Predictability Is Not Provenance: Rethinking Next-Word Prediction in LLM-Based Neural Encoding

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM)-based semantic encoding studies often interpret pre-onset performance in target-aligned lag–r curves as evidence that the brain predicts upcoming words. We show that this inference has two limitations. First, contextual representations of successive words are statistically dependent, while responses elicited by those words overlap in the measured signal. Conventional target-word encoding can therefore conflate the target contribution with covariance-weighted contributions from historical sources, making predictive success ambiguous with respect to neural provenance. Second, held-out correlation quantifies aggregate predictive adequacy but does not identify which word generated predictable activity or how its contribution evolves from onset. We formalize this ambiguity with a constrained-availability source-time finite impulse response model (CAST-FIR) and derive Target Isolation from it as a source-adjustment procedure. A strictly causal simulation and a passive acoustic benchmark produce prospective-looking performance without anticipatory computation. Across English ECoG and Chinese SEEG, Target Isolation contracts the $r > 0.075$ cross-onset temporal-generalization area by $98.1\%/100.0\%$, respectively. Fitted source-relative kernels peak at 75 ms in English and 225 ms in Chinese after source onset. These results do not rule out neural prediction. They show that target-aligned predictability alone does not establish target-word provenance: performance should establish predictive adequacy, whereas temporal interpretation requires explicit source-time structure.

1 INTRODUCTION

Neural encoding models predict brain activity from stimulus features, providing a computational framework for relating external information to neural responses (Naselaris et al., 2011; Holdgraf et al., 2017). Early language encoding studies focused on acoustic envelopes, spectrograms, phonemes, and articulatory features, with subsequent work extending to lexical and syntactic representations (Mesgarani et al., 2014; Brodbeck et al., 2018). More recently, LLM-based encoding has extended this progression to semantic representations. By integrating long-range context, token-level embeddings enable the study of mappings between contextual semantic information and neural activity (Jain & Huth, 2018; Schrimpf et al., 2021; Goldstein et al., 2022; 2024; 2025).

This paradigm treats word i as the target word and uses its representation x_i to predict the neural response at time $O_i+\tau$, where O_i denotes its onset and τ is the relative lag (Fig. 1A). Fitting a separate model for each τ and computing the held-out Pearson correlation yields a lag–r curve. Such lag-resolved analyses are used to infer whether and when information about the target word is represented in the measured neural activity. One striking pattern in prior work is that encoding performance begins to rise and may peak before target-word onset (Fig. 1B). Such pre-onset performance has been interpreted as a predictive signal, suggesting continuous next-word prediction in the brain (Goldstein et al., 2022). However, we identify two fundamental limitations of this paradigm.

The first, termed **Limitation I: Temporal Source Mixing**, arises from two interacting factors. First, **sequential dependence**: contextual embeddings explicitly integrate preceding context, thereby introducing statistical dependence among representations of successive words (Fig. 1C). Second, **neural-response superposition**: neural responses elicited by preceding stimuli persist and overlap in the measured signal (Fig. 1D). Together, these factors allow the target-word representation to predict a lingering response elicited by preceding words through cross-word covariance, thereby inflating

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

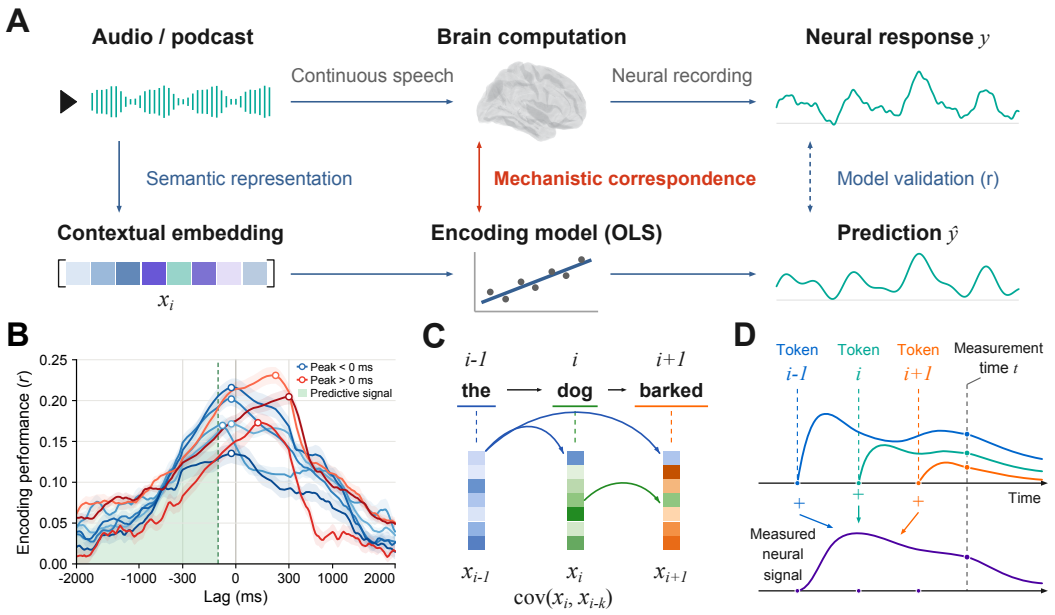


Figure 1: **Semantic encoding paradigm and its limitation. A**, the encoding model utilizes contextual embeddings to predict neural responses, and held-out performance validates predictions rather than their provenance. **B**, Lag–r curves for eight ECoG subjects using public data from Bhattacharjee et al. (2026). The three lag intervals are shown at equal width to clarify peak timing. **C**, sequential dependence mixes information from preceding words into each contextual embedding. **D**, overlapping responses superpose neural contributions evoked by preceding words.

apparent target-specific encoding performance. Several studies have proposed geometric projection or linear residualization controls (Goldstein et al., 2022; Azizpour et al., 2026; Schönmann et al., 2026), but these feature-side transformations constrain only particular forms or ranges of sequential dependence and may compromise the semantic information encoded in the target representation.

The second, termed **Limitation II: Performance or Model?**, concerns the object of scientific interpretation. Prior work has often treated the rise and peak of a lag–r curve as a temporal signature of neural processing. However, encoding performance only measures how well the model linearly predicts held-out neural responses from the target word's semantic representation. The curve therefore reflects only the model's predictive performance across lags, not the neural mechanisms that generate the observed response. Predictive adequacy is nevertheless a prerequisite for interpretation: a model must capture generalizable structure in neural responses before its fitted structure can meaningfully inform hypotheses about neural computation. This creates a trade-off between predictive capacity and interpretability. Flexible models may achieve high performance while obscuring how their predictions are formed, whereas simple models may be transparent but insufficiently expressive to capture the data. An informative encoding model must therefore be both predictively adequate and structurally interpretable. Performance validates a model; model structure determines what it can explain.

In this work, we show that conventional target-word encoding conflates target and covariance-weighted historical contributions, so a lag–r curve cannot by itself establish neural provenance. We formalize this temporal source mixing and introduce CAST-FIR, which preserves contextual representations while jointly modeling all sources available at each query time in source-relative coordinates. In a strictly causal ground-truth simulation, sequential dependence and response superposition generate a smooth pre-onset predictive signal without anticipatory computation; a passive acoustic benchmark shows that source adjustment attenuates prospective-looking performance without removing inter-token dependence. Across English ECoG and Chinese SEEG, Target Isolation markedly contracts apparent cross-onset temporal generalization. Fitted source-relative kernels peak only after source onset. Together, these results do not rule out neural prediction; they show that performance should establish predictive adequacy, whereas temporal interpretation requires model structure that makes source identity and source-relative time explicit.

2 THEORY: TEMPORAL MIXING AND SOURCE-TIME ATTRIBUTION

2.1 WHAT CONVENTIONAL SEMANTIC ENCODING ESTIMATES

Our analysis begins with a premise: neural responses elicited by successive words can overlap in time and superpose additively in the measured signal. Conventional analysis is anchored to target word i, whereas j indexes any source word that may contribute to the signal at absolute time t. Let O_j and $x_j \in \mathbb{R}^p$ denote its onset time and centered semantic representation. We formalize this premise as

$$S_e(t) = \sum_j x_j^\top h_e(t - O_j) + u_e(t) + \epsilon_e(t), \qquad h_e(a) = 0 \quad \text{for } a < 0, \tag{1}$$

where $h_e(a) \in \mathbb{R}^p$ is the response kernel at contact e, specifying how semantic features contribute linearly to the expected neural response at source-relative lag $a = t - O_j$, the elapsed time since that source's onset. The terms $u_e(t)$ and $\epsilon_e(t)$ collect unmodeled systematic activity and residual noise.

To mirror the conventional encoding pipeline used in prior ECoG and MEG studies, neural activity is averaged over a fixed-width sliding window and a separate target-word encoder is fit using ordinary least squares (OLS) at each target-relative lag τ (Goldstein et al., 2022; 2024; 2025; Azizpour et al., 2026; Schönmann et al., 2026). Let $y_{i,e}(\tau)$ denote the resulting response at contact e, and let $\bar{h}_{e,W_y}(a)$ denote $h_e(a)$ averaged over the same centered window of width W_y. Defining the onset gap $\Delta_{i,k} = O_i - O_{i-k}$ and retaining the preceding K sources gives

$$y_{i,e}(\tau) = x_i^\top \bar{h}_{e,W_y}(\tau) + \sum_{k=1}^{K} x_{i-k}^\top \bar{h}_{e,W_y}(\tau + \Delta_{i,k}) + \eta_{i,e}^{(K)}(\tau). \tag{2}$$

Thus, the neural response is the sum of the target-word response, the window-averaged responses to preceding sources at their corresponding source-relative lags, and a remainder $\eta_{i,e}^{(K)}(\tau)$ collecting all remaining sources, unmodeled systematic activity, and residual noise.

