Decide Before You Record: Heterogeneity-Guided Pre-Acquisition Stimulus Selection for Cross-Day Neuroprostheses

Di Wu* **Binglun Huang*** **Chengxi Xie** **Zhenxi Song** **Zhiguo Zhang†**
School of Intelligence Science and Engineering
Harbin Institute of Technology, Shenzhen
zhiguozhang@hit.edu.cn

Abstract

Decoding language intentions from intracranial electroencephalography signals enables direct communication for patients with severe speech impairments. However, practical deployment remains severely limited by cross-day performance degradation caused by physiological and instrumental heterogeneity, requiring patients to record extensive calibration data daily to maintain decoding performance. Here we introduce **He**terogeneity-Guided **D**ata **A**cquisition and **A**daptation (**HeDA^2**), a novel calibration paradigm that **determines which stimuli to record** for optimal calibration **before any target-day neural data are acquired**, fundamentally departing from existing methods that passively adapt to whatever data happen to be available. HeDA^2 operates through three stages: it first quantifies cross-day neural heterogeneity with a proposed **S**entence-level **H**eterogeneity via **A**lignment **P**ath **E**ntropy that captures how different phonetic phenomena exhibit varying degrees of heterogeneity; second, an estimator learns to predict this heterogeneity from historical patterns, enabling strategic selection of calibration sentences before target-day recording begins; third, heterogeneity-aware adaptation updates the decoder using minimal acquired data. Experiments on intracranial recordings from two participants demonstrate consistent gains over random acquisition under matched sentence budgets. On T12, using approximately one fifth of the calibration sentences, HeDA^2 recovers most of the improvement obtained by full-data fine-tuning, supporting more data-efficient calibration of speech neuroprostheses.

1 Introduction

Brain-computer interfaces that decode language from neural activity offer a direct route toward restoring communication for patients with severe speech impairments, such as those caused by amyotrophic lateral sclerosis or locked-in syndrome [1, 2, 3]. Recent advances in high-resolution invasive recordings have enabled continuous decoding of attempted speech into text [2, 4, 5] and even directly into synthesized voice [6, 7], approaching clinically meaningful throughput.

Despite this progress, a fundamental practical bottleneck still prevents the long-term clinical deployment of such systems. A decoder trained on historical neural recordings suffers from pronounced performance degradation when used on subsequent days, owing to electrode micro-displacement, fluctuations in neural state, and broader physiological variability across recording sessions (Figure 1A) [2]. To counteract cross-day heterogeneity, current speech neuroprostheses require tens of minutes of attempted-speech recordings each day to recalibrate the decoder before communication

*These authors contributed equally.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

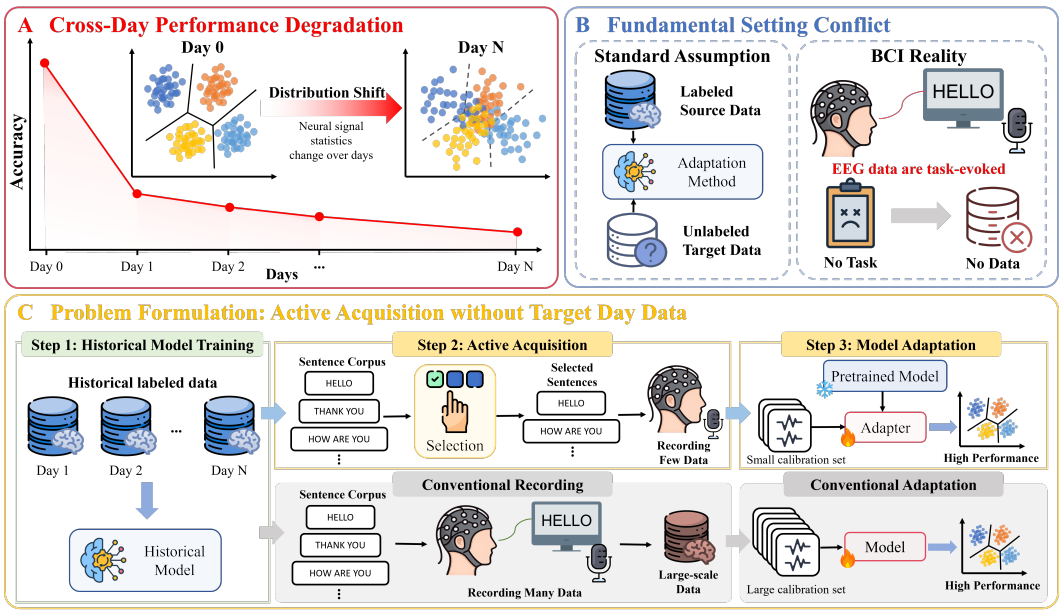


Figure 1: Overview of the problem formulation. (A) Decoder accuracy drops daily due to neural signal drift. (B) Target-day adaptation and pool-based selection require recordings that are unavailable before stimulus acquisition. (C) Active Acquisition selects a few informative sentences before recording for calibration, replacing conventional large-scale daily recording.

resumes [2, 4, 1, 5]. This daily recording burden strains patients' limited attention and endurance, restricting the practical use of neural language prostheses.

Existing attempts to alleviate this burden include transfer learning, unsupervised domain adaptation, and active learning [8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. Target-day fine-tuning and domain adaptation require recordings for adaptation, while pool-based active learning requires candidate recordings to rank (Figure 1B). These methods can operate after acquisition, but do not determine which stimuli should elicit the recordings in the first place. The truly missing capability is therefore not a better way to exploit already recorded trials, but a principled answer to a logically prior question: **when no target-day data yet exist, which stimuli should be prioritized for recording so that minimal acquisition effort translates into maximal adaptation benefit?**

We propose Active Acquisition (**AA**), a calibration paradigm that selects stimuli before neural recording (Figure 1C). We further realize this paradigm by proposing **HeDA^2**, which operates in three integrated stages. First, we quantify cross-day neural heterogeneity through **S**entence-level **H**eterogeneity via **A**lignment **P**ath **E**ntropy (SHAPE), which captures how confidently the historical model explains a neural signal under a given phoneme sequence, and serves as a proxy for cross-day representational mismatch. SHAPE itself requires a neural recording. Second, we train a textual estimator on historical neural-derived SHAPE targets to predict relative SHAPE from linguistic features, scoring candidate stimuli without target-day recordings. Third, once the chosen sentences have been recorded, a heterogeneity-aware adaptation step decouples the spatial heterogeneity induced by electrode displacement and semantic heterogeneity associated with neural state changes, leading to stable and data-efficient decoder recalibration.

We evaluate HeDA^2 on intracranial recordings from two participants across diverse cross-day recalibration settings. On T12, HeDA^2 recovers most of the improvement obtained by the full-calibration baseline using approximately one fifth of the calibration sentences. Evaluation on T15 further shows gains over random acquisition with hyperparameters transferred from T12, supporting data-efficient calibration across these two participants. Our contributions are summarized as follows:

(1) We introduce Active Acquisition, a new calibration paradigm for neural language decoders that determines which stimuli to record before any target-day neural data yet exist, thereby shifting the calibration decision from sample selection to stimulus selection, resolving the mismatch between conventional calibration approaches with real-world long-term Neuroprostheses.

(2) We instantiate this paradigm as HeDA^2, a unified framework that couples SHAPE with a textual estimator for stratified pre-acquisition stimulus selection, and a heterogeneity-aware adaptation step that decouples spatial and semantic corrections to deliver stable cross-day decoding gains.
(3) Experiments across multiple recalibration settings show that HeDA^2 improves over random acquisition under matched sentence budgets and, on T12, recovers most of FullFT's improvement using approximately one fifth of the calibration sentences.

2 Related Work

2.1 Neural Language Decoding

Decoding language intentions from neural activity offers a promising route toward restoring communication for patients with severe speech impairments. Existing studies have demonstrated the feasibility of decoding language and speech intentions from both non-invasive neural recordings [18, 19, 20, 21, 22, 23] and invasive neural signals [24, 25, 1, 26]. Early efforts predominantly focused on the classification of isolated language components, including syllables, phonemes, and words, based on non-invasive modalities such as functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS), and electroencephalography (EEG) [20, 21, 22]. In comparison, decoding based on invasive recordings has achieved markedly superior performance, owing to the high spatial resolution and favorable signal-to-noise ratio of cortical-level signals. Three complementary modalities have been widely adopted, namely electrocorticography (ECoG) and stereoelectroencephalography (sEEG) that measure local field potentials from surface or depth electrodes [27, 4, 28, 29], and intracortical microelectrode arrays such as Utah arrays that record single-neuron spiking activity from penetrating electrodes [2, 5, 6]. Building on the pioneering sentence-level decoding framework of Anumanchipalli *et al.* [27], subsequent studies have progressively scaled neural decoding toward continuous, real-time communication, producing high-performance text neuroprostheses [2, 4, 5] as well as voice-synthesis systems capable of closed-loop audio feedback and paralinguistic modulation [6, 7]. Complementary research has further broadened the deployment scope of such systems, including the decoding of inner speech from motor cortex [30, 31] and the extension of intracortical speech decoding to participants with longstanding anarthria and locked-in syndrome [3]. Despite these encouraging advances, decoders trained on historical recordings suffer from substantial cross-day performance degradation and therefore rely on extensive daily recalibration recordings to remain usable [2, 4, 1, 5], a burden that severely limits the long-term clinical deployment of neural language prostheses.

2.2 Heterogeneity in Neural Decoding

Heterogeneity across subjects, sessions, devices, and tasks is a well-recognized property of neural recordings in BCIs, and has motivated a substantial body of transfer learning research [32, 33]. A widely adopted strategy is unified modeling that aggregates heterogeneous recordings into a shared representation, exemplified by subject-independent decoders in classical EEG-based BCIs [11, 34, 35, 36, 37], channel-agnostic or topology-unified large-scale pre-training for brain signals [12, 38, 13, 39], and cross-subject neural decoders built from distributed intracranial recordings tailored to speech [40, 14]. Complementary to such unified modeling, unsupervised domain adaptation seeks to align source and target feature distributions without target-day labels [15, 16] or adversarial learning of invariant representations [17]. Active learning [8, 9, 10] reduces calibration effort by selecting the most informative samples under a fixed annotation budget, and has also been combined with transfer learning for further calibration reduction in BCIs [32]. At the clinical deployment level, current speech neuroprostheses continue to rely on extensive daily recalibration, with tens of minutes of newly recorded data typically collected at the start of each day to preserve decoding accuracy [1, 2, 4, 5]. Target-day fine-tuning and domain adaptation operate on recordings after collection, while pool-based active learning ranks existing candidate recordings. Active Acquisition addresses the preceding decision of which stimuli to prioritize when no target-day recordings yet exist, and is complementary to these adaptation methods.

