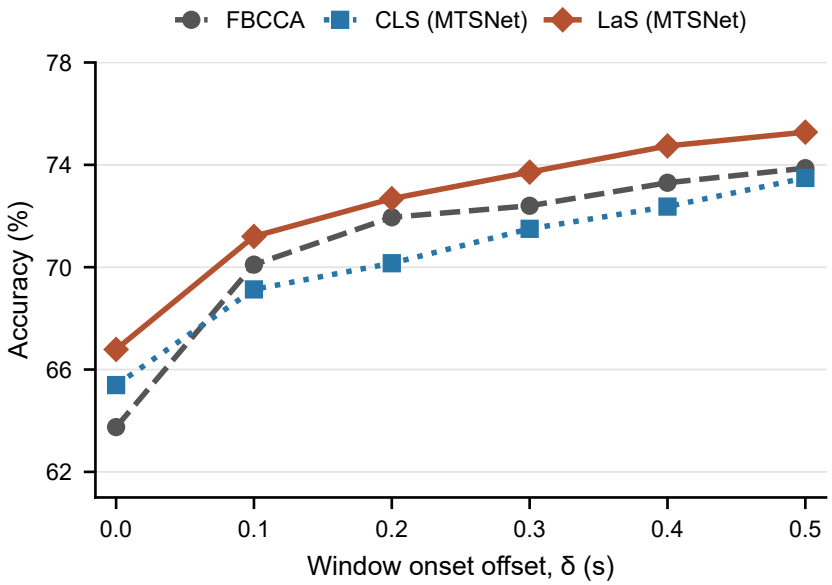


Figure D.1: Forty-class accuracy of a single 1.0-s Physical window at six offsets from the valid analysis onset $t_0 = 0.14$ s. The neural models are trained with RW; FBCCA directly classifies each window.

All three curves rise through the last measured offset, $\delta = 0.5$ s.

Conditional unseen-class decoding after IR training. Figure 4c of the main paper reports five-class unseen-only results after RW training. Table D.7 reports IR-trained MTSNet results under the same five-candidate evaluation format, organized by text encoder, adaptation state, and window length. IR produces approximately 300,000 training windows with $\rho = 0.3$; each result covers eight class-held-out splits with five unseen candidates per split.

Table D.7: Five-class unseen-only ZSL accuracy (%) after IR training.

Text encoder	State	0.6 s	0.8 s	1.0 s
BERT-base-uncased	Frozen	27.0200	42.3800	40.0100
BERT-base-uncased	LoRA	34.1500	59.3900	55.8700
EmbeddingGemma-300M	Frozen	34.3333	65.7857	65.6071
EmbeddingGemma-300M	LoRA	32.0119	58.0952	60.2619

The shared early-window penalty is consistent with a response that remains nonstationary after latency correction. We hypothesize that supervised CLS can combine stimulus frequency with phase and class-specific onset transients, whereas LaS finds these additional early-window cues harder to learn. The gain from IR augmentation for LaS (Table D.5) is consistent with this difference in access to early-window cues. Early nonstationarity may also impair the frequency representation needed for unseen-class transfer. At 0.6, 0.8, and 1.0 s, IR-trained MTSNet with EmbeddingGemma-300M obtains 34.3333%, 65.7857%, and 65.6071% five-class ZSL accuracy (Table D.7), compared with 80.6786%, 87.9762%, and 90.8929% for its RW-trained counterpart in Figure 4c.