At each lag, conventional encoding regresses $y_{i,e}(\tau)$ on the target representation x_i. Define

$$C_x = \mathrm{Cov}(x_i, x_i), \qquad \Gamma_k = \mathrm{Cov}(x_i, x_{i-k}), \qquad g_{e,k}(\tau) = \mathbb{E}_{\Delta_{i,k}}\left[\bar{h}_{e,W_y}(\tau + \Delta_{i,k})\right].$$

C_x is the covariance of the representations, Γ_k is their cross-covariance with the k-th preceding source, and $g_{e,k}(\tau)$ is that source's window-averaged response kernel averaged over its onset gaps. Under the idealized conditions detailed in Appendix A.3, the population OLS coefficient is

$$\beta_e^*(\tau) \approx \underbrace{\bar{h}_{e,W_y}(\tau)}_{\text{target source}} + \underbrace{C_x^{-1} \sum_{k=1}^{K} \Gamma_k g_{e,k}(\tau)}_{\text{history mixture}}. \tag{3}$$

We can further derive the lag–r curve. Let $c_e(\tau) = \mathrm{Cov}(x_i, y_{i,e}(\tau))$, and define the target-source and history-mixture components as $A_e(\tau) = C_x \bar{h}_{e,W_y}(\tau)$ and $B_e(\tau) = \sum_{k=1}^{K} \Gamma_k g_{e,k}(\tau)$, respectively. Under the same idealized conditions, we can further derive the theoretical lag–r curve:

$$r_e^2(\tau) = \frac{c_e(\tau)^\top C_x^{-1} c_e(\tau)}{\mathrm{Var}(y_{i,e}(\tau))} \approx \frac{[A_e(\tau) + B_e(\tau)]^\top C_x^{-1} [A_e(\tau) + B_e(\tau)]}{\mathrm{Var}(y_{i,e}(\tau))}. \tag{4}$$

Thus, both the OLS coefficient and the lag–r curve contain the same history mixture. Crucially, the form of sequential dependence relevant to linear encoding is the cross-covariance Γ_k, rather than the pairwise cosine similarity targeted by some prior analyses (Goldstein et al., 2022; Azizpour et al., 2026). Consequently, conventional encoding can overestimate the predictive performance attributed to the target word by conflating target and history contributions. More importantly, the apparent pre-onset predictive signal may itself be an artifact of temporal source mixing. This ambiguity motivates controls that separate the target contribution from the history mixture before interpreting pre-onset predictability. Complete derivations are provided in Appendix A.3.

2.2 WHY EXISTING PREDICTOR-SIDE CONTROLS ARE INSUFFICIENT

Existing controls attempt to suppress history dependence by transforming the target-word predictor. Goldstein et al. (2022) project each target representation onto the orthogonal complement of the

previous-word representation, and Azizpour et al. (2026), recognizing that pre-onset activity may reflect neural-response superposition from earlier words, extend this geometric projection to the preceding eight words. However, pairwise cosine similarity is not the form of sequential dependence that generates the history mixture in Equations 3 and 4; the relevant quantity is the cross-covariance. Moreover, enforcing orthogonality to the preceding eight embeddings does not guarantee removal of temporal source mixing, because more distant historical sources remain uncontrolled.

Schönmann et al. (2026) instead remove the component of each target representation that is linearly predictable from the previous-word representation. This method directly suppresses linear dependence on the previous-word representation but does not account for neural-response superposition across the full history. Even if extended to all preceding words, the residualization operator would change as the history expands, progressively redefining the target representation. At the limit, this redefinition can collapse the representation to zero, eliminating its representational capacity entirely (Appendix A.6).

More fundamentally, all three controls adopt predictor-side removal: they transform contextual embeddings to reduce temporal source mixing. However, transforming the predictor changes the semantic representation itself and can attenuate semantic information carried by the target token in its current context (Appendix D.2, Table 6). Therefore, rather than mitigating temporal source mixing at the cost of semantic information, a safer approach is to jointly model the temporal evolution of contributions from all available source tokens. Full details are provided in Appendix A.6.

2.3 CAST-FIR: Source-Time Encoding

Rather than transforming x_i, CAST-FIR models all words in source-relative time. Let $y_{q,e}$ denote the window-averaged neural response centered at absolute query time t_q, so that $y_{q,e} = y_{i,e}(\tau)$ when $t_q = O_i + \tau$. We partition the nonnegative source-relative-lag range $[0, L)$ into M bins of width δ:

$$B_m = [m\delta, (m+1)\delta), \qquad m = 0, \ldots, M-1, \qquad L = M\delta.$$

We set $\delta = 25\,\mathrm{ms}$ to match the stride of conventional lag-resolved encoding and place both methods on a common temporal grid. With $m(a) = \lfloor a/\delta \rfloor$ denoting the bin containing source-relative lag $a \in [0, L)$, we aggregate source representations by their lag at each query time:

$$z_{q,m} = \sum_j x_j \, \mathbf{1}[t_q - O_j \in B_m], \qquad \widehat{y}_{q,e} = \widehat{\alpha}_e + \sum_{m=0}^{M-1} z_{q,m}^\top \widehat{\theta}_{e,m}. \tag{5}$$

Each $\theta_{e,m} \in \mathbb{R}^p$ parameterizes the neural contribution of sources in lag bin B_m, with all bins jointly estimated by ridge regression as a single source-relative-lag kernel. Conceptually, this kernel draws on the FIR framework to approximate the underlying source-locked neural response mechanism. Target Isolation uses these modeled contributions for response-side adjustment (Appendix A.8):

$$\widehat{y}_{i,e}^{\,\text{non-target}}(\tau) = \sum_{\substack{j \neq i \\ 0 \le t_q - O_j < L}} x_j^\top \widehat{\theta}_{e,m(t_q - O_j)}, \qquad \widetilde{y}_{i,e}^{\,\text{target}}(\tau) = y_{i,e}(\tau) - \widehat{y}_{i,e}^{\,\text{non-target}}(\tau). \tag{6}$$

Equation 6 defines Target Isolation by subtracting cross-fitted modeled contributions from available non-target words. We do not claim that this procedure recovers a clean neural response generated solely by the target word. The adjusted response remains a model-dependent residual containing unmodeled activity, noise, and source-estimation error, and should therefore be interpreted as a model-based characterization of the observed neural response. Because this adjustment leaves the full contextual embedding unchanged, it preserves semantic information while attenuating temporal source mixing. Complete details are provided in Appendix A.7.

3 Limitation I: Temporal Source Mixture

Equations 3 and 4 show that, under conventional encoding, the fitted coefficient and lag–r curve conflate the target contribution with covariance-weighted contributions from preceding words. This raises a critical question: could the reported pre-onset predictive signal (Goldstein et al., 2022) be an artifact of temporal source mixing? In principle, sequential dependence provides a statistical path from the target-word representation to neural responses elicited by preceding words, while neural-response superposition allows these responses to persist and overlap in the measured signal.

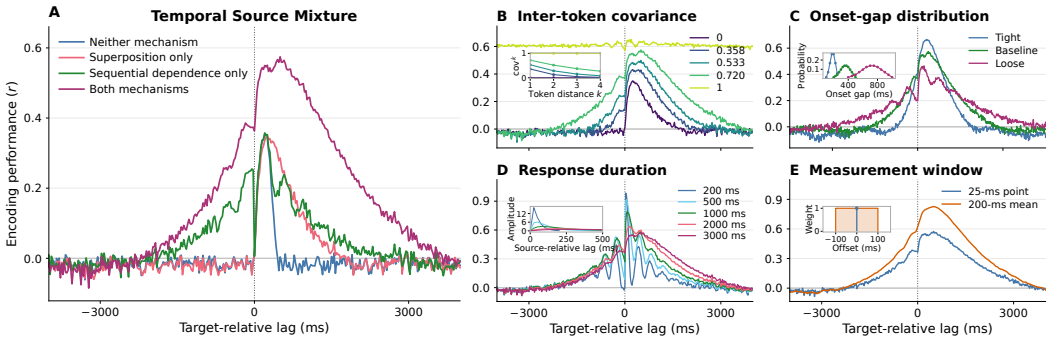


Figure 2: **Synthetic data showed that pre-onset encoding can be enabled by sequential dependence, and amplified by response superposition. A**, the two mechanisms underlying temporal source mixing are isolated in a prediction-free generator. **B–E**, stronger covariance, tighter onsets, longer responses, and temporal averaging each increase source mixing.

In empirical neural recordings, these factors covary and the true source contributions are not directly observed. We therefore use a ground-truth simulation for two purposes. First, it allows each determinant of temporal source mixture to be manipulated while all other variables are held fixed. Second, it constructs a strictly causal, prediction-free system in which no word contributes before its own onset. We sequentially generate word onsets and six-dimensional source representations along an absolute time axis, with a dependence parameter ρ controlling adjacent-word covariance. Each word that has already occurred is then mapped into a synthetic neural response through a one-sided response kernel, ensuring that the simulated system contains no predictive contribution from future words. Full details of the generator and parameterization are provided in Appendix B.

If the conventional paradigm still reproduces the so-called predictive signal in such a prediction-free system, it is insufficient to support claims of continuous next-word prediction in the brain.

3.1 How Temporal Source Mixing Generates an Apparent Predictive Signal

To isolate the two components of temporal source mixture, we separately manipulated sequential dependence and neural-response superposition. Sequential dependence was either removed by setting $\rho = 0$ or retained at a fixed nonzero level ($\rho = .72$ by default), while neural-response superposition was suppressed by truncating responses at the next-word onset or retained by allowing them to persist and overlap. We then used the target-word representation to perform conventional OLS encoding.