3 Preliminaries

3.1 Neural Language Decoding

We consider the task of decoding phoneme sequences from multichannel intracranial signals. Let $\mathbf{X} \in \mathbb{R}^{C \times T}$ denote the neural signal, where C is the number of recording channels and T is the

number of time steps. The decoding target is a phoneme sequence $\mathbf{y} = (y_1, \ldots, y_L)$ where each phoneme y_l belongs to a finite vocabulary $\mathcal{V}$ and L denotes the sequence length. A neural decoder parameterized by θ maps $\mathbf{X}$ to a sequence of probability distributions over $\mathcal{V} \cup \{\varnothing\}$, where $\varnothing$ denotes the blank symbol. At each time step t, the decoder outputs $p_\theta(v \mid t)$ for $v \in \mathcal{V} \cup \{\varnothing\}$.

Since the alignment between neural time steps and phonemes is unknown, the Connectionist Temporal Classification (CTC) objective is often adopted. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ denote an alignment path where $\pi_t \in \mathcal{V} \cup \{\varnothing\}$. Define $\mathcal{B}(\cdot)$ as the CTC collapse function that removes consecutive repeated labels and blank symbols. The set of valid alignment paths is

$$\mathcal{A}(\mathbf{y}) = \{\boldsymbol{\pi} \mid \mathcal{B}(\boldsymbol{\pi}) = \mathbf{y}, \boldsymbol{\pi} \in (\mathcal{V} \cup \{\varnothing\})^T\}. \tag{1}$$

The probability of an alignment path is then $P(\boldsymbol{\pi} \mid \mathbf{X}) = \prod_{t=1}^{T} p_\theta(\pi_t \mid t)$. The conditional probability of the phoneme sequence is obtained by marginalizing over all valid paths:

$$P(\mathbf{y} \mid \mathbf{X}) = \sum_{\boldsymbol{\pi} \in \mathcal{A}(\mathbf{y})} P(\boldsymbol{\pi} \mid \mathbf{X}). \tag{2}$$

The CTC objective is then defined as

$$\mathcal{L}_{\mathrm{CTC}}(\mathbf{X}, \mathbf{y}) = -\log P(\mathbf{y} \mid \mathbf{X}). \tag{3}$$

3.2 Problem Formulation: Active Acquisition without Target Day Data

We address cross-day degradation in neural language decoding caused by electrode displacement and neural state heterogeneity. Whereas target-day fine-tuning and pool-based active learning operate on existing recordings, we study the preceding decision, called Active Acquisition: **selecting informative stimuli before any target domain data exist.** We describe the setting:

1 **Historical Model Training.** Assuming we have access only to historical neural recordings $\mathcal{D}_{\text{hist}} = \bigcup_{i=1}^{N} \mathcal{D}_i$ from days 1 through N where $\mathcal{D}_i = \{(\mathbf{X}_j^{(i)}, \mathbf{y}_j^{(i)})\}_{j=1}^{n_i}$ denote the labeled neural data recorded on day i and a pretrained decoder $f_{\theta_0} : \mathbb{R}^{C \times T} \to \mathcal{V}^*$ parameterized by θ_0.

2 **Active Acquisition.** No labeled or unlabeled neural data $\mathcal{D}_{N'}$ from the future days are available. Let $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$ denote a predefined corpus of M sentences, each associated with a known phoneme sequence. The goal is to select a subset $\mathcal{S}_k \subset \mathcal{S}$ with $|\mathcal{S}_k| = k \ll M$ such that recording neural responses to these sentences yields maximal benefit for model adaptation. Without access to any data from day N' satisfying $N' > N$, we must design a selection function $\pi : \mathcal{S} \to \{0, 1\}$ that identifies which sentences to record. After acquisition, we obtain a small calibration set $\mathcal{D}_{\text{cal}} = \{(\mathbf{X}_j^{(N')}, \mathbf{y}_j^{(N')})\}_{j=1}^{k}$ corresponding to the selected sentences.

3 **Model Adaptation.** Using $\mathcal{D}_{\text{cal}}$, we adapt the historical model f_{θ_0} to obtain an updated decoder f_{θ^*} that mitigates cross-day heterogeneity.

Selecting informative sentences reduces patients' recording burden during cross-day adaptation. A carefully chosen calibration subset enables efficient adaptation instead of hundreds of daily recordings.

4 Heterogeneity-Guided Data Acquisition and Adaptation

4.1 Overview of HeDA^2

We propose HeDA^2, a framework that dramatically reduces data collection burden during cross-day neural decoder calibration (Figure 2). By intelligently selecting which sentences to record, HeDA^2 achieves effective adaptation with minimal calibration data, a critical requirement for clinical deployment. HeDA^2 operates through three integrated stages. First, we quantify cross-day neural heterogeneity by introducing SHAPE, an alignment uncertainty metric used as a proxy for cross-day representational mismatch (Section 4.2). Second, we train a predictor on historical heterogeneity patterns, enabling strategic sentence selection without any target day recordings (Section 4.3). Third, we perform adaptation using the acquired data, with specialized updates targeting the specific heterogeneity sources (Section 4.4).

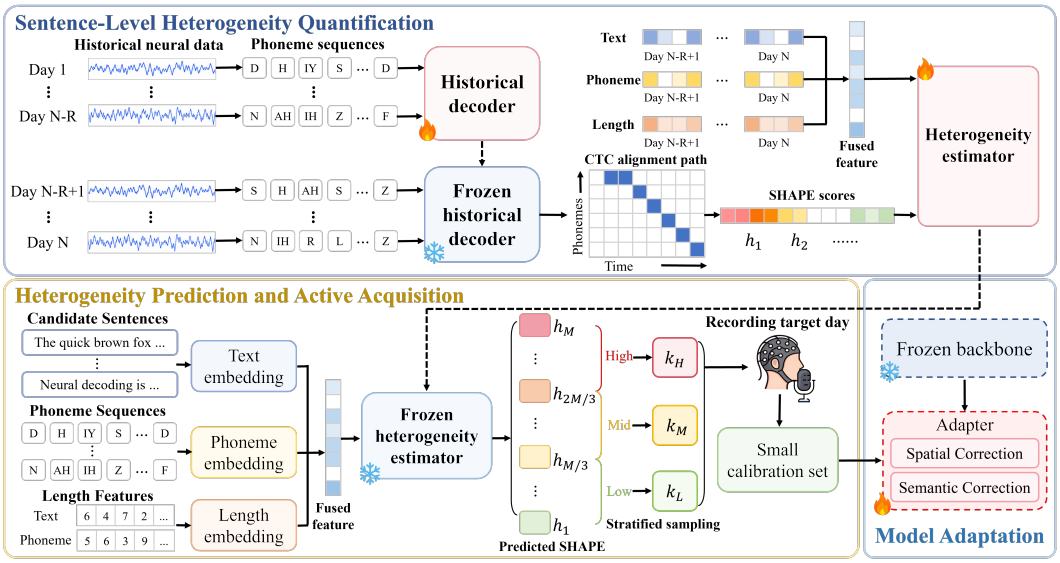


Figure 2: Overview of HeDA^2. Top: SHAPE quantifies cross-day heterogeneity from CTC alignment paths of a frozen historical decoder on past recordings. Bottom: an estimator predicts SHAPE for every candidate sentence, and stratified sampling selects sentences for the target day. Lightweight spatial and semantic correction adapters adapt the frozen backbone to the small calibration set.

4.2 Sentence-Level Heterogeneity Quantification

We quantify cross-day neural heterogeneity through a novel entropy-based metric that measures alignment uncertainty in pre-trained neural decoders. For a neural signal $\mathbf{X} \in \mathbb{R}^{C\times T}$ and phoneme sequence $\mathbf{y}$, we define SHAPE as the Shannon entropy over the posterior distribution of CTC alignment paths:

$$\text{SHAPE}(\mathbf{X}, \mathbf{y}; \theta) = -\sum_{\boldsymbol{\pi}\in\mathcal{A}(\mathbf{y})} P(\boldsymbol{\pi} \mid \mathbf{X}, \mathbf{y}) \log P(\boldsymbol{\pi} \mid \mathbf{X}, \mathbf{y}) \tag{4}$$

where the posterior probability of each alignment path is $P(\boldsymbol{\pi} \mid \mathbf{X}, \mathbf{y}) = \frac{P(\boldsymbol{\pi}|\mathbf{X};\theta)}{P(\mathbf{y}|\mathbf{X};\theta)}$, and $P(\boldsymbol{\pi} \mid \mathbf{X}; \theta) = \prod_{t=1}^{T} p_\theta(\pi_t \mid t)$ represents the alignment path probability under decoder parameters θ.

SHAPE captures how confidently the historical model explains the neural signal using the given phoneme sequence. Larger values indicate higher alignment ambiguity, which can reflect mismatch between the current neural signal and the decoder's learned representations, as well as trial difficulty and decoder uncertainty. SHAPE is computed from a neural recording and its phoneme sequence. Historical recordings provide the SHAPE targets for a textual predictor; only this predictor is available for scoring candidate stimuli before target-day acquisition.

4.3 Heterogeneity Prediction and Active Acquisition

We leverage historical heterogeneity patterns to predict future heterogeneity, enabling strategic data acquisition without target domain recordings.

Heterogeneity Prediction. We first partition historical data $\mathcal{D}_{\text{hist}}$ into two disjoint sets. Days 1 to $N - R$ are used to train the historical decoder:

$$\theta_0 = \arg\min_{\theta} \sum_{i=1}^{N-R} \sum_{(\mathbf{X},\mathbf{y})\in\mathcal{D}_i} \mathcal{L}_{\text{CTC}}(\mathbf{X}, \mathbf{y}; \theta). \tag{5}$$

Days $(N - R + 1)$ to N are reserved for characterizing heterogeneity evolution. Using the frozen decoder f_{θ_0}, we compute SHAPE values for all recordings from days $(N - R + 1)$ to N:

$$\mathcal{H} = \bigcup_{i=N-R+1}^{N} \{(\mathbf{z}_j, h_j) : (\mathbf{X}_j^{(i)}, \mathbf{y}_j^{(i)}) \in \mathcal{D}_i\}, \tag{6}$$

where $h_j = \text{SHAPE}(\mathbf{X}_j^{(i)}, \mathbf{y}_j^{(i)}; \theta_0)$ and the feature tuple is $\mathbf{z}_j = (s_j, \ell_s^{(j)}, \mathbf{y}_j, \ell_y^{(j)})$, consisting of sentence text s_j, text length $\ell_s^{(j)} = |s_j|$, phoneme sequence $\mathbf{y}_j$, and phoneme length $\ell_y^{(j)} = |\mathbf{y}_j|$. We then train an estimator $g_\phi : \mathcal{S} \times \mathbb{N} \times \mathcal{V}^* \times \mathbb{N} \to \mathbb{R}$ to predict SHAPE from textual features:

$$\hat{h} = g_\phi(\mathbf{z}) = \text{MLP}(\text{Enc}_{\text{LM}}(s, \ell_s, \mathbf{y}, \ell_y)). \tag{7}$$

Using ordered pairs $\mathcal{P} = \{(i, j) : h_i > h_j, (\mathbf{z}_i, h_i), (\mathbf{z}_j, h_j) \in \mathcal{H}\}$, we optimize:

$$\phi^* = \arg\min_\phi \sum_{(i,j)\in\mathcal{P}} \max(0, m - (g_\phi(\mathbf{z}_i) - g_\phi(\mathbf{z}_j))), \tag{8}$$

where $m > 0$ is the margin parameter ensuring correct relative ordering.

Stratified Acquisition. For deployment on day N', we apply g_{ϕ^*} to corpus $\mathcal{S}$ and obtain predicted heterogeneity scores $\{\hat{h}_i\}_{i=1}^M$. Let $\hat{h}_{(1)} \le \hat{h}_{(2)} \le \ldots \le \hat{h}_{(M)}$ denote the sorted scores. We partition sentences into three strata:

$$\mathcal{S}_L = \{s_i : \hat{h}_i < \hat{h}_{(\lfloor M/3 \rfloor)}\} \quad \text{(low heterogeneity)} \tag{9}$$

$$\mathcal{S}_M = \{s_i : \hat{h}_{(\lfloor M/3 \rfloor)} \le \hat{h}_i < \hat{h}_{(\lfloor 2M/3 \rfloor)}\} \quad \text{(medium heterogeneity)} \tag{10}$$

$$\mathcal{S}_H = \{s_i : \hat{h}_i \ge \hat{h}_{(\lfloor 2M/3 \rfloor)}\} \quad \text{(high heterogeneity)} \tag{11}$$

To acquire k sentences, we sample $k_H = \lfloor \alpha k \rfloor$, $k_M = \lfloor \beta k \rfloor$, and $k_L = k - k_H - k_M$ sentences from $\mathcal{S}_H$, $\mathcal{S}_M$, and $\mathcal{S}_L$ respectively, where $\alpha, \beta \in [0, 1]$ are predefined sampling probabilities with $\alpha + \beta \le 1$. Clarification of SHAPE prediction and acquisition parameters appears in Appendix B.

4.4 Heterogeneity-aware Model Adaptation

After acquiring the selected sentences $\mathcal{S}_k$ and recording corresponding neural responses on day N', we obtain the calibration set $\mathcal{D}_{\text{cal}} = \{(\mathbf{X}_j^{(N')}, \mathbf{y}_j^{(N')})\}_{j=1}^k$. We perform targeted adaptation that addresses two primary sources of cross-day heterogeneity: electrode displacement affecting local neural patterns, and neural state changes affecting high-level semantic representations. Modern neural decoders typically contain convolutional components for local feature extraction and linear modules for modeling long-range dependencies. We design distinct adaptation strategies for these two component types based on their roles in neural representation, and illustrate both on one Conformer block in Figure 3.

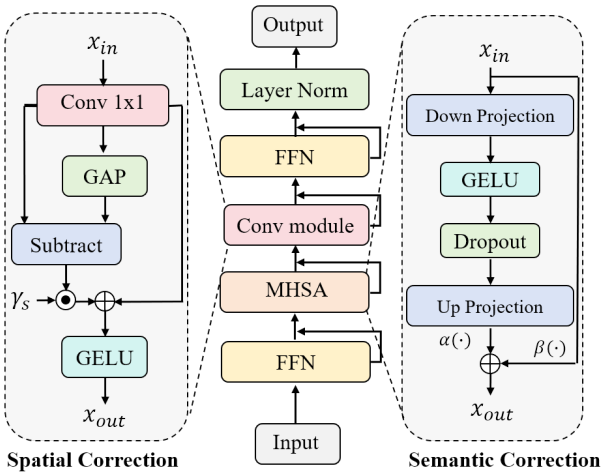


Figure 3: Illustration of heterogeneity-aware adaptation on one Conformer block. Only the parameters inside the highlighted boxes are updated during adaptation; the original decoder weights remain frozen.

Spatial Heterogeneity Correction. Convolutional layers capture local spatiotemporal patterns that are directly affected by electrode displacement and local neural pattern changes. For latent features $\mathbf{X} \in \mathbb{R}^{C' \times T'}$ in a convolutional layer, we construct a spatiotemporal heterogeneity template $\mathbf{Y} = \text{Conv}_{1\times1}(\mathbf{X})$ and compute the corrected feature as:

$$\mathbf{Z} = \text{GELU}(\mathbf{Y} + \boldsymbol{\gamma}_s \odot (\mathbf{Y} - \text{GAP}(\mathbf{Y}))) \tag{12}$$

where $\text{GAP}(\cdot)$ denotes global average pooling and $\boldsymbol{\gamma}_s \in \mathbb{R}^{C'}$ is a learnable channel-wise scaling vector. This design explicitly models and removes spatial shifts caused by electrode displacement.

Semantic Heterogeneity Correction. Attention mechanisms and feed-forward networks capture high-level semantic relationships that exhibit different heterogeneity patterns from local features. For input features $\mathbf{h} \in \mathbb{R}^D$ in these linear modules, we apply low-dimensional residual adaptation $\mathbf{h}' = \beta \cdot \mathbf{h} + \alpha \cdot \mathbf{h}_{\text{adapt}}$, where the adaptation term is:

$$\mathbf{h}_{\text{adapt}} = \mathbf{W}_{\text{up}} \cdot \text{Dropout}_{\rho}(\text{GELU}(\mathbf{W}_{\text{down}} \cdot \mathbf{h})) \tag{13}$$

Here, $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d \times D}$ and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{D \times d}$ are projection matrices with $d \ll D$, and $\alpha, \beta \in \mathbb{R}$ are learnable scalars. We optimize only the adaptation parameters $\Theta_{\text{adapt}} = \{\boldsymbol{\gamma}_s, \mathbf{W}_{\text{down}}, \mathbf{W}_{\text{up}}, \alpha, \beta\}$ while keeping original decoder parameters frozen:

$$\Theta^*_{\text{adapt}} = \arg\min_{\Theta_{\text{adapt}}} \sum_{(\mathbf{X},\mathbf{y}) \in \mathcal{D}_{\text{cal}}} \mathcal{L}_{\text{CTC}}(\mathbf{X}, \mathbf{y}; \theta_0, \Theta_{\text{adapt}}) + \lambda \|\Theta_{\text{adapt}}\|_2^2. \tag{14}$$

5 Experiments

5.1 Experimental Setup

Dataset. We evaluate HeDA^2 on the Brain-to-Text Benchmark '24 [2], a publicly available intracortical dataset collected from participant T12 who has amyotrophic lateral sclerosis. The dataset contains 12,100 attempted-speech sentences recorded over 25 sessions spanning approximately four months, covering both a 50-word vocabulary and a 125,000-word vocabulary. Neural activity is acquired from four 64-channel Utah microelectrode arrays implanted in the ventral motor cortex; we use threshold crossings and spiking-band power [6] binned in 20 ms windows from 128 of the 256 available electrodes, a configuration empirically found to maximize decoding accuracy. Neural features are z-scored on a per-day basis. Acquisition uses no target-day normalization statistics. After the selected k sentences are recorded, each feature's mean and variance are estimated only from those calibration trials and held fixed for evaluation; neither unselected target-day training trials nor test trials enter these estimates.

Cross-day transfer protocols. We evaluate HeDA^2 on T12 under three cross-day transfer protocols that differ in the length of the historical window and the distance between the historical window and the target deployment day. We consider: (i) **Protocol 1** ($N=2$, $R=1$, $N'=3$): a short-horizon setting with historical decoder trained with data from Day 1, Day 2 trains the heterogeneity estimator, and Day 3 is the target day; (ii) **Protocol 2** ($N=7$, $R=2$, $N' \in \{10, 11, 12, 13, 14\}$): a medium-horizon setting in which recalibration occurs three to seven days after the historical window; (iii) **Protocol 3** ($N=12$, $R=3$, $N' \in \{15, \ldots, 21\}$): a long-horizon setting in which recalibration occurs three to nine days after the historical window. For each protocol, the historical decoder θ_0 is trained on all sentences from Days 1 through $N-R$, while Days $(N-R+1)$ through N are used for training the textual SHAPE estimator. At target day N', the candidate corpus $\mathcal{S}$ consists of training sentences, from which $k=100$ sentences are selected for calibration unless otherwise specified. We additionally evaluate participant T15 from Brain-to-Text '25 [5], training a T15-specific decoder on Days 1–5 and estimator on Days 6–7, with Days 10–14 held out for evaluation. We use $k = 50$ because T15 sessions contain 100–200 sentences; all other hyperparameters are transferred from T12 without retuning.

Implementation details. We use a compact Conformer [41] (Appendix A), trained for 200 epochs with AdamW [42], batch size 64, learning rate 2×10^{-2} under a cosine schedule, and weight decay 10^{-5}. The SHAPE estimator g_ϕ combines pre-trained BERT-base sentence embeddings and phoneme-length features through a two-layer MLP, trained with margin ranking loss ($m = 1.0$) for 10 epochs with batch size 16 and learning rate 2×10^{-5}; stratified acquisition uses $\alpha = 0.8$ and $\beta = 0.15$. For heterogeneity-aware adaptation, the semantic adapter uses down-projection dimension $d = 128$, dropout $\rho = 0.4$, and ℓ_2 regularization $\lambda = 10^{-4}$, and is jointly updated with the spatial scaling vector $\boldsymbol{\gamma}_s$, initialized to 0.01, for 200 epochs at learning rate 3×10^{-4} while the original decoder remains frozen. We repeat all experiments with five random seeds and report mean and standard deviation of phoneme error rate (PER) on the target-day test set.

Baselines. We compare HeDA^2 with acquisition and adaptation baselines, distinguishing their access to target-day recordings. (1) NoFT: apply θ_0 directly to day N' without calibration. (2) RandomFT: randomly acquire k sentences from $\mathcal{S}$ and fully fine-tune θ_0, providing a within-subject cross-day fine-tuning baseline. (3) BADGE [8]: select k recordings via k-means++ on historical-decoder gradient embeddings. It is granted the candidate target-day recording pool unavailable

Table 1: Cross-day decoding PER (↓) on Protocol 1 and 2, reported per target day as mean ± std over five random seeds. Δ reports the absolute PER reduction of HeDA^2 (full) over RandomFT.