As shown in Figure 2**A**, when word representations were independent, the target representation predicted only the response elicited by the target word itself (blue and pink curves). Negative-lag peaks therefore remained near zero regardless of response overlap ($r = .010$ and $.007$, respectively), showing that neural-response superposition alone does not create cross-word predictability.

Once sequential dependence was introduced, substantial pre-onset performance emerged even without neural-response superposition (green curve), with its maximum rising from .010 to .255. This occurs because cross-covariance allows the target-word representation to predict neural responses elicited by preceding words. When neural-response superposition was additionally introduced (purple curve), persistent responses from multiple source words overlapped in absolute time. This overlap smoothed the distinct source-specific peaks seen in the green curve and allowed multiple source contributions to jointly support target-aligned predictability. As a result, the maximum pre-onset performance and the peak over the full curve increased by 51.4% and 61.1%, respectively.

Crucially, when both mechanisms were present, conventional encoding yielded a lag–r curve resembling the empirical pattern in Figure 1**B**, including the predictive signal. Sequential dependence enables the target-word representation to predict neural responses elicited by other words, while neural-response superposition causes these responses to overlap and sum in the measured signal. Together, these mechanisms caused conventional encoding performance to reflect a conflation of contributions from temporally distinct sources, thereby inflating the apparent target-specific predictability. Pre-onset encoding performance is therefore not, by itself, sufficient evidence for neural prediction.

3.2 What Shapes a Lag–r Curve?

The preceding experiment establishes that temporal source mixing is sufficient to generate what has been interpreted as a predictive signal. We next used the same ground-truth generator to dissect how different components of the conventional encoding pipeline shape the resulting lag–r curve.

Sequential dependence gates access to responses from other sources. Figure 2**B** varies only the dependence between adjacent word representations. We tested five dependence levels: $\rho = 0$ for independence; $.358$ and $.533$ empirically estimated from the English and Chinese datasets; $.720$ as the simulation default; and 1 for maximal dependence. As covariance increased, encoding performance rose throughout the lag–r curve. At $\rho = 1$, all word representations within each story are identical, making responses elicited by every source predictable from the target representation and flattening the lag–r curve into a nearly horizontal line centered at $r = .610$. This loss of target-aligned temporal structure exposes the limits of source attribution under the conventional paradigm: high performance can persist even when the target is indistinguishable from all other sources.

Neural-response superposition shapes how source contributions accumulate. Figure 2**C–E** examines three factors that control the effective degree of neural-response superposition. Onset timing serves as an indirect proxy for the degree of neural-response superposition (Figure 2**C**): increasing the mean onset gap from 186 ms to 360 ms and 719 ms progressively reduces overlap, separates source-specific peaks, and makes the lag–r curve approach the no-superposition pattern in Figure 2**A**.

Response duration provides a direct manipulation of neural-response superposition (Figure 2**D**). With integrated response strength fixed, extending the response from 200 to 3000 ms spreads the same contribution over time (see Appendix B.3, Figure 6**A**, for results under an alternative assumption). This increases overlap between successive-word responses and merges distinct source-specific peaks, producing a progressively smoother lag–r curve. At the same time, the response becomes less concentrated at any time point, reducing post-onset peak encoding performance from $.985$ to $.575$.

The measurement operator can further increase temporal mixing (Figure 2**E**). Replacing a 25-ms point sample with the 200-ms centered average commonly used in conventional encoding mixes activity across ± 100 ms, producing a smoother and uniformly higher lag–r curve: the maximum negative-lag r increases from $.386$ to $.596$, and the full-curve peak from $.575$ to $.827$.

Together, these manipulations show that stronger neural-response superposition merges source-specific contributions into a smoother lag–r curve and can increase target-aligned performance as contributions accumulate. This superposition obscures source attribution: both the temporal profile and magnitude of encoding performance may reflect responses to other words rather than the target.

Summary and Implications. Overall, the height, width, smoothness, and peak location of a lag–r curve are jointly shaped by temporal source mixing. Encoding analyses should therefore explicitly account for this confound when making target-specific claims. This also motivates a re-evaluation of prior findings: how much of the reported target-aligned performance can be attributed to temporal source mixing, and how do the conclusions change after controlling for it? More fundamentally, we ask whether temporal interpretation should rely on changes in predictive performance or on the interpretable structure of the model itself. These considerations lead to Limitation II.

4 Limitation II: Performance or Model?

This section concerns the distinction between model validation and interpretation. At each lag, $r(\tau)$ quantifies linear predictive accuracy and thus validates the encoding model, but does not identify whether the predicted response is attributable to word i or how predictive credit is distributed across sources and response ages (Figure 1**A**). We therefore argue that mechanistic interpretation should instead come from model components that explicitly parameterize the dimensions of interest. CAST-FIR follows this principle by using aggregate performance for validation and source identity and source-relative lag for temporal interpretation.

We first analyzed the public English ECoG dataset, comprising eight participants who listened to the same approximately 30-minute podcast, with 5,013 analyzable word events and 184 valid contacts (Goldstein et al., 2022; Bhattacharjee et al., 2026). We then collected a Chinese SEEG

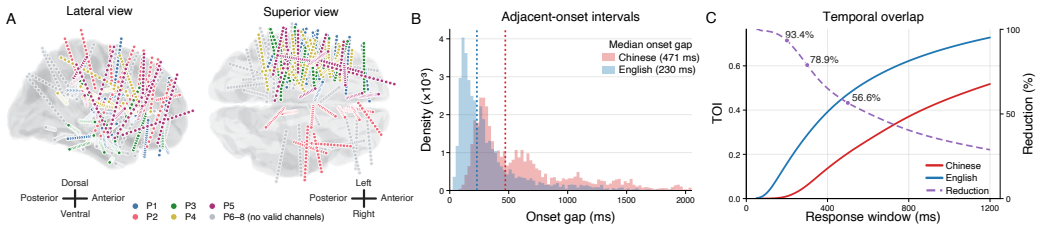


Figure 3: **Slower Chinese speech reduces neural-response superposition relative to English. A**, template-space coverage of the Chinese SEEG recordings. **B**, Chinese stimuli have longer adjacent-onset intervals. **C**, longer gaps yield a lower Temporal Overlap Index across response windows.

dataset to extend the analysis across language, recording modality, and degree of neural-response superposition. Eight participants completed two listening sessions; the primary analysis included five participants and 46 valid contacts (Figure 3**A**). Motivated by Figure 2**C**, the six narrative stories (2,634 words) were presented at $0.45\times$ the original speed to increase inter-word onset intervals (Figure 3**B**) while preserving continuous speech, thereby reducing expected neural-response superposition.

We quantify this timing difference using the Temporal Overlap Index (TOI), a proxy for expected neural-response superposition. Let Δ_i denote the onset gap between adjacent words and W an assumed single-event response duration:

$$\mathrm{TOI}(W) = \frac{1}{N}\sum_{i=1}^{N}\frac{\max(0, W-\Delta_i)}{W}. \tag{7}$$

At $W = 500$ ms, TOI was 20.5% for Chinese and 47.2% for English. Intuitively, under a 500-ms response duration, the Chinese stimulus timing yields 56.6% less expected neural-response superposition than the English stimulus timing (Figure 3**C**). Full details are provided in Appendix C.

4.1 A Passive Control System Without Prediction

We first evaluate the controls introduced in Section 2 in a passive system with no anticipatory computation. Following Schönmann et al. (2026), we use speech acoustics as a passive benchmark to test whether putative predictive signals arise in the stimuli underlying both the English ECoG and Chinese SEEG recordings. Unlike neural activity, speech acoustics provide a passive control that preserves stimulus dependencies but contains no intrinsic mechanism for predicting future acoustic features. We represent the audio using onset-held, nine-dimensional word-level Mel features and apply the same target alignment as in the neural analysis. In this benchmark, lower pre-onset r indicates less residual temporal source mixture after control. Full details are provided in Appendix D.

Figure 4 compares conventional OLS, three predictor-side controls, and the Target Isolation defined in Equation 6. Conventional OLS predicted pre-onset acoustic responses in both languages, with mean held-out $r = .0713/.1376$ for Chinese/English. Target Isolation and linear residualization yielded the strongest reductions, to **.0295**/.0693 and .0314/**.0555**, respectively.

Despite their comparable pre-onset attenuation, linear residualization has three limitations: its operator changes as history expands, the transformation can degrade semantic information, and extending it to the full history may collapse the representation (Section 2.2). Accordingly, lower performance under a predictor-side control is ambiguous: it may reflect reduced temporal dependence, attenuation of semantic information, or both; frozen-representation benchmarks confirm the latter (Appendix D.2). By contrast, CAST-FIR accounts for all available non-target words through a common source-time model, allowing the resulting Target Isolation procedure to achieve similar attenuation without changing the semantic representation (Section 2.3). These properties motivate its use for the neural source audits in Sections 4.2 and 4.3.