Method	Protocol 1	Protocol 2 ($N=7$, $R=2$, $N' \in \{10\text{–}14\}$)					
	Day 3	Day 10	Day 11	Day 12	Day 13	Day 14	Avg.
NoFT (lower bound)	$0.6550_{\pm.000}$	$0.5708_{\pm.000}$	$0.5897_{\pm.000}$	$0.6241_{\pm.000}$	$0.6366_{\pm.000}$	$0.6557_{\pm.000}$	$0.6154_{\pm.000}$
RandomFT ($k=100$)	$0.4949_{\pm.003}$	$0.4485_{\pm.003}$	$0.4332_{\pm.009}$	$0.4673_{\pm.006}$	$0.4433_{\pm.002}$	$0.4208_{\pm.008}$	$0.4426_{\pm.006}$
Badge [8] ($k=100$)	$0.4894_{\pm.003}$	$0.4393_{\pm.010}$	$0.4245_{\pm.010}$	$0.4717_{\pm.005}$	$0.4609_{\pm.010}$	$0.4206_{\pm.006}$	$0.4434_{\pm.008}$
Oracle SHAPE + FullFT	$0.4745_{\pm.002}$	$0.4210_{\pm.002}$	$0.4165_{\pm.002}$	$0.4550_{\pm.003}$	$0.4470_{\pm.002}$	$0.4016_{\pm.002}$	$0.4282_{\pm.002}$
HeDA^2-AA + FullFT	$0.4716_{\pm.008}$	$0.4283_{\pm.004}$	$0.4208_{\pm.003}$	$0.4412_{\pm.002}$	$0.4219_{\pm.006}$	$0.3997_{\pm.003}$	$0.4224_{\pm.002}$
HeDA^2 (full)	$0.4460_{\pm.004}$	$0.3892_{\pm.006}$	$0.4042_{\pm.002}$	$0.4128_{\pm.006}$	$0.4145_{\pm.001}$	$0.3588_{\pm.005}$	$0.3959_{\pm.003}$
FullFT (all data, $k=M$)	$0.3933_{\pm.004}$	$0.3902_{\pm.003}$	$0.3780_{\pm.002}$	$0.4200_{\pm.003}$	$0.3673_{\pm.003}$	$0.3592_{\pm.004}$	$0.3829_{\pm.003}$
Δ	$-0.0489_{\pm.005}$	$-0.0593_{\pm.007}$	$-0.0290_{\pm.009}$	$-0.0545_{\pm.011}$	$-0.0288_{\pm.002}$	$-0.0620_{\pm.005}$	$-0.0467_{\pm.003}$

Table 2: Cross-day decoding PER (↓) on Protocol 3, reported per target day as mean ± std over five random seeds. Δ reports the absolute PER reduction of HeDA^2 (full) over RandomFT.

Method	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20	Day 21	Avg.
NoFT (lower bound)	$0.4701_{\pm.000}$	$0.5469_{\pm.000}$	$0.5259_{\pm.000}$	$0.5410_{\pm.000}$	$0.5211_{\pm.000}$	$0.5439_{\pm.000}$	$0.5618_{\pm.000}$	$0.5301_{\pm.000}$
RandomFT ($k=100$)	$0.3343_{\pm.007}$	$0.3619_{\pm.008}$	$0.3778_{\pm.007}$	$0.3743_{\pm.005}$	$0.3711_{\pm.006}$	$0.3692_{\pm.009}$	$0.3459_{\pm.008}$	$0.3621_{\pm.007}$
Badge [8] ($k=100$)	$0.3456_{\pm.007}$	$0.3715_{\pm.008}$	$0.3812_{\pm.005}$	$0.3785_{\pm.005}$	$0.3745_{\pm.011}$	$0.3795_{\pm.009}$	$0.3399_{\pm.005}$	$0.3672_{\pm.007}$
Oracle SHAPE + FullFT	$0.3202_{\pm.007}$	$0.3493_{\pm.002}$	$0.3672_{\pm.002}$	$0.3653_{\pm.005}$	$0.3592_{\pm.004}$	$0.3583_{\pm.002}$	$0.3226_{\pm.002}$	$0.3489_{\pm.003}$
HeDA^2-AA + FullFT	$0.3227_{\pm.003}$	$0.3486_{\pm.007}$	$0.3695_{\pm.002}$	$0.3664_{\pm.001}$	$0.3572_{\pm.002}$	$0.3883_{\pm.004}$	$0.3433_{\pm.005}$	$0.3566_{\pm.003}$
HeDA^2 (full)	$0.3133_{\pm.007}$	$0.3121_{\pm.004}$	$0.3308_{\pm.001}$	$0.3500_{\pm.003}$	$0.3421_{\pm.003}$	$0.3443_{\pm.001}$	$0.3091_{\pm.003}$	$0.3288_{\pm.003}$
FullFT (all data, $k=M$)	$0.2986_{\pm.002}$	$0.3146_{\pm.001}$	$0.3365_{\pm.003}$	$0.3583_{\pm.002}$	$0.3226_{\pm.002}$	$0.3127_{\pm.002}$	$0.2844_{\pm.004}$	$0.3182_{\pm.002}$
Δ	$-0.0210_{\pm.007}$	$-0.0498_{\pm.009}$	$-0.0470_{\pm.007}$	$-0.0243_{\pm.006}$	$-0.0290_{\pm.006}$	$-0.0249_{\pm.007}$	$-0.0368_{\pm.007}$	$-0.0333_{\pm.007}$

before acquisition and is therefore an advantaged reference. (4) HEDA^2-AA + FULLFT: select k sentences using stratified predicted SHAPE and apply the same full fine-tuning as RandomFT, isolating acquisition from adaptation. (5) ORACLE SHAPE + FULLFT: compute SHAPE from actual target-day recordings and apply full fine-tuning; this reference also requires additional recording access unavailable before acquisition. (6) FULLFT: fine-tune θ_0 on all M target-day training trials as a full-data reference. Budget-matched methods use $k = 100$ on T12 and $k = 50$ on T15. Appendix B provides further details.

5.2 Main Results

We evaluate cross-day decoding under the three T12 transfer protocols. The per-target-day PER results in Tables 1 and 2 highlight five observations. First, NoFT has higher PER than every calibration-based baseline on every target day, with absolute gaps over FullFT ranging from 0.1715 to 0.2965, illustrating the benefit of cross-day recalibration. Second, replacing random selection with HeDA^2-AA while keeping full fine-tuning unchanged reduces average PER in all three protocols and improves 12 of the 13 target days, demonstrating the benefit of SHAPE-guided acquisition under a fixed adaptation procedure. Third, HeDA^2-AA achieves lower average PER than BADGE in all three protocols, supporting the utility of historical SHAPE prediction for stimulus selection. Fourth, full HeDA^2 improves over the budget-matched non-oracle alternatives. Across the 12 target days in Protocols 2 and 3, it recovers approximately 95% of FullFT's improvement relative to NoFT using 100 sentences, approximately one fifth of the full calibration set. The corresponding mean PERs are 0.3568 and 0.3452, respectively, quantifying the accuracy–acquisition trade-off (Appendix E). Fifth, the advantage over RandomFT is consistent across individual days, as reflected by the negative per-day ΔPER values.

To examine whether the gains extend beyond T12, we additionally evaluate T15 from Brain-to-Text '25 [5], a second participant with a different recording configuration. As shown in Table A2 in Appendix D, HeDA^2 reduces mean PER from 0.3861 (RandomFT) to 0.3458 with $k = 50$, improving on all five target days without hyperparameter retuning. FullFT achieves 0.3227 with all calibration trials. The appendix also provides the statistical comparison.

5.3 Ablation Studies

To dissect the gains of HeDA^2 and to understand which design choices are responsible for them, we conduct a series of ablation studies organized around the following research questions: (i) can the textual estimator approximate ground-truth SHAPE without any target-day recording? (ii) how much does each adaptation component contribute to cross-day performance? and (iii) how does HeDA^2 scale with the calibration budget k? Two further ablations, on the regression target of the textual

Table 3: Ablation on the heterogeneity estimator.

Selection	Day 10	Day 11	Day 12	Avg.
Oracle SHAPE (unavail.)	$0.4210_{\pm.002}$	$0.4165_{\pm.002}$	$0.4550_{\pm.003}$	$0.4308_{\pm.003}$
Predicted SHAPE (ours)	$0.4283_{\pm.004}$	$0.4208_{\pm.003}$	$0.4412_{\pm.002}$	$0.4301_{\pm.003}$
Overlap rate	78%	80%	76%	78%

Table 4: Component ablation on T12, Days 10–12 ($k = 100$; PER ↓).

Configuration	Day 10	Day 11	Day 12	Avg.
AA (full FT)	$0.4283_{\pm.004}$	$0.4208_{\pm.003}$	$0.4412_{\pm.002}$	$0.4301_{\pm.003}$
AA + Se.	$0.4127_{\pm.002}$	$0.4037_{\pm.001}$	$0.4430_{\pm.003}$	$0.4198_{\pm.002}$
AA + Sp.	$0.3995_{\pm.006}$	$0.4112_{\pm.004}$	$0.4346_{\pm.011}$	$0.4151_{\pm.005}$
AA + Sp. + Se.	$0.3892_{\pm.006}$	$0.4042_{\pm.002}$	$0.4128_{\pm.006}$	$0.4021_{\pm.003}$

estimator (Appendix F.2) and on the acquisition strategy (Appendix F.3), are deferred to the appendix due to space constraints. All ablation experiments are conducted on T12 Protocol 2 with $k = 100$ unless otherwise specified, following the same training protocol described in Section 5.1.

Does the heterogeneity estimator approximate ground-truth SHAPE? A central premise of HeDA^2 is that an estimator trained on historical SHAPE can substitute for ground-truth SHAPE when no target-day recordings exist. Table 3 validates this along two axes: under matched $k = 100$ and full fine-tuning, Predicted SHAPE matches Oracle SHAPE in downstream PER (0.4301 vs. 0.4308 on average), and the two selected subsets overlap by 78% sentences on average (up to 80% on Day 11), supporting useful agreement in stimulus selection without any target-day observation.

Contribution of each adaptation component. We next assess whether the two adaptation components separately and jointly contribute to cross-day decoding gains. Sharing HeDA^2 selection, we compare four configurations: (i) Active Acquisition with vanilla full fine-tuning (AA); (ii) AA with only semantic correction (AA + Se.); (iii) AA with only spatial correction (AA + Sp.); and (iv) AA with both corrections (full HeDA^2). Table 4 shows that semantic correction alone reduces average PER from 0.4301 to 0.4198, improving two of the three days. Spatial correction alone reduces average PER to 0.4151, and combining both corrections further reduces it to 0.4021. These results support the semantic adapter's contribution both alone and on top of spatial correction.
Additional interpretation of the component ablation appears in Appendix F.1.

How does performance scale with the acquisition budget k? We vary the acquisition budget and compare HeDA^2 with RandomFT under the same budget. As shown in Table 5, the PER of HeDA^2 decreases monotonically as k grows. HeDA^2 with $k = 150$ (approximately 30% of M) already outperforms RandomFT with $k = 300$. At $k = 200$, HeDA^2 achieves 0.3741 PER, compared with 0.3829 for FullFT using the entire approximately 500-sentence corpus. This comparison combines stimulus selection with a different adaptation procedure; the acquisition-only comparisons hold full fine-tuning fixed. Together, these results support reducing the number of calibration sentences through acquisition and adaptation.

Table 5: Ablation on the acquisition budget k on Protocol 2. Avg. PER (↓) over Days 10–14. At $k = M$, RandomFT is FullFT, HeDA^2 uses all trials with the proposed adaptation.

k	RandomFT	HeDA^2
50	$0.4792_{\pm.010}$	$0.4459_{\pm.020}$
100	$0.4426_{\pm.006}$	$0.3959_{\pm.003}$
150	$0.4232_{\pm.011}$	$0.3846_{\pm.016}$
200	$0.4087_{\pm.009}$	$0.3741_{\pm.020}$
300	$0.3889_{\pm.006}$	$0.3542_{\pm.004}$
M	$0.3829_{\pm.003}$	$0.3480_{\pm.004}$

5.4 Linguistic-based Acquisition Comparison

To assess acquisition quality beyond simple linguistic heuristics, we compare four text-only baselines on T12 Protocol 2: sentence length (Len), phoneme coverage (Cov), phoneme-bigram diversity (Bi), and text-embedding diversity (Emb). All methods select sentences before target-day recordings are available, and receive identical downstream full fine-tuning. This matched setting isolates the contribution of sentence selection. Appendix C provides implementation details. As shown in Table 6, SHAPE-guided acquisition achieves 0.4224 mean PER, compared with 0.4387 for the strongest text-only heuristic, Cov, an absolute reduction of 0.0163. It outperforms all four heuristics on each of the five target days. Among the heuristics, Cov slightly improves on RandomFT on average, whereas Len, Bi, and Emb do not. Thus, prioritizing sentence length or diversity alone does not consistently improve decoding in this setting. The gains from SHAPE-guided acquisition suggest that a ranking learned from historical neural-derived alignment uncertainty provides useful selection information beyond these linguistic criteria.

5.5 Interpretability Analysis

We examine whether predicted SHAPE is associated with decoder gradients relevant to target-day adaptation on T12 Protocols 2 and 3. We group candidates into score deciles and, within each decile,

Table 6: T12 Protocol 2 acquisition-only comparison: PER (↓), $k = 100$, and identical downstream full fine-tuning. Daily values are mean ± standard deviation over five seeds; Avg. is the mean over target days.

Method	Day 10	Day 11	Day 12	Day 13	Day 14	Avg.
RandomFT	$0.4485_{\pm.0031}$	$0.4332_{\pm.0087}$	$0.4673_{\pm.0062}$	$0.4433_{\pm.0024}$	$0.4208_{\pm.0079}$	$0.4426_{\pm.0057}$
Len	$0.4457_{\pm.0025}$	$0.4431_{\pm.0022}$	$0.4603_{\pm.0016}$	$0.4492_{\pm.0020}$	$0.4442_{\pm.0081}$	$0.4485_{\pm.0033}$
Cov	$0.4492_{\pm.0007}$	$0.4319_{\pm.0030}$	$0.4621_{\pm.0027}$	$0.4277_{\pm.0010}$	$0.4224_{\pm.0023}$	$0.4387_{\pm.0019}$
Bi	$0.4600_{\pm.0077}$	$0.4425_{\pm.0068}$	$0.4767_{\pm.0080}$	$0.4532_{\pm.0100}$	$0.4174_{\pm.0105}$	$0.4500_{\pm.0086}$
Emb	$0.4593_{\pm.0041}$	$0.4420_{\pm.0080}$	$0.4550_{\pm.0029}$	$0.4668_{\pm.0080}$	$0.4377_{\pm.0116}$	$0.4522_{\pm.0069}$
HeDA^2-AA + FullFT	$0.4283_{\pm.004}$	$0.4208_{\pm.003}$	$0.4412_{\pm.002}$	$0.4219_{\pm.006}$	$0.3997_{\pm.003}$	$0.4224_{\pm.002}$

average the cosine similarities of individual pre-adaptation output-head gradients to the mean gradient of 40 same-day test sentences. This test-derived reference is used only for retrospective interpretation; acquisition uses the historical textual estimator.

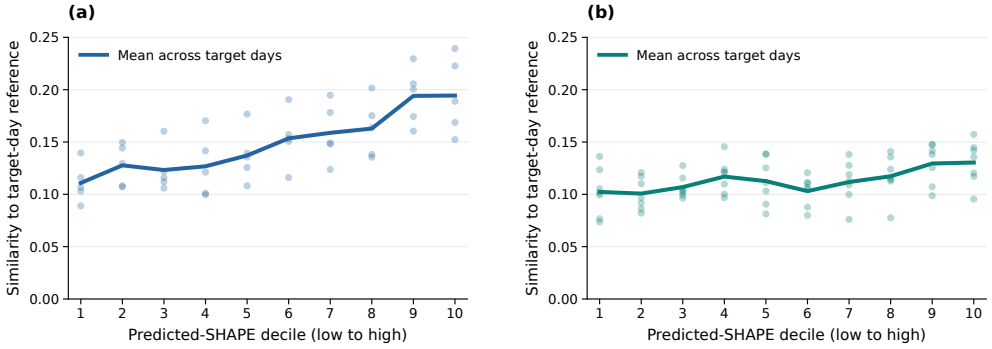


Figure 4: Gradient alignment across predicted-SHAPE deciles for participant T12 under Protocol 2 (a) and Protocol 3 (b). Faint dots show daily mean cosine similarities to the retrospective target-day reference; solid lines show averages with equal weight across target days.

As shown in Figure 4, higher-scoring groups generally exhibit greater target-day gradient alignment, with a more pronounced increase under Protocol 2 (Fig. 4-(a)). Additional diagnostic comparisons in Appendix G show that high-scoring sentences contain more phonemes and have longer recording windows, but fewer recording frames per phoneme. They also cover more adjacent phoneme pairs and exhibit more concentrated gradient directions. These observations associate high predicted SHAPE with richer phonetic content and update directions more closely aligned with target-day decoding. Together with the acquisition-strategy results in Table A4, they help explain prioritizing high-SHAPE sentences while retaining samples from the remaining strata.

6 Discussion and Conclusion

We introduced Active Acquisition, a calibration paradigm that decides which stimuli to record before any target-day recording occurs, and instantiated it with HeDA^2, which quantifies alignment uncertainty through Sentence-level Heterogeneity via Alignment Path Entropy, anticipates it with a textual estimator, and adapts the decoder through decoupled spatial and semantic corrections. Active Acquisition aligns calibration with the real BCI data acquisition workflow, and the resulting framework recovers most of FullFT's improvement on T12 while using approximately one fifth of the calibration sentences. Evaluation on T15 further shows gains over random acquisition without hyperparameter retuning, extending the cross-day calibration results to a second participant. A more detailed discussion of limitations, future work, and broader impact is provided in Appendix H.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No.62606002), Guangdong Provincial Key Laboratory of Artificial General Intelligence and Large Models (2026B1212010005), Lingang Laboratory (No. LGL-1987-03), and Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions (2023SHIBS0003).

References

[1] 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.

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

[3] Justin J Jude, Stephanie Haro, Hadar Levi-Aharoni, Hiroaki Hashimoto, Alexander J Acosta, Nicholas S Card, Maitreyee Wairagkar, David M Brandman, Sergey D Stavisky, Ziv M Williams, Sydney S Cash, John D Simeral, Leigh R Hochberg, and Daniel B Rubin. Decoding intended speech with an intracortical brain-computer interface in a person with longstanding anarthria and locked-in syndrome. *bioRxiv*, 2025.

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

[5] Nicholas S Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R Willett, Erin M Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R Deo, Aparna Srinivasan, Eun Young Choi, Matthew F Glasser, Leigh R Hochberg, Jaimie M Henderson, Kiarash Shahlaie, Sergey D Stavisky, and David M Brandman. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, 2024.

[6] Maitreyee Wairagkar, Nicholas S Card, Tyler Singer-Clark, Xianda Hou, Carrina Iacobacci, Lee M Miller, Leigh R Hochberg, David M Brandman, and Sergey D Stavisky. An instantaneous voice-synthesis neuroprosthesis. *Nature*, 644:145–154, 2025.

[7] Jiawei Li, Chunxu Guo, Chao Zhang, Edward F Chang, and Yuanning Li. High-fidelity neural speech reconstruction through an efficient acoustic-linguistic dual-pathway framework. *bioRxiv*, 2025.

[8] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[9] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8505–8514, 2021.

[10] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8708–8716, 2022.

[11] Siamac Fazli, Cristian Grozea, Marton Danoczy, Benjamin Blankertz, Florin Popescu, and Klaus-Robert Müller. Subject independent eeg-based bci decoding. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[12] Weibang Jiang, Liming Zhao, and Bao liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024.

[13] Chaoqi Yang, M Brandon Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[14] Aditya Singh, Tessy Thomas, Jinlong Li, Greg Hickok, Xaq Pitkow, and Nitin Tandon. Transfer learning via distributed brain recordings enables reliable speech decoding. *Nature Communications*, 16:8749, 2025.

[15] Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107–1116, 2017.

[16] He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2019.

[17] Ozan Ozdenizci, Ye Wang, Toshiaki Koike-Akino, and Deniz Erdoğmuş. Learning invariant representations from EEG via adversarial inference. *IEEE Access*, 8:27074–27085, 2020.

[18] Alborz Rezazadeh Sereshkeh, Rozhin Yousefi, Andrew T Wong, and Tom Chau. Online classification of imagined speech using functional near-infrared spectroscopy signals. *Journal of neural engineering*, 16(1):016005, 2018.

[19] Xiaopeng Si, Sicheng Li, Shaoxin Xiang, Jiayue Yu, and Dong Ming. Imagined speech increases the hemodynamic response and functional connectivity of the dorsal motor cortex. *Journal of Neural Engineering*, 18(5):056048, 2021.

[20] Siyi Deng, Ramesh Srinivasan, Tom Lappas, and Michael D'Zmura. Eeg classification of imagined syllable rhythm using hilbert spectrum methods. *Journal of neural engineering*, 7(4):046006, 2010.

[21] Charles S DaSalla, Hiroyuki Kambara, Makoto Sato, and Yasuharu Koike. Single-trial classification of vowel speech imagery using common spatial patterns. *Neural networks*, 22(9):1334–1339, 2009.

[22] Xuemin Chi, John B Hagedorn, Daniel Schoonover, and Michael D'Zmura. Eeg-based discrimination of imagined speech phonemes. *International Journal of Bioelectromagnetism*, 13(4):201–206, 2011.

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

[24] David A Moses, Matthew K Leonard, Joseph G Makin, and Edward F Chang. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature communications*, 10(1):3096, 2019.

[25] Joseph G Makin, David A Moses, and Edward F Chang. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582, 2020.