4.2 Temporal Generalization Is Overestimated Under Conventional Encoding

The passive benchmark provides operational validation that CAST-FIR can attenuate temporal source mixture by separating modeled contributions from other words. We therefore focus the neural

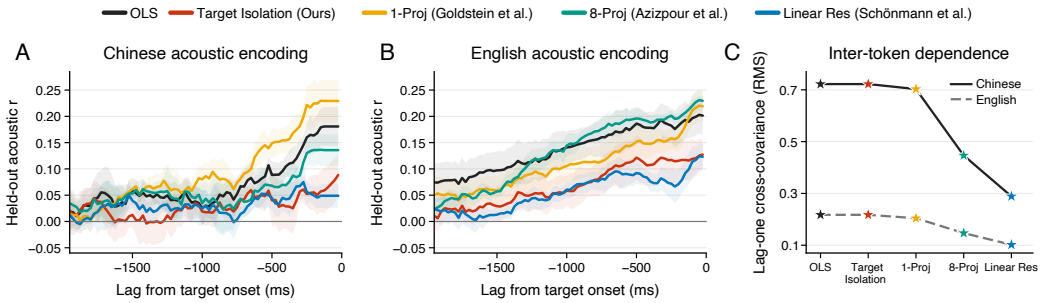


Figure 4: **Target Isolation attenuates passive pre-onset prediction while preserving inter-token dependence.** **A,B**, held-out acoustic encoding under OLS, Target Isolation, and predictor-side controls. **C**, only predictor-side controls suppress representation cross-covariance.

source audit on whether Target Isolation changes the apparent temporal generalization of target-word representations (Goldstein et al., 2022; 2024; 2025). Both neural analyses used 70–200 Hz high gamma, centered 200 ms response windows, and a 25 ms stride. The English analysis used GPT-2 XL last-layer representations and ten contiguous folds, whereas the Chinese analysis used Qwen2.5-1.5B last-layer representations and leave-one-story-out cross-validation. Detailed preprocessing and cross-validation procedures are provided in Appendix C.

Because conventional encoding fits an independent mapping at each lag, pre-onset performance alone cannot establish that the same neural representation is engaged before and after word onset. Azizpour et al. (2026) therefore used temporal generalization to test the stronger pre-activation hypothesis: a mapping trained at τ_{train} is held fixed and evaluated against neural responses at every τ_{test}. If a post-onset word representation is already active before onset, its mapping should generalize into the pre-onset region. Following this logic, we evaluated

$$M(\tau_{\text{train}}, \tau_{\text{test}}) = \text{Corr}(\hat{y}_{\tau_{\text{train}}}, y_{\tau_{\text{test}}}) .$$

Repeating this procedure over all lag pairs yields a train-by-test heat map, where red indicates stronger temporal generalization and green indicates little or none. In the raw matrices (Figure 5**A**, left column), temporal generalization extends broadly along the diagonal. Under conventional OLS, the red regions in the second and fourth quadrants indicate cross-onset temporal generalization, suggesting a shared neural response pattern before and after target onset and potentially supporting next-word prediction in the brain (Appendix B.6). However, temporal generalization collapsed once temporal source mixing was mitigated by isolating the target-word contribution with the cross-fitted CAST-FIR model. After Target Isolation, the cross-onset generalization largely disappeared: using $r > .075$ as a descriptive threshold, the high-generalization area (red) contracted by $98.1\%/100.0\%$ for English/Chinese. The small remaining region is concentrated around onset, indicating generalization within the target-token processing window rather than broad predictive transfer across words.

Temporal source mixing causes conventional target-aligned analyses to substantially overestimate apparent cross-onset temporal generalization in both neural datasets. Conclusions drawn primarily from encoding performance can therefore be markedly misleading: Figure 5**A** does not show that the brain never predicts future input; it shows that conventional encoding performance and temporal generalization are insufficient to establish next-word prediction in the brain. Once contributions from available sources are accounted for, the conventional paradigm provides little source-specific evidence that the target word is predictively represented before onset. Accordingly, we next shift our focus from the aggregate score to the interpretability afforded by the fitted model itself.

4.3 Interpreting the Model Rather Than the Score

As illustrated in Figure 1**A**, encoding performance is a validation metric whose role is to quantify how well model predictions match neural responses. Across the two parallel pipelines—the brain processing a podcast and the encoding model transforming semantic representations—the fitted model, rather than its performance, is the object placed in mechanistic correspondence with the brain. Because the brain is effectively a black box whose neural computations are difficult to observe directly,

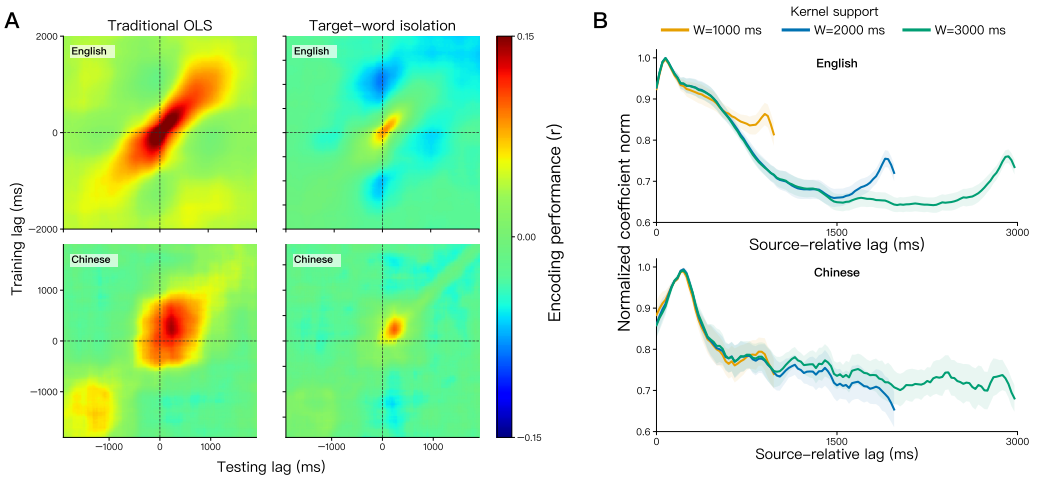


Figure 5: **Target Isolation attenuates cross-onset temporal generalization, while source-relative-lag kernels peak after onset. A**, native and Target Isolated temporal generalization for English and Chinese. **B**, last-layer source-relative-lag kernels across three response supports; ribbons show SEM.

an encoding model serves as a tractable surrogate that approximates its behavior; interrogating this approximation illuminates neural mechanisms and constitutes the central purpose of neural encoding. Performance should therefore be used to establish whether a model fits the data well enough to support interpretation, not to explain neural mechanisms directly; such explanation additionally requires the model itself to be interpretable. This creates a trade-off between predictive accuracy and interpretability: an overly flexible model may be accurate yet opaque, whereas an overly simple model may underfit neural responses, rendering its interpretation unreliable. We therefore turn to the internal structure of CAST-FIR. By explicitly parameterizing source identity and source-relative time, its source-relative-lag kernel reveals when each word contributes to the neural response.

For contextual word predictors, the coefficient-block norm traces how strongly the fitted model maps contextual-semantic information onto neural activity over time from word onset. Across $W = 1000/2000/3000$ ms supports, the participant-averaged profiles rose after onset, peaked at 75 ms in English and 225 ms in Chinese, and then declined (Figure 5**B**). These post-onset peaks broadly align with established cortical speech-processing timescales of approximately 60–200 ms (Brodbeck et al., 2018; Travis et al., 2013). This correspondence motivates the hypothesis that neural sensitivity to semantic information rises after word onset, peaks at approximately 75–225 ms, and then declines.

5 Conclusion

This work shows that pre-onset predictability in conventional target-aligned encoding does not establish target-word provenance. Theory and a strictly causal simulation demonstrate how sequential dependence and response superposition mix contributions from temporally distinct sources, while a passive acoustic benchmark shows that the so-called predictive signal can arise without anticipatory computation. Across English ECoG and Chinese SEEG, Target Isolation substantially attenuated apparent cross-onset temporal generalization. These findings do not rule out neural prediction; they show that aggregate performance cannot identify which word generated predictable activity or when that source began to contribute. Target Isolation provides an auditable route from predictive adequacy to temporal attribution. We therefore encourage future work to use performance to establish predictive adequacy, while grounding mechanistic interpretation in model components that explicitly parameterize the dimensions of interest.

ETHICS STATEMENT

The English-language analyses were secondary analyses of the publicly released dataset (Bhattacharjee et al., 2026; Bhattacharjee, 2025), derived from the Podcast ECoG dataset (Zada et al., 2025a;b). The source studies reported institutional ethics approval and participant informed consent.

The Chinese SEEG experiment was approved by the institutional ethics committee of the recruiting hospital (committee name and approval number withheld for double-blind review). All participants provided written informed consent before participation and could withdraw at any time without affecting their clinical care. The number, type, anatomical targets, implantation, and removal of all electrodes were determined exclusively by the treating clinical team based on clinical needs. The research introduced no additional electrodes, did not prolong clinical monitoring, and did not affect diagnosis or treatment. Research participation was limited to passive story-listening tasks conducted during clinically required monitoring. Data were pseudonymized before analysis, and raw recordings and identifiable clinical or anatomical information will not be publicly released.

REPRODUCIBILITY STATEMENT

The reproducibility materials include arrays, tables, manifests, and validation reports for the main simulation, acoustic, and neural results. We will submit these materials as an anonymized ZIP archive in the supplementary materials, together with a code snapshot, environment specification, fixed random seeds, figure- and table-level analysis manifests, and instructions for accessing the English dataset. If the raw Chinese recordings cannot be released because of clinical privacy constraints, we will provide de-identified derived statistics and complete reproduction scripts to the extent permitted by the ethics approval.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, writing portions of the code, and drafting initial versions of parts of the manuscript. All AI-assisted content was reviewed, revised, and verified by the authors against the primary literature, study records, analysis code, and experimental results. The authors take full responsibility for the final content of this submission.

REFERENCES

Sahel Azizpour, Britta U. Westner, Jakub Szewczyk, Umut Güçlü, and Linda Geerligs. Reassessing prediction in the brain: Pre-onset neural encoding during natural listening does not reflect pre-activation. *eLife*, 2026. doi: 10.7554/eLife.110085.1. Reviewed Preprint, version 1.

Arnab Bhattacharjee. ECoG data of 8 subjects listening to a podcast, 2025. URL `https://doi.org/10.5281/zenodo.15220273`.