[26] Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski, and Tanja Schultz. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of neural engineering*, 16(3):036019, 2019.

[27] Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

[28] Chen Feng, Lu Cao, Di Wu, En Zhang, Ting Wang, Xiaowei Jiang, Heng Ding, Chenhao Zhou, Jinbo Chen, Hui Wu, et al. A high-performance brain-sentence communication designed for logosyllabic language. *bioRxiv*, pages 2023–11, 2023.

[29] Youkun Qian, Changjiang Liu, Peixi Yu, Xingchen Ran, Shurui Li, Qinrong Yang, Yang Liu, Lei Xia, Yijie Wang, Jianxuan Qi, et al. Real-time decoding of full-spectrum chinese using brain-computer interface. *Science Advances*, 11(45):eadz9968, 2025.

[30] Erin M Kunz, Benyamin Abramovich Krasa, Foram Kamdar, Donald T Avansino, Nick Hahn, Seonghyun Yoon, Akansha Singh, Samuel R Nason-Tomaszewski, Nicholas S Card, Justin J Jude, Payton G Bechefsky, Carrina Iacobacci, Leigh R Hochberg, Daniel B Rubin, Ziv M Williams, David M Brandman, Sergey D Stavisky, Nicholas AuYong, Chethan Pandarinath, Shaul Druckmann, Jaimie M Henderson, and Francis R Willett. Inner speech in motor cortex and implications for speech neuroprostheses. *Cell*, 188:4658–4673, 2025.

[31] Zehao Zhao, Zhenjie Wang, Yan Liu, Youkun Qian, Yuan Yin, Xiaowei Gao, Binke Yuan, Shelley Xiuli Tong, Xing Tian, Gao Chen, Yuanning Li, Junfeng Lu, and Jinsong Wu. Neural hierarchy for coding articulatory dynamics in speech imagery and production. *bioRxiv*, 2025.

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

[33] Di Wu, Jie Yang, and Mohamad Sawan. Transfer learning on electromyography (emg) tasks: approaches and beyond. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[34] Phairot Autthasan, Rattanaphon Chaisaen, Thapanun Sudhawiyangkul, Phurin Rangpong, Suktipol Kiatthaveephong, Nat Dilokthanakul, Gun Bhakdisongkhram, Huy Phan, Cuntai Guan, and Theerawit Wilaiprasitporn. Min2net: End-to-end multi-task learning for subject-independent motor imagery eeg classification. *IEEE Transactions on Biomedical Engineering*, 69(6):2105–2118, 2022.

[35] Shuaiqi Liu, Xu Wang, Ling Zhao, Jie Zhao, Qi Xin, and Shui-Hua Wang. Subject-independent emotion recognition of eeg signals based on dynamic empirical convolutional neural network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(5):1710–1721, 2021.

[36] Han Wei Ng and Cuntai Guan. Subject-independent meta-learning framework towards optimal training of eeg-based classifiers. *Neural Networks*, page 106108, 2024.

[37] Theekshana Dissanayake, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Geometric deep learning for subject independent epileptic seizure prediction using scalp eeg signals. *IEEE Journal of Biomedical and Health Informatics*, 26(2):527–538, 2021.

[38] Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[39] Antonis Antoniades, Yiyi Yu, Joseph Canzano, William Wang, and Spencer LaVere Smith. Neuroformer: Multimodal and multitask generative pretraining for brain data. *arXiv preprint arXiv:2311.00136*, 2023.

[40] Steven M Peterson, Zoe Steine-Hanson, Nathan Davis, Rajesh P N Rao, and Bingni W Brunton. Generalized neural decoders for transfer learning across participants and recording modalities. *Journal of Neural Engineering*, 18(2):026014, 2021.

[41] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

A Conformer Decoder Architecture

We use a compact Conformer [41] as the default phoneme decoder on participant T12, with its architecture detailed in Table A1. The decoder consists of three components: (i) an *embedding layer* that projects the 256-dimensional neural feature stream into the encoder's hidden space via a 1-D convolution with 256 output channels, kernel size 32, and stride 4, yielding a temporally downsampled token sequence; (ii) a *Conformer encoder* of three blocks with hidden dimension 256, each block combining a macaron feed-forward module with expansion factor 4, a local multi-head self-attention module with 8 heads and window size 64, and a convolution module with expansion factor 2 and kernel size 31; and (iii) an *output* stage that applies an input projection (Linear $256 \times 32 \to 256$) over the concatenated encoder outputs before a linear prediction head maps to the 41-class phoneme vocabulary (including the CTC blank). Convolution modules use BatchNorm1d and a dropout rate of 0.4 is applied throughout the encoder. The local self-attention window keeps the per-step complexity linear in sequence length, which matches the short-latency requirement of online neural decoding. All baselines and variants of HeDA^2 share this identical backbone; only the adaptation modules and the sentence selection strategy differ across conditions.

Table A1: Conformer decoder architecture used on participant T12. All variants of HeDA^2 and all baselines share this identical backbone.

Stage	Module	Configuration
Input	Neural features	256-dim threshold-crossing + spike-band power
Embedding	Conv1d	in=256, out=256, kernel=32, stride=4
	Activation	ReLU
Encoder ($\times 3$)	FFN (macaron, $\frac{1}{2}$)	hidden=256, expansion=4, dropout=0.4
	Local MHSA	heads=8, window=64, dropout=0.4
	Conv module	expansion=2, kernel=31, BatchNorm1d
	FFN (macaron, $\frac{1}{2}$)	hidden=256, expansion=4, dropout=0.4
Output	Input projection	Linear $256 \times 32 \to 256$ over concatenated encoder outputs
	Classification head	Linear $256 \to 41$ (40 phonemes + CTC blank)
Objective	Loss	CTC (Section$\sim$ 3)

B Additional Experimental Details

Data access for each baseline. We distinguish budget-matched acquisition methods from references with additional data access. NoFT evaluates the historical decoder without adaptation. RandomFT randomly acquires k sentences and fully fine-tunes the historical decoder, providing a within-subject cross-day transfer baseline. HeDA^2-AA + FullFT uses SHAPE-guided acquisition followed by identical full fine-tuning, isolating acquisition from adaptation. BADGE [8] selects recordings using historical-decoder gradient embeddings; it is granted a target-day recording pool unavailable at the pre-acquisition decision point and is therefore an advantaged, non-deployable reference. Oracle SHAPE + FullFT likewise uses target-day recordings to compute SHAPE and is unavailable before acquisition. FullFT uses the entire target-day training set ($k = M$) as a full-data reference. Finally, four deployable text-only acquisition rules share RandomFT's budget and full fine-tuning: longest sentences by character count (Len), greedy inverse-frequency phoneme coverage with harmonically discounted repetitions (Cov), farthest-first selection in TF–IDF-weighted, L2-normalized phoneme-bigram space (Bi), and k-means++ selection with squared Euclidean distance between RoBERTa-base CLS embeddings (Emb).

Normalization and sentence overlap. Acquisition uses only candidate text and the participant-specific estimator trained on historical neural-derived SHAPE targets. After acquisition, each target-day feature's mean and variance are estimated from exactly the selected k calibration trials and held fixed for evaluation. Neither unselected target-day training recordings nor test recordings enter these statistics. Calibration sentences come from the target-day training split and are disjoint from test sentences. T12 sentences are almost entirely distinct across sessions, so candidates are generally historically unseen.

Table A2: T15 held-out-day evaluation with the Protocol 2 split. PER (↓) is reported as mean ± standard deviation. Budget-matched methods share $k = 50$; FullFT uses all target-day training trials.

Method	Day 10	Day 11	Day 12	Day 13	Day 14	Avg.
NoFT	$0.4627_{\pm.000}$	$0.6169_{\pm.000}$	$0.4871_{\pm.000}$	$0.6378_{\pm.000}$	$0.5927_{\pm.0000}$	$0.5594_{\pm.0000}$
RandomFT	$0.3109_{\pm.008}$	$0.4083_{\pm.008}$	$0.3274_{\pm.014}$	$0.4424_{\pm.006}$	$0.4413_{\pm.0066}$	$0.3861_{\pm.0085}$
BADGE (recording pool)	$0.3085_{\pm.006}$	$0.4152_{\pm.007}$	$0.3316_{\pm.010}$	$0.4361_{\pm.007}$	$0.4385_{\pm.0067}$	$0.3860_{\pm.0073}$
HeDA^2	$0.2813_{\pm.001}$	$0.3654_{\pm.004}$	$0.2602_{\pm.003}$	$0.4037_{\pm.002}$	$0.4185_{\pm.0015}$	$0.3458_{\pm.0023}$
HeDA^2 + Oracle SHAPE	$0.2811_{\pm.002}$	$0.3522_{\pm.003}$	$0.2585_{\pm.002}$	$0.4008_{\pm.003}$	$0.4222_{\pm.0009}$	$0.3430_{\pm.0022}$
FullFT (full-data reference)	$0.2607_{\pm.001}$	$0.3352_{\pm.001}$	$0.2526_{\pm.002}$	$0.3828_{\pm.001}$	$0.3822_{\pm.0022}$	$0.3227_{\pm.0014}$

What SHAPE predicts. SHAPE measures the frozen historical decoder's uncertainty over neural-to-phoneme alignment paths for a recorded trial and its attempted phoneme sequence. Its value depends on both the neural recording and the historical decoder. Higher entropy can reflect cross-day representational mismatch, but also trial difficulty and decoder uncertainty; we therefore use it as a proxy for cross-day representational mismatch. Historical recordings supply the SHAPE targets used to train the textual surrogate below. Our assumption is that the learned relationship between linguistic content and relative SHAPE remains informative on future days, while absolute SHAPE values may vary across days.

Budget, strata, and randomness. Here, M is determined by the available corpus, while k is a user-specified recording budget. The two boundaries are empirical tertiles recomputed from candidate scores without target-day neural statistics; three strata are a fixed design choice. We use $\alpha = 0.8$ and $\beta = 0.15$ throughout, fixed without inspecting Protocol 2 target-day data or results and transferred to T15 without retuning. Given an estimator and corpus, stratum assignment is deterministic; randomness enters through sampling within strata. The quotas remain fixed while the sampled sentence list may vary, allowing selection to exploit an imperfect ranking without repeatedly prescribing a fixed list.

Future-day predictability. We evaluate whether the historical estimator predicts relative SHAPE on future days. On held-out Days 10–12, Table 3 compares predicted and oracle SHAPE under identical full fine-tuning: mean PER is 0.4301 versus 0.4308, and selected-subset overlap averages 78%. The subset overlap measures agreement between the selected calibration sets, with the estimator fixed across target days. Separately, replacing historical SHAPE targets with historical sentence-level PER while keeping textual inputs, architecture, and downstream full fine-tuning fixed increases five-day mean PER from 0.4224 to 0.4454 (Appendix F.2), supporting alignment entropy over this simpler neural-derived target.