Arnab Bhattacharjee, Zaid Zada, Haocheng Wang, Bobbi Aubrey, Werner Doyle, Patricia Dugan, Daniel Friedman, Orrin Devinsky, Adeen Flinker, Peter J. Ramadge, Uri Hasson, Ariel Goldstein, and Samuel A. Nastase. Aligning brains into a shared space improves their alignment with large language models. *Nature Computational Science*, 6:169–178, 2026. doi: 10.1038/s43588-025-00900-y. Published online 18 November 2025.

Christian Brodbeck, L. Elliot Hong, and Jonathan Z. Simon. Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, 28(24):3976–3983.e5, 2018. doi: 10.1016/j.cub.2018.10.042.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022. doi: 10.1038/s42003-022-03036-1.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity—multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pp. 1–14, 2017. doi: 10.18653/v1/S17-2001.

Michael J. Crosse, Giovanni M. Di Liberto, Adam Bednar, and Edmund C. Lalor. The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, 10:604, 2016. doi: 10.3389/fnhum.2016.00604.

Michael J. Crosse, Nathaniel J. Zuk, Giovanni M. Di Liberto, Aaron R. Nidiffer, Sophie Molholm, and Edmund C. Lalor. Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. *Frontiers in Neuroscience*, 15:705621, 2021. doi: 10.3389/fnins.2021.705621.

Benedikt V. Ehinger and Olaf Dimigen. Unfold: An integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *PeerJ*, 7:e7838, 2019. doi: 10.7717/peerj.7838.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. doi: 10.1038/s41593-022-01026-4.

Ariel Goldstein, Avigail Grinstein-Dabush, Mariano Schain, Haocheng Wang, Zhuoqiao Hong, Bobbi Aubrey, Samuel A. Nastase, Zaid Zada, Eric Ham, Amir Feder, Harshvardhan Gazula, Eliav Buchnik, Werner Doyle, Sasha Devore, Patricia Dugan, Roi Reichart, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Orrin Devinsky, Adeen Flinker, and Uri Hasson. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, 15:2768, 2024. doi: 10.1038/s41467-024-46631-y.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A. Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Yossi Matias, Orrin Devinsky, Noam Siegelman, Adeen Flinker, Omer Levy, Roi Reichart, and Uri Hasson. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, 16:10529, 2025. doi: 10.1038/s41467-025-65518-0.

Nuno R. Gonçalves, Robert Whelan, John J. Foxe, and Edmund C. Lalor. Towards obtaining spatiotemporally precise responses to continuous sensory stimuli in humans: A general linear modeling approach to EEG. *NeuroImage*, 97:196–205, 2014. doi: 10.1016/j.neuroimage.2014.04.012.

Liberty S. Hamilton and Alexander G. Huth. The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582, 2020. doi: 10.1080/23273798.2018.1499946.

Christopher R. Holdgraf, Jochem W. Rieger, Cristiano Micheli, Stephanie Martin, Robert T. Knight, and Frédéric E. Theunissen. Encoding and decoding models in cognitive electrophysiology. *Frontiers in Systems Neuroscience*, 11:61, 2017. doi: 10.3389/fnsys.2017.00061.

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile the human cerebral cortex. *Nature*, 532:453–458, 2016. doi: 10.1038/nature17637.

Shailee Jain and Alexander G. Huth. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6628–6637, 2018.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4563–4568, 2020. doi: 10.18653/v1/2020.emnlp-main.369.

Edmund C. Lalor and John J. Foxe. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, 31(1):189–193, 2010. doi: 10.1111/j.1460-9568.2009.07055.x.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F. Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014. doi: 10.1126/science.1245994.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011. doi: 10.1016/j.neuroimage.2010.07.073.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Jona Sassenhagen. How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression. *Language, Cognition and Neuroscience*, 34(4):474–490, 2019. doi: 10.1080/23273798.2018.1502458.

Inés Schönmann, Jakub Szewczyk, Floris P. de Lange, and Micha Heilbron. Stimulus dependencies—rather than next-word prediction—can explain pre-onset brain encoding in naturalistic listening designs. *eLife*, 14:RP106543, 2026. doi: 10.7554/eLife.106543.3.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118.

Cory Shain and William Schuler. Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition*, 215:104735, 2021. doi: 10.1016/j.cognition.2021.104735.

Nathaniel J. Smith and Marta Kutas. Regression-based estimation of ERP waveforms: I. the rERP framework. *Psychophysiology*, 52(2):157–168, 2015a. doi: 10.1111/psyp.12317.

Nathaniel J. Smith and Marta Kutas. Regression-based estimation of ERP waveforms: Ii. nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, 52(2):169–181, 2015b. doi: 10.1111/psyp.12320.

Frédéric E. Theunissen, Stephen V. David, Nima C. Singh, Anne Hsu, William E. Vinje, and Jack L. Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, 12(3):289–316, 2001. doi: 10.1080/net.12.3.289.316.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32, pp. 14954–14964, 2019.

Mariya Toneva, Tom M. Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757, 2022. doi: 10.1038/s43588-022-00354-6.

Katherine E. Travis, Matthew K. Leonard, Alexander M. Chan, Christina Torres, Marisa L. Sizemore, Zhe Qu, Emad Eskandar, Anders M. Dale, Jeffrey L. Elman, Sydney S. Cash, and Eric Halgren. Independence of early speech processing from word meaning. *Cerebral Cortex*, 23(10):2370–2379, 2013. doi: 10.1093/cercor/bhs228.

Marcel A. J. van Gerven. A primer on encoding models in sensory neuroscience. *Journal of Mathematical Psychology*, 76:172–183, 2017. doi: 10.1016/j.jmp.2016.06.009.

Marty G. Woldorff. Distortion of ERP averages due to overlap from temporally adjacent ERPs: Analysis and correction. *Psychophysiology*, 30(1):98–119, 1993. doi: 10.1111/j.1469-8986.1993.tb03209.x.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3687–3692, 2019. doi: 10.18653/v1/D19-1382.

Zaid Zada, Samuel A. Nastase, Bobbi Aubrey, Itamar Jalon, Ariel Goldstein, Sebastian Michelmann, Haocheng Wang, Liat Hasenfratz, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Sasha Devore, Orrin Devinsky, Adeen Flinker, and Uri Hasson. The "podcast" ECoG dataset for modeling neural activity during natural language comprehension. *Scientific Data*, 12:1135, 2025a. doi: 10.1038/s41597-025-05462-2.

Zaid Zada et al. The "podcast" ECoG dataset, 2025b. URL `https://doi.org/10.18112/openneuro.ds005574.v1.0.0`.

A Assumptions and Full Theory Derivations

A.1 Notation

Table 1: Main notation used in the theory and source-time model.

Symbol	Definition
i, j, k	target word, source word, and backward word offset ($j = i - k$)
e, q	neural contact and absolute-time query index
O_j	source-word onset time
t, τ, a	absolute time, target-relative lag $t - O_i$, and source-relative lag $t - O_j$
$\Delta_{i,k}$	onset gap $O_i - O_{i-k}$
$x_j \in \mathbb{R}^p$	centered source-word representation
$S_e(t)$	instantaneous absolute-time neural signal
$y_{q,e}$	window-averaged neural response centered at absolute query time t_q
$y_{i,e}(\tau)$	the same response indexed relative to target word i, with $y_{i,e}(\tau) = y_{q,e}$ when $t_q = O_i + \tau$
$h_e(a), \bar{h}_{e,W_y}(a)$	source-relative-lag kernel and its window-averaged version
$u_e(t), \epsilon_e(t)$	unmodeled systematic component and residual error
C_x, Γ_k	current covariance and current–history cross-covariance
$B_m, z_{q,m}$	source-relative-lag bin and representation aggregated within that bin
$\theta_{e,m}, \Theta_e$	contact-specific age-bin coefficient and stacked coefficient vector

Unless otherwise stated, representations and responses are population-centered and the regression includes an intercept. Sample analogues use centering parameters fitted on the training split.

A.2 Modeling Assumptions and Temporal Windowing

Equation 1 is a minimal mechanism sufficient to generate temporal source mixing. It assumes that (i) source contributions are approximately additive at the temporal scale of analysis; (ii) for a given contact, the average representation-to-response mapping is indexed by source-relative lag and is approximately stationary across training and test segments; (iii) population second moments exist; and (iv) the effective temporal support induced by neural preprocessing and response averaging is known or explicitly bounded. The model does not claim to exhaust neural nonlinearities, state dependence, inter-regional interactions, attention, or unobserved acoustic and phonetic factors.

CAST-FIR remains a forward encoding model in the standard sense (Naselaris et al., 2011; van Gerven, 2017; Holdgraf et al., 2017), and its use of contextual feature spaces follows the neural language-encoding literature (Huth et al., 2016; Jain & Huth, 2018; Toneva & Wehbe, 2019; Caucheteux & King, 2022). Additive lagged-response models have a long history in receptive-field estimation and continuous-stimulus encoding (Theunissen et al., 2001; Lalor & Foxe, 2010; Gonçalves et al., 2014; Crosse et al., 2016; 2021), while regression-based deconvolution methods explicitly address responses that overlap across adjacent events (Woldorff, 1993; Smith & Kutas, 2015a;b; Ehinger & Dimigen, 2019; Shain & Schuler, 2021; Sassenhagen, 2019). CAST-FIR uses this lineage for a different inferential target: it indexes each lagged contribution by source identity and source-relative time so that held-out predictions can be audited by provenance.

A centered response window of width W_y defines

$$y_{q,e} = \frac{1}{W_y} \int_{-W_y/2}^{W_y/2} S_e(t_q + u)\, du, \qquad y_{i,e}(\tau) = y_{q,e}\big|_{t_q = O_i + \tau}, \tag{8}$$

with windowed kernel

$$\bar{h}_{e,W_y}(a) = \frac{1}{W_y} \int_{-W_y/2}^{W_y/2} h_e(a + u)\, du. \tag{9}$$

Even if $h_e(a) = 0$ for $a < 0$, the windowed kernel can be nonzero for $-W_y/2 < a < 0$, because centered window averaging mixes post-onset activity into nominally pre-onset lags.