C Linguistic-based Acquisition Comparison Details

We compare text-only acquisition rules on T12 Protocol 2 to isolate sentence-selection quality, with daily results reported in Table 6 in the main paper. All methods use only candidate linguistic information before recording and receive the same full fine-tuning afterward. Len prioritizes character count. Cov greedily maximizes inverse-frequency phoneme coverage with harmonic discounting of repeated occurrences. Bi uses farthest-first selection on TF–IDF-weighted, L2-normalized phoneme-bigram vectors. Emb uses k-means++ on RoBERTa-base CLS embeddings with squared Euclidean distance. SHAPE-guided acquisition improves on the strongest of these heuristics, Cov, by 0.0163 mean PER.

D Evaluation on Participant T15

T15 from Brain-to-Text '25 [5] provides a second participant with a different implant configuration, corpus, and feature dimensionality. We train the decoder on Days 1–5 and the estimator on Days 6–7, and evaluate on Days 10–14. Sessions contain 100–200 sentences, so the acquisition budget is adjusted to $k = 50$; all other hyperparameters transfer from T12 without tuning. We assess decoding gains across the five held-out T15 target days. As shown in Table A2, HeDA^2 improves on RandomFT on all five days, reducing mean PER from 0.3861 to 0.3458, and approaches its oracle-acquisition

counterpart (0.3430). The decoder and estimator are trained on T15's historical recordings, using the hyperparameter settings from T12. The five paired day-level RandomFT-minus-HeDA^2 differences yield a 95% confidence interval of $[0.0192, 0.0613]$ ($t(4) = 5.30$, $p = 0.006$). The gains are consistent across the five target days for T15, with both models trained on this participant's historical recordings.

E Comparison with FullFT

Across the 12 target days in Protocols 2 and 3, HeDA^2 averages 0.3568 PER versus 0.3452 for FullFT, a mean gap of 0.0116. Relative to NoFT, this recovers approximately 94.75% of FullFT's improvement using 100 attempted sentences, approximately one fifth of the full training set. HeDA^2 has lower PER on six days: the differences on Days 10, 14, and 16 are below 0.003, while Days 12, 17, and 18 show larger reductions (0.4128 vs. 0.4200, 0.3308 vs. 0.3365, and 0.3500 vs. 0.3583). Defining paired differences as FullFT minus HeDA^2, the mean is -0.0116 (95% CI $[-0.023179, 0.000046]$, $t(11) = -2.19$, $p = 0.0508$); the exact paired Wilcoxon test gives $p = 0.1763$. These descriptive tests summarize the paired variability across repeated target-day evaluations within T12. Together with the mean gap, they characterize the accuracy–acquisition trade-off underlying the recovery of most of FullFT's improvement.

We vary the number of attempted calibration sentences to examine the accuracy–acquisition trade-off. As shown in Table 5, HeDA^2 with $k = 150$ achieves 0.3846 PER, compared with 0.3889 for RandomFT at $k = 300$. At $k = 200$ on Protocol 2, it reaches 0.3741, below the FullFT reference of 0.3829. These budget-specific comparisons demonstrate the combined benefit of selection and adaptation in reducing the number of attempted calibration sentences. Sentence count is the acquisition-burden measure used in this evaluation.

F Additional Ablation Studies

F.1 Independent Contribution of the Semantic Adapter

To isolate the contribution of semantic correction, we compare AA with full fine-tuning, AA with semantic correction alone, AA with spatial correction alone, and AA with both corrections on Days 10–12. As shown in Table 4 in the main paper, semantic correction alone reduces mean PER from 0.4301 to 0.4198 and improves two of the three days. Adding semantic correction to the spatial configuration reduces mean PER from 0.4151 to 0.4021. Thus, semantic correction improves decoding both on its own and when combined with spatial correction. "Semantic" denotes the adapter applied to attention and feed-forward modules.

F.2 Ablation on the Regression Target of the Textual Estimator

In the main paper we used **S**entence-level **H**eterogeneity via **A**lignment **P**ath **E**ntropy (SHAPE) as the regression target for training the textual estimator g_ϕ. A natural alternative is to train g_ϕ against sentence-level PER computed on historical recordings, thereby directly predicting downstream decoder error instead of alignment uncertainty. We instantiate this alternative by replacing SHAPE with PER as the regression target of g_ϕ while keeping the identical architecture and training recipe; the downstream pipeline (stratified acquisition followed by full fine-tuning) is also unchanged.

We compare HeDA^2-Select + FullFT driven by the two proxies on Protocol 2 to assess the effect of the estimator's training target. As shown in Table A3, SHAPE yields lower downstream PER on every reported target day and reduces mean PER from 0.4454 to 0.4224. With the estimator inputs, architecture, and downstream adaptation held fixed, this improvement supports alignment-path uncertainty as an informative training target for pre-acquisition stimulus ranking.

F.3 Ablation on the Acquisition Strategy

In the main paper we used a stratified sampling scheme that allocates the acquisition budget across three SHAPE strata with quotas (k_L, k_M, k_H) governed by probabilities α and β (see Section 4.3). A natural question is whether such stratification is truly necessary, or whether simpler alternatives such as selecting the k sentences with the highest SHAPE (top-k) would suffice.

Table A3: Ablation on the regression target of g_ϕ on Protocol 2 with $k = 100$. Both variants use HeDA^2-Select + FullFT; only the training target of g_ϕ differs. Avg. is computed over Days 10–14.

Target of g_ϕ	Day 10	Day 11	Day 12	Day 13	Day 14	Avg.
PER proxy (alternative)	$0.4369_{\pm.004}$	$0.4298_{\pm.003}$	$0.4576_{\pm.003}$	$0.4848_{\pm.005}$	$0.4177_{\pm.003}$	$0.4454_{\pm.003}$
SHAPE proxy (ours)	$0.4283_{\pm.004}$	$0.4208_{\pm.003}$	$0.4412_{\pm.002}$	$0.4219_{\pm.006}$	$0.3997_{\pm.003}$	$0.4224_{\pm.002}$

We compare three acquisition rules at $k = 100$ on Protocol 2: deterministic Top-k selects the highest-scoring sentences, Uniform allocates equal quotas across the three SHAPE strata, and the default stratified rule uses $\alpha = 0.8$ and $\beta = 0.15$. Predicted SHAPE and downstream heterogeneity-aware adaptation are held fixed. Top-k selects the highest k scores deterministically, whereas sampling within the high stratum draws from all sentences in that stratum.

We compare downstream PER to assess the effect of the three acquisition rules. As shown in Table A4, the default stratified allocation achieves 0.3959 mean PER, compared with 0.4300 for Top-k and 0.4437 for Uniform. This improvement supports allocating most of the budget to high predicted heterogeneity while retaining samples from the other strata. The default assigns 80% of the budget to the high stratum, 15% to the medium stratum, and the remainder to the low stratum.

Table A4: Acquisition-strategy comparison on T12 Protocol 2 with $k = 100$. All methods share predicted SHAPE and heterogeneity-aware adaptation. Mean PER (↓) is over Days 10–14.

Acquisition rule	Mean PER
Top-k	0.4300
Uniform ($\alpha = \beta = 1/3$)	0.4437
Stratified ($\alpha = 0.8$, $\beta = 0.15$)	**0.3959**

G Additional Interpretability Analysis

We characterize the 100 sentences with the highest predicted SHAPE on each T12 target day in Protocols 2 and 3 and compare them with 100 randomly selected sentences. This diagnostic top-100 selection describes the high-scoring portion of the candidate corpus; the deployed HeDA^2 acquisition rule samples across three score strata. We examine phoneme counts and recording time, phoneme distributions and coverage, and gradient alignment and within-set structure. These analyses complement the score-decile gradient alignment in Figure 4.

G.1 Phoneme Counts and Recording Time

We compare phoneme counts and recording-time characteristics to characterize the content of the two selections. As shown in Figure A1, high-SHAPE sentences contain longer phoneme sequences and have longer recording windows, while their mean recording frames per phoneme is lower on every target day. These measurements indicate more phonetic content per recorded frame, alongside longer mean recording windows.

G.2 Phoneme Distribution and Coverage

We examine phoneme-frequency preferences by comparing token shares between the two selections under Protocols 2 and 3, as shown in Figure A2(a-i,a-ii). For each target day, phonemes are ranked from most to least frequent in the candidate pool. High-SHAPE selection modestly favors common phonemes: the four most frequent phonemes together account for an additional 2.21 and 1.16 percentage points of selected tokens under Protocols 2 and 3, respectively. Vowel and consonant token shares remain similar to random selection (Figure A2b). Individual-phoneme coverage is nearly complete for both selections (Figure A2(c-i)), while high-SHAPE sentences cover more adjacent phoneme pairs (Figure A2(c-ii)). High-SHAPE selection therefore retains broad single-phoneme coverage while covering more phoneme transitions.

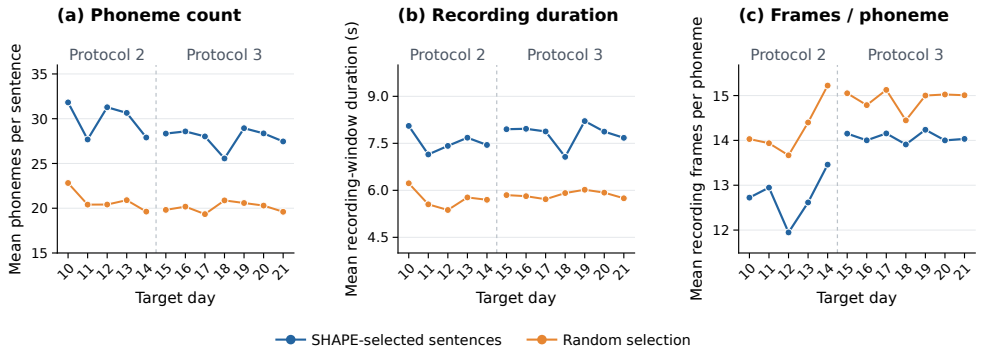


Figure A1: Phoneme counts and recording time of the 100 highest-predicted-SHAPE and 100 randomly selected sentences per target day. (a) Mean phoneme count per sentence. (b) Mean recording-window duration. (c) Mean recording frames per phoneme.

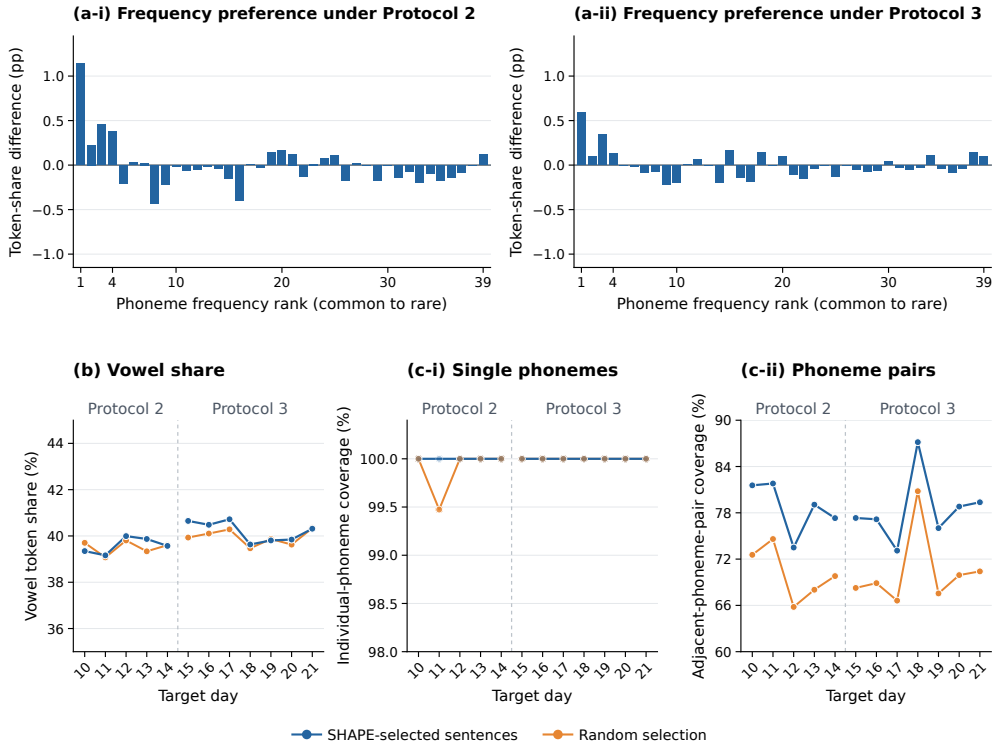


Figure A2: Phoneme distribution and coverage for the diagnostic top-100 and random selections. (a-i,a-ii) Token-share differences (High-SHAPE minus Random) for phonemes ranked by candidate-pool frequency under Protocols 2 and 3, respectively. (b) Vowel token share. (c-i) Individual-phoneme coverage. (c-ii) Adjacent-phoneme-pair coverage.

G.3 Gradient Alignment and Within-Set Structure

For each sentence, we compute the output-head gradient before adaptation and measure its cosine similarity to the mean gradient of the 40 same-day test sentences. These test gradients supply a retrospective reference for interpretation; pre-acquisition scoring uses the historical textual estimator. Figure 4 groups candidate sentences into predicted-SHAPE deciles and averages the individual gradient similarities within each decile. Faint dots represent daily decile means, and solid lines average them with equal weight across target days. Higher-scoring groups generally show greater

similarity, with a more pronounced increase under Protocol 2 (panel a) than Protocol 3 (panel b). This pattern associates predicted SHAPE with gradient directions relevant to target-day decoding.

We compare gradient properties to characterize the update directions of the top-100 and random selections. As shown in Figure A3, high-SHAPE sentences have smaller mean gradient norms, greater mean cosine similarity to the target-day reference, and lower within-set effective ranks across the evaluated days. Effective rank is the participation ratio of the Gram matrix formed from unit-normalized sentence gradients. Together, these measurements indicate that high-SHAPE selection emphasizes a more concentrated set of gradient directions with greater target-day alignment. The stratified rule combines emphasis on this high-scoring region with coverage of the remaining score strata, consistent with the acquisition-strategy results in Table A4.

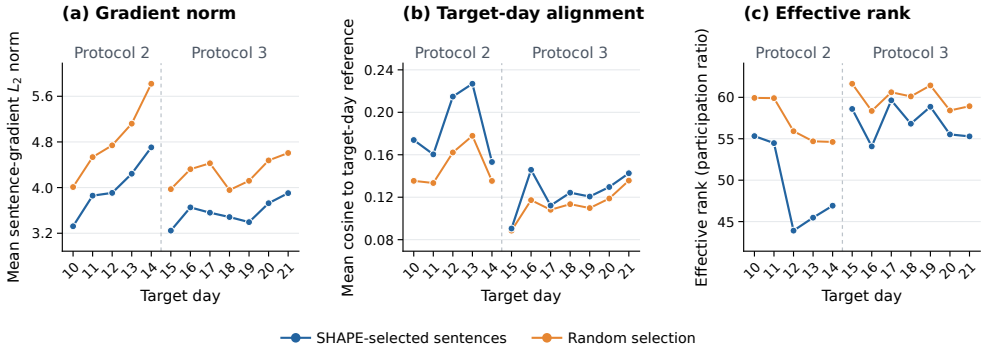


Figure A3: Gradient properties of the 100 highest-predicted-SHAPE and 100 randomly selected sentences per day. (a) Mean sentence-gradient L_2 norm. (b) Mean cosine similarity to the retrospective target-day reference. (c) Effective rank of the unit-normalized sentence-gradient Gram matrix.

H Extended Discussion

Why Active Acquisition is the right framing. Cross-day degradation motivates both efficient acquisition and effective adaptation. Fine-tuning and unsupervised adaptation operate on recordings after they exist, while Active Acquisition chooses the stimuli that elicit them. It is therefore complementary to within-subject transfer learning. The acquisition-only experiments hold full fine-tuning fixed; the complete system additionally changes adaptation. These comparisons distinguish the gain from stimulus selection from the combined gain of selection and adaptation.

Limitations. First, our experiments cover two participants, with historical training performed separately for each. Evaluation on additional participants and other recording modalities is left to future work. Second, the estimator depends on historical relationships between linguistic content and neural-derived alignment uncertainty. Abrupt changes, including channel loss, surgical revision, or device replacement, can invalidate these relationships; spatial correction cannot reconstruct a missing channel. Third, SHAPE requires CTC path distributions, while the adapters require suitable convolutional and attention/feed-forward modules. Fourth, offline phoneme error rate does not measure conversational throughput or user-perceived latency. Calibration effort is reported as attempted sentences, with recording-window diagnostics in Appendix G; actual full-session duration is unavailable, and equal sentence counts need not imply equal recording time. Closed-loop operation and inference overhead have not been benchmarked. A sustained increase in alignment-path entropy on prescribed stimuli could motivate an online failure monitor, but its sensitivity and thresholds require validation.

Future work. Several directions follow naturally from the paradigm. (i) Extending Active Acquisition to multi-participant settings, where a textual estimator trained across participants could amortize the historical-data requirement for a new user. (ii) Coupling Active Acquisition with online closed-loop adaptation, so that the acquired calibration set itself is refined as new sentences are recorded

within a session. (iii) Generalizing the heterogeneity proxy beyond CTC alignment uncertainty, for instance to encoder–decoder architectures with attention-based alignments, to support a broader class of neuroprosthetic applications.

Broader impact. Reducing the number of attempted calibration sentences could make long-term speech neuroprostheses more practical. The present offline experiments measure the trade-off between the number of calibration sentences and decoding PER. Deployment still requires informed consent, appropriate oversight, patient-data protection, and monitoring for shifts outside the historical training distribution. These considerations apply to both reused participant datasets.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state three claims: (a) Active Acquisition is a new calibration paradigm that decides which stimuli to record before any target-day recording occurs; (b) HeDA^2 instantiates this paradigm via Sentence-level Heterogeneity via Alignment Path Entropy, a textual estimator with stratified acquisition, and decoupled spatial/semantic adaptation; and (c) on T12, HeDA^2 recovers most of FullFT's improvement using approximately one fifth of the calibration sentences, with additional budget-matched gains on T15. These claims are supported by Sections 4.4–5.2, Tables 1–6, and Appendix D.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix H (Limitations paragraph) explicitly discusses evaluation on two participants with participant-specific training, the dependence of the textual estimator on stationary historical heterogeneity, the assumption of a decoder containing convolutional and linear/attention modules, and the use of PER as a proxy for communication quality.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not introduce new theoretical results; it proposes an empirical paradigm (Active Acquisition) and a framework (HeDA^2) evaluated experimentally.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5.1 specifies the datasets (T12 from Brain-to-Text '24 and T15 from Brain-to-Text '25), the T12 cross-day transfer protocols and T15 split, the decoder backbone (with full architectural details in Appendix A), and the implementation details for the historical decoder, the textual SHAPE estimator, the stratified acquisition, and the heterogeneity-aware adaptation, all of which are necessary to reproduce the reported numbers.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The Brain-to-Text Benchmark '24 dataset [2] is publicly available through the DRYAD repository as cited in Section 5.1. We released code at https://github.com/xxchengxi/HeDA2; all hyperparameters required to reproduce the experiments are specified in the main paper and Appendix A.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Section 5.1 reports the data splits induced by the three T12 protocols and the T15 evaluation, the optimizer (AdamW), the learning-rate schedule, the number of epochs, the regularization coefficients, and the stratified acquisition probabilities α, β. Additional details about the Conformer backbone are in Appendix A.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Day-level main and ablation results report mean and standard deviation across five distinct random seeds for each (method, target day) cell. Table 6 reports daily results and cross-day means for the text-only acquisition comparison. The standard deviations capture variability due to model initialization and stochastic optimization while the data splits are held fixed.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The Conformer backbone described in Appendix A contains roughly 8M parameters and runs on a single consumer GPU; per-day calibration with $k = 100$ takes on the order of a few minutes. Total compute, including all baselines, ablations, and seed repetitions, is on the order of a few hundred GPU-hours and is reported alongside the released code.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses existing, IRB-approved, publicly released intracranial datasets [2, 5], and does not collect new human-subject data. No personally identifying information is used.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix H (Broader impact paragraph) discusses positive impact in the form of reduced daily recording burden for patients with severe communication impairments, and notes that practitioners deploying the paradigm in clinical settings should monitor the textual estimator for distribution shift; we do not foresee direct negative societal impacts beyond the standard considerations that apply to all medical-grade BCIs.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any high-risk dataset or generative model; it releases task-specific decoder code that operates on a previously released, IRB-approved intracranial dataset and has no plausible misuse pathway beyond the original data.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The Brain-to-Text Benchmark '24 dataset is cited via [2] and used in accordance with its DRYAD terms of use. The Conformer architecture is cited via [41].

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The HeDA^2 framework is documented in Sections 4.2–4.4 and the released codebase will include configuration files, training scripts, and a README that documents all hyperparameters listed in Section 5.1 and Appendix A.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not collect any new human-subject data and does not involve crowdsourcing; it uses previously released data from T12 and T15 [2, 5].

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We do not collect new human-subject data. The IRB approvals and informed consent for the reused T12 and T15 data are documented in the original publications [2, 5].

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology of HeDA^2. A pre-trained language-model encoder is used only as the textual feature extractor inside the SHAPE estimator g_ϕ (Section 5.1); it is a standard, fixed component and not a core methodological contribution.