A.3 Exact Conventional Encoding Mixture

Retaining the preceding K sources, Equation 1 gives

$$y_{i,e}(\tau) = x_i^\top \bar{h}_{e,W_y}(\tau) + \sum_{k=1}^{K} x_{i-k}^\top \bar{h}_{e,W_y}(\tau + \Delta_{i,k}) + \eta_{i,e}^{(K)}(\tau), \tag{10}$$

where $\eta_{i,e}^{(K)}(\tau)$ collects sources outside the displayed history, subsequent sources available at positive lags, unmodeled systematic activity, and residual noise.

Define

$$c_e(\tau) = \mathrm{Cov}(x_i, y_{i,e}(\tau)), \qquad C_x = \mathrm{Cov}(x_i, x_i). \tag{11}$$

The minimum-norm population least-squares coefficient is

$$\beta_e^*(\tau) = C_x^\dagger c_e(\tau), \tag{12}$$

where † denotes the Moore–Penrose pseudoinverse, which reduces to the ordinary inverse when C_x is full rank. Substituting Equation 10 yields the exact identity

$$c_e(\tau) = C_x \bar{h}_{e,W_y}(\tau) + H_e^{(K)}(\tau) + b_{\eta,e}^{(K)}(\tau), \tag{13}$$

where

$$H_e^{(K)}(\tau) = \sum_{k=1}^{K} \mathbb{E}\big[x_i x_{i-k}^\top \bar{h}_{e,W_y}(\tau + \Delta_{i,k})\big], \tag{14}$$

$$b_{\eta,e}^{(K)}(\tau) = \mathrm{Cov}\Big(x_i, \eta_{i,e}^{(K)}(\tau)\Big). \tag{15}$$

The term $H_e^{(K)}(\tau)$ is the exact response-lag-weighted history cross-moment. It jointly depends on the representation pair (x_i, x_{i-k}), the onset gap $\Delta_{i,k}$, and the response kernel evaluated at the corresponding source-relative lag. Equation 14 therefore does not require onset gaps to be independent of the associated word pairs.

If the conditional cross-moment

$$\mathbb{E}\big[x_i x_{i-k}^\top \mid \Delta_{i,k}\big]$$

varies weakly with the onset gap, the representation dependence and source-relative-lag dependence can be approximately separated. Specifically,

$$\begin{aligned} H_e^{(K)}(\tau) &\approx \sum_{k=1}^{K} \Gamma_k g_{e,k}(\tau), \\ \Gamma_k &= \mathrm{Cov}(x_i, x_{i-k}), \\ g_{e,k}(\tau) &= \mathbb{E}_{\Delta_{i,k}}\left[\bar{h}_{e,W_y}(\tau + \Delta_{i,k})\right]. \end{aligned} \tag{16}$$

Here, Γ_k captures sequential dependence across representations, whereas $g_{e,k}(\tau)$ captures the neural response at the source-relative lags associated with the k-th preceding source. This factorization exposes the interaction underlying temporal source mixing:

$$\underbrace{\Gamma_k}_{\text{sequential dependence}} \quad \underbrace{g_{e,k}(\tau)}_{\text{source-lag response}} \quad \longrightarrow \quad \underbrace{\Gamma_k g_{e,k}(\tau)}_{\text{history contribution}} .$$

The factorization is used for interpretation and is not required for the exact OLS identity. When the separability condition does not hold, the appropriate history contribution remains Equation 14.

Combining Equations 12 and 13 gives

$$\beta_e^*(\tau) = C_x^\dagger C_x \bar{h}_{e,W_y}(\tau) + C_x^\dagger H_e^{(K)}(\tau) + C_x^\dagger b_{\eta,e}^{(K)}(\tau). \tag{17}$$

Under the separability approximation and the residual-orthogonality condition

$$b_{\eta,e}^{(K)}(\tau) \approx 0,$$

this becomes

$$\beta_e^*(\tau) \approx C_x^\dagger C_x \bar{h}_{e,W_y}(\tau) + C_x^\dagger \sum_{k=1}^{K} \Gamma_k g_{e,k}(\tau). \tag{18}$$

When C_x is full rank, $C_x^\dagger = C_x^{-1}$ and $C_x^{-1} C_x = I$, so Equation 18 reduces to the form used in Equation 3 of the main text.

When C_x is rank deficient, however, the target component recovered by population OLS is

$$C_x^\dagger C_x \bar{h}_{e,W_y}(\tau), \tag{19}$$

rather than the full kernel $\bar{h}_{e,W_y}(\tau)$. Because C_x is symmetric and positive semidefinite,

$$P_x = C_x^\dagger C_x$$

is the orthogonal projector onto the column space of C_x, namely the subspace along which the representations exhibit nonzero population variation. To make this explicit, decompose the kernel as

$$\bar{h}_{e,W_y}(\tau) = \bar{h}^{\parallel}_{e,W_y}(\tau) + \bar{h}^{\perp}_{e,W_y}(\tau), \tag{20}$$

where

$$\bar{h}^{\parallel}_{e,W_y}(\tau) = C_x^\dagger C_x \bar{h}_{e,W_y}(\tau), \tag{21}$$

$$\bar{h}^{\perp}_{e,W_y}(\tau) = \left(I - C_x^\dagger C_x\right) \bar{h}_{e,W_y}(\tau). \tag{22}$$

The component $\bar{h}^{\parallel}_{e,W_y}(\tau)$ lies in the column space of C_x and is identifiable from variation in x_i. By contrast, $\bar{h}^{\perp}_{e,W_y}(\tau)$ lies in the null space of C_x, so

$$C_x \bar{h}^{\perp}_{e,W_y}(\tau) = 0. \tag{23}$$

Because the representations are centered,

$$\mathbb{E}\left[\left(x_i^\top \bar{h}^{\perp}_{e,W_y}(\tau)\right)^2\right] = \bar{h}^{\perp}_{e,W_y}(\tau)^\top \mathbb{E}[x_i x_i^\top] \bar{h}^{\perp}_{e,W_y}(\tau) \tag{24}$$

$$= \bar{h}^{\perp}_{e,W_y}(\tau)^\top C_x \bar{h}^{\perp}_{e,W_y}(\tau) \tag{25}$$

$$= 0. \tag{26}$$

It follows that

$$x_i^\top \bar{h}^{\perp}_{e,W_y}(\tau) = 0 \qquad \text{almost surely.} \tag{27}$$

Consequently,

$$x_i^\top \bar{h}_{e,W_y}(\tau) = x_i^\top \bar{h}^{\parallel}_{e,W_y}(\tau) \qquad \text{almost surely.} \tag{28}$$

Thus, the null-space component does not affect the encoding prediction and cannot be identified from the observed variation in x_i. Equivalently, we may define $\bar{h}_{e,W_y}(\tau)$ as its identifiable projection without changing any model prediction, recovering the notation used in the main text. When C_x is full rank, $C_x^\dagger C_x = I$, and the identifiable projection is the full kernel.

A.4 Exact Lag–r Decomposition

The finite-sample held-out metric is

$$\widehat{r}_e(\tau) = \mathrm{Corr}_{i \in \mathcal{I}_{\text{test}}} \left(x_i^\top \widehat{\beta}_e(\tau), y_{i,e}(\tau) \right). \tag{29}$$

In the population setting, with identically distributed training and test data, the squared correlation between the linear projection and the response is

$$r_e^2(\tau) = \frac{c_e(\tau)^\top C_x^\dagger c_e(\tau)}{\mathrm{Var}(y_{i,e}(\tau))}. \tag{30}$$

Let $A_e(\tau) = C_x \bar{h}_{e,W_y}(\tau)$ and $B_e(\tau) = H_e^{(K)}(\tau) + b_{\eta,e}^{(K)}(\tau)$. The numerator expands as

$$A_e^\top C_x^\dagger A_e + 2A_e^\top C_x^\dagger B_e + B_e^\top C_x^\dagger B_e. \tag{31}$$

These terms capture accessible target signal, target–history interference, and accessible history signal. The interference term can have either sign, so subtracting estimated history need not reduce r monotonically at every lag. When the target source has no causal contribution, $A_e(\tau) = 0$, yet $B_e^\top C_x^\dagger B_e$ can remain positive.

A.5 Exact No-History Conditions

Let v_i be the predictor produced by a control and let s_{i-k} denote the source feature that drives the k-th historical response. At a fixed contact and lag, the exact history contribution to its cross-covariance with the response is

$$H_{e,v}^{\text{hist}}(\tau) = \sum_{k=1}^{K} \mathbb{E}\big[(v_i - \mu_v) s_{i-k}^\top \bar{h}_{e,W_y}(\tau + \Delta_{i,k})\big]. \tag{32}$$

The exact no-history condition is $H_{e,v}^{\text{hist}}(\tau) = 0$. It may hold through a zero cross-moment, a null response direction, or cancellation across historical offsets; thus zero marginal covariance is not necessary at one particular lag. Conversely, if response age and word-pair statistics are dependent, $\mathrm{Cov}(v_i, s_{i-k}) = 0$ need not be sufficient because age-specific conditional cross-covariances can cancel marginally and reappear after weighting by the kernel.

A stronger condition that is sufficient independently of the unknown kernel is

$$\mathbb{E}\big[(v_i - \mu_v) s_{i-k}^\top \mid \tau + \Delta_{i,k} = a\big] = 0 \quad \text{for every relevant } a \text{ and } k. \tag{33}$$

Under the additional separability assumption used in Equation 16, this reduces to the simpler sufficient condition $\mathrm{Cov}(v_i, s_{i-k}) = 0$ for every response-relevant k.

A.6 Algebraic Scope of Predictor-Side Controls

One-word geometric projection. For each word, Goldstein et al. (2022) project the target-word representation onto the orthogonal complement of the immediately preceding representation:

$$v_i = x_i - \frac{x_i^\top x_{i-1}}{\|x_{i-1}\|_2^2} x_{i-1}. \tag{34}$$

This construction ensures that the transformed target representation has zero inner product with the previous-word representation:

$$v_i^\top x_{i-1} = 0.$$

However, sequential dependence is characterized by $\mathrm{Cov}(v_i, x_{i-1})$, or more exactly by Equation 32. A zero scalar inner product for each pair constrains a trace-like within-pair quantity but does not set every entry of this cross-covariance matrix to zero, nor does it address $k \geq 2$.

Eight-word geometric projection. Recognizing the limitation of restricting geometric projection to the previous-word representation, Azizpour et al. (2026) extend the procedure to the preceding eight words. Let

$$G_i = [x_{i-1} \quad \cdots \quad x_{i-8}] \in \mathbb{R}^{p\times 8}.$$

The transformed representation is

$$v_i = \left(I - G_i G_i^\dagger\right) x_i, \tag{35}$$

where $G_i G_i^\dagger$ is the orthogonal projector onto the column space of G_i. Similarly, because G_i, and hence the projection subspace, varies across occurrences, this within-occurrence orthogonality does not generally imply $\mathrm{Cov}(v_i, x_{i-k}) = 0$ across the dataset. The procedure also leaves sources outside its fixed eight-word support uncontrolled and changes the contextual representation being evaluated.

One-word linear residualization. Schönmann et al. (2026) instead estimate a linear mapping from the immediately preceding representation to the target representation. Let all variables be centered. At the population level, the corresponding linear residual is

$$v_i = x_i - \mathrm{Cov}(x_i, x_{i-1})\,\mathrm{Cov}(x_{i-1}, x_{i-1})^\dagger x_{i-1}. \tag{36}$$

This construction satisfies

$$\mathrm{Cov}(v_i, x_{i-1}) = 0$$

at the population level, while sample OLS provides the analogous orthogonality on the centered training set. It therefore targets across-sample dependence more directly than geometric projection. Nevertheless, it does not constrain histories at $k \geq 2$ and does not mechanically guarantee zero covariance after the estimated residualization operator is applied to held-out data.

Joint linear residualization over an expanding history. A natural extension is to residualize the target representation jointly against the preceding K representations. Define

$$Z_i^{(K)} = \left[x_{i-1}^\top \quad \cdots \quad x_{i-K}^\top\right]^\top, \tag{37}$$

and let

$$\begin{aligned} v_i^{(K)} &= x_i - C_{xZ}^{(K)} \left(C_Z^{(K)}\right)^\dagger Z_i^{(K)}, \\ C_{xZ}^{(K)} &= \mathrm{Cov}\left(x_i, Z_i^{(K)}\right), \\ C_Z^{(K)} &= \mathrm{Cov}\left(Z_i^{(K)}, Z_i^{(K)}\right). \end{aligned} \tag{38}$$

The population residual satisfies

$$\mathrm{Cov}\left(v_i^{(K)}, Z_i^{(K)}\right) = 0.$$

Its covariance is

$$C_v^{(K)} = C_x - C_{xZ}^{(K)} \left(C_Z^{(K)}\right)^\dagger C_{Zx}^{(K)}. \tag{39}$$

Because the history spaces are nested, orthogonal projection gives

$$C_v^{(K+1)} \preceq C_v^{(K)}, \qquad \mathbb{E}\left[\left\|v_i^{(K+1)}\right\|_2^2\right] \leq \mathbb{E}\left[\left\|v_i^{(K)}\right\|_2^2\right]. \tag{40}$$

As K increases, the residual converges to

$$x_i - \mathrm{Proj}_{\mathcal{H}_\infty}(x_i),$$

the component of x_i that cannot be linearly predicted from its full representation history, where $\mathcal{H}_\infty$ denotes the closed linear span of that history. This limit is zero if and only if

$$x_i \in \mathcal{H}_\infty.$$

For a centered finite training set, let $X \in \mathbb{R}^{N\times p}$ contain the target representations and let Z_K denote the stacked history design. The sample residual is

$$V_K = (I - P_{Z_K})\, X, \qquad P_{Z_K} = Z_K Z_K^{\dagger}. \tag{41}$$

It collapses exactly when

$$\mathrm{col}(X) \subseteq \mathrm{col}(Z_K),$$

which can occur when a high-dimensional history design spans the centered sample space. Such training-set collapse can reflect interpolation and need not generalize to held-out data. Even without complete collapse, increasing K progressively removes components of the original contextual representation that are linearly predictable from prior context. The resulting predictor therefore represents a history-conditioned transformation rather than the original semantic representation.

A.7 Full Source-Time Design and Estimation

Let

$$B_m = [m\delta, (m+1)\delta), \qquad m = 0, \ldots, M-1,$$

partition the nonnegative source-relative-lag support $[0, L)$, where $L = M\delta$. At query time t_q, define

$$z_{q,m} = \sum_j x_j\, \mathbf{1}[t_q - O_j \in B_m] \in \mathbb{R}^p. \tag{42}$$

Because $B_m \subseteq [0, L)$, the condition $t_q - O_j \in B_m$ simultaneously assigns source word j to its current source-relative-lag bin and enforces onset availability: it implies $O_j \le t_q$. No separate feature-availability time is therefore required.

Let

$$Z_q = \begin{bmatrix} z_{q,0}^\top & \cdots & z_{q,M-1}^\top \end{bmatrix} \in \mathbb{R}^{1\times Mp}, \qquad \Theta_e = \begin{bmatrix} \theta_{e,0}^\top & \cdots & \theta_{e,M-1}^\top \end{bmatrix}^\top. \tag{43}$$

Then

$$\widehat{y}_{q,e} = \widehat{\alpha}_e + Z_q \widehat{\Theta}_e = \widehat{\alpha}_e + \sum_{m=0}^{M-1} z_{q,m}^\top \widehat{\theta}_{e,m}. \tag{44}$$

For an outer-training query set $\mathcal{Q}_{-f}$, ridge estimation solves

$$(\widehat{\alpha}_e^{(-f)}, \widehat{\Theta}_e^{(-f)}) = \arg\min_{\alpha,\Theta} \sum_{q\in\mathcal{Q}_{-f}} [y_{q,e} - \alpha - Z_q\Theta]^2 + \lambda\|\Theta\|_2^2. \tag{45}$$

All source-relative-lag coefficients are therefore estimated jointly within each outer-training split. All data-driven centering, scaling, PCA, and regularization choices for this branch are fitted within the corresponding training split and applied unchanged to its held-out queries. The response target is the window-averaged quantity $y_{q,e}$ defined in Equation 8, matching the target used by conventional encoding.

For $a \in [0, L)$, let

$$m(a) = \left\lfloor \frac{a}{\delta} \right\rfloor \tag{46}$$

denote its source-relative-lag-bin index. The contribution of source word j to a held-out query is

$$\widehat{c}_{q,e,j}^{(-f)} = \mathbf{1}[0 \le t_q - O_j < L]\, x_j^\top \widehat{\theta}_{e,m(t_q-O_j)}^{(-f)}. \tag{47}$$

Hence the held-out prediction decomposes as

$$\widehat{y}_{q,e}^{(-f)} = \widehat{\alpha}_e^{(-f)} + \sum_j \widehat{c}_{q,e,j}^{(-f)}, \tag{48}$$

and can be aggregated by source identity, source set, or source-relative lag without refitting the model.

For an anchor word i, define the available historical and non-target source sets at query time $t_q = O_i + \tau$ as

$$\mathcal{H}_i(t_q) = \{j < i : 0 \le t_q - O_j < L\}, \tag{49}$$

$$\mathcal{N}_i(t_q) = \{j \neq i : 0 \le t_q - O_j < L\}. \tag{50}$$

Their cross-fitted modeled contributions are

$$\widehat{y}_{i,e}^{\,\text{history},(-f)}(\tau) = \sum_{j \in \mathcal{H}_i(t_q)} \widehat{c}_{q,e,j}^{\,(-f)}, \tag{51}$$

$$\widehat{y}_{i,e}^{\,\text{non-target},(-f)}(\tau) = \sum_{j \in \mathcal{N}_i(t_q)} \widehat{c}_{q,e,j}^{\,(-f)}. \tag{52}$$

The corresponding adjusted responses are

$$\widetilde{y}_{i,e}^{\,\text{history},(-f)}(\tau) = y_{i,e}(\tau) - \widehat{y}_{i,e}^{\,\text{history},(-f)}(\tau), \tag{53}$$

$$\widetilde{y}_{i,e}^{\,\text{target},(-f)}(\tau) = y_{i,e}(\tau) - \widehat{y}_{i,e}^{\,\text{non-target},(-f)}(\tau). \tag{54}$$

The first removes only sources preceding the anchor in token order. The second implements Equation 6: it removes every non-target source that has become available by t_q, including words with $j > i$ once their onsets have passed. Because token order follows onset order, the two source sets coincide before target onset. At positive target-relative lags, their difference is exactly the set of post-target words already available on the absolute timeline.

A.8 Dataset-Specific Cross-Fitting for Target Isolation

To prevent direct self-fitting, all responses used by Target Isolation are constructed out of sample under dataset-specific splits. For English, the source-time model uses ten purged physical-time folds; the resulting out-of-fold adjusted responses are evaluated with the author-compatible tenfold word-index OLS/TG pipeline. For Chinese, source-time fitting and downstream evaluation use leave-one-story-out folds. Preprocessing and hyperparameter selection for each source-time model are fitted only on its corresponding training fold.

B Complete Simulation Setup and Controlled Factors

B.1 Simulation Generator and Evaluation

The simulation is defined on a 25 ms absolute-time grid. Unless otherwise stated, each condition contains 12 stories with 180 tokens per story. Word onsets are generated sequentially from onset gaps supported on $200, 225, \ldots, 500$ ms with weights $1, 2, 3, 5, 7, 9, 10, 10, 9, 7, 5, 3, 1$. The same onset-gap realization is reused across paired conditions so that changes in the resulting lag–r curves can be attributed to the manipulated factor rather than to different stimulus timing. For each story, six-dimensional source representations are generated as

$$x_i = \rho x_{i-1} + \sqrt{1 - \rho^2}\, \xi_i, \tag{55}$$

where ξ_i is an independent Gaussian innovation and ρ controls sequential dependence between adjacent word representations. Thus, $\rho = 0$ produces independent representations, whereas larger ρ induces stronger cross-word dependence. Because the realized finite-sample dependence need not equal the requested ρ exactly, we also report the empirical adjacent-word correlation $r(x_i, x_{i-1})$.

Each source representation is mapped to the simulated response through shared linear weights sampled elementwise from $\mathcal{N}(0, 1/6)$. The response elicited by a source word depends only on its nonnegative source-relative lag $a = t - O_i$, through the one-sided kernel

$$k(a) \propto \left(1 - e^{-a/180}\right) e^{-a/550}, \qquad a \ge 0, \tag{56}$$

with $k(a) = 0$ for $a < 0$. Unless otherwise stated, the kernel has 3000 ms support and is normalized as specified for the corresponding controlled sweep. This one-sided support guarantees that no word contributes to the simulated response before its own onset. Independent observation noise is added at each time point from $\mathcal{N}(0, 1.65^2)$.

All paired conditions use factorized random streams and therefore share the same onset gaps, Gaussian innovations, linear weights, and noise realization. Only the variable under study is changed. Conventional target-word encoding is then applied exactly as in the main experiment: at each target-relative lag, the target representation x_i is used to fit a separate OLS model to the simulated response. Unless otherwise stated, lags are scanned from -4000 to 4000 ms in 25 ms increments and evaluated with leave-one-story-out cross-validation. All controlled experiments use seed 20260721.

B.2 Factor Settings for the Mechanistic Isolation Experiment

Figure 2**A** and Section 3 isolate the two components of temporal source mixture by independently manipulating sequential dependence and neural-response superposition. Table 2 lists the exact settings used for the four curves. Sequential dependence is removed by setting $\rho = 0$ and introduced by setting $\rho = .72$. Neural-response superposition is suppressed by truncating each source response at the onset of the next word, or retained by allowing the full 3000 ms response to persist.

Table 2: **Factor settings for the mechanistic isolation experiment in Figure 2A.** All four conditions use the same onset gaps, innovations, linear weights, noise realization, and random seed.

Sequential dependence	ρ	Response superposition	Response handling
Absent	0	Suppressed	Truncate at next onset
Absent	0	Retained	Full 3000 ms response
Present	.72	Suppressed	Truncate at next onset
Present	.72	Retained	Full 3000 ms response

In the superposition-suppressed condition, truncation prevents responses from multiple source words from adding within the same absolute-time bin. It does not remove the preceding source's own response from pre-target time. The manipulation therefore isolates the additional accumulation produced by neural-response superposition.

B.3 Neural Response Duration Under Two Normalizations

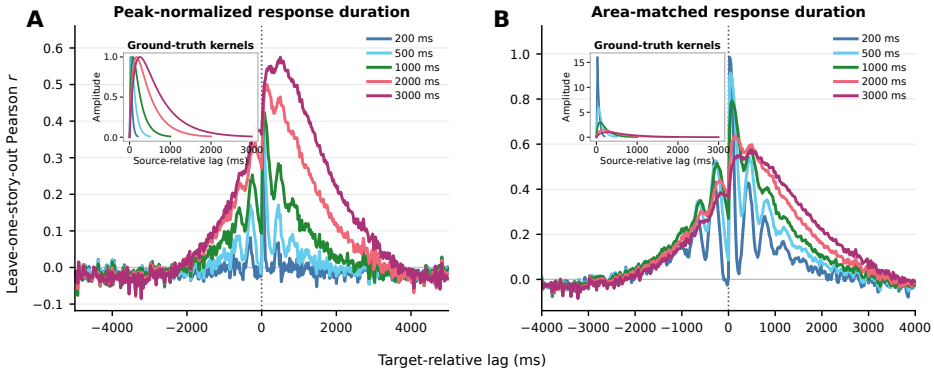


Figure 6: **Response-duration sweeps under two kernel normalizations. A**, Peak-normalized kernels: each kernel has peak amplitude 1, so extending the response increases both source overlap and total kernel area. **B**, Area-matched kernels: integrated kernel mass is held fixed, so short responses have higher local amplitude while long responses spread the same mass across time. Insets show the corresponding ground-truth kernels.

Response duration controls how long a source word remains present in the measured signal. We therefore evaluated two duration sweeps under the default correlated-plus-superposition condition ($\rho = .72$, full response overlap, and no anticipation). Both sweeps used response supports of 200, 500, 1000, 2000, and 3000 ms and kept the onset gaps, source representations, response weights, and noise realization paired across durations. The difference is the normalization applied to the

ground-truth kernel: one hypothesis keeps the local response peak fixed, whereas the other keeps the integrated kernel mass fixed. The peak-normalized audit was scanned over ± 5000 ms for display; all extrema reported here lie within the ± 4000 ms range used by the other sweeps.

Peak-normalized duration. In the first sweep, each duration-specific rise–decay kernel was normalized to have peak amplitude 1 (Figure 6**A**). The rise and decay constants scale with the selected support, so the family preserves the same relative shape while changing its absolute duration. This holds the local response maximum fixed while allowing the integrated response area to increase with duration. Under this assumption, longer responses produce progressively broader and stronger lag–r curves. The maximum negative-lag correlation increases from .081 at 200 ms to .386 at 3000 ms, while the maximum nonnegative-lag correlation increases from .366 to .575 (Table 3). This pattern has two causes. First, longer kernels keep each source active for more target-relative lags, so preceding words can be predicted from the current word across a wider pre-onset interval. Second, because the peak is held fixed, increasing duration also increases total signal energy; the sweep therefore combines greater temporal superposition with a larger cumulative response.

Area-matched duration. In the second sweep, the duration-specific kernels were rescaled to have the same integrated area, 864.27 ms (Figure 6**B**). This fixed-area, or fixed-kernel-mass, assumption holds the total response mass fixed while changing how concentrated that response is in time. Under this assumption, response duration still changes the shape of the lag–r curve, but the peak magnitude is no longer monotonic. The maximum negative-lag correlation is .435, .527, .515, .442, and .386 for 200, 500, 1000, 2000, and 3000 ms supports, respectively. The maximum nonnegative-lag correlation decreases from .985 to .575 as the same integrated response is spread over longer time.

This sweep separates two effects that are confounded in the peak-normalized setting. Short kernels concentrate the fixed response mass into a narrow interval, producing high local SNR and sharp target-relative peaks. Intermediate kernels create enough overlap for preceding-source contributions to appear at negative lags while retaining relatively high local amplitude. Long kernels merge many source-specific responses into a smoother curve, but their lower instantaneous amplitude reduces the peak correlation. Thus, response duration should be interpreted primarily as a determinant of curve width, smoothness, and source accumulation, not as a monotonic determinant of encoding magnitude.

Table 3: **Response-duration summaries under two kernel normalizations.** Values are maximum held-out Pearson r within the negative-lag and nonnegative-lag regions. Peak-normalized kernels keep the local maximum fixed at 1; area-matched kernels keep the integrated response mass fixed.

Duration	Peak-normalized			Area-matched		
(ms)	Kernel peak	Neg. peak r	Nonneg. peak r	Kernel peak	Neg. peak r	Nonneg. peak r
200	1.00	.081	.366	16.09	.435	.985
500	1.00	.170	.383	6.06	.527	.916
1000	1.00	.253	.423	3.00	.515	.792
2000	1.00	.341	.500	1.50	.442	.637
3000	1.00	.386	.575	1.00	.386	.575

B.4 Signal-to-Noise Ratio

Table 4: **SNR sweep summaries.** SNR is the pooled RMS of the noiseless response divided by the pooled RMS of observation noise.

Target SNR	Noise RMS	Neg. peak lag	Neg. peak r	Nonneg. peak r
0.25	5.354	−175 ms	.142	.227
0.50	2.677	−175 ms	.274	.414
1.00	1.338	−175 ms	.432	.641
2.00	0.669	−25 ms	.556	.810

We next varied measurement SNR while keeping the underlying source process fixed. The noiseless response, onset sequence, source representations, and response weights were identical across con-

ditions. The noise direction was also shared; only its amplitude was scaled. SNR is defined as the pooled RMS of the noiseless past response divided by the pooled RMS of observation noise.

Increasing SNR raises the entire lag–r curve without changing the causal structure of the generator (Figure 7). Target SNRs of 0.25, 0.5, 1, and 2 produced maximum negative-lag correlations of $.142$, $.274$, $.432$, and $.556$, respectively, and maximum nonnegative-lag correlations of $.227$, $.414$, $.641$, and $.810$ (Table 4). The negative-lag peak remains near the same part of the curve for the lower three SNRs, while the highest-SNR condition exposes a near-onset maximum at -25 ms.

This manipulation shows that SNR changes detectability and effect size, not provenance. When noise is large, the covariance-mediated contribution from already-onset sources is partially obscured. As noise is reduced, the same source mixture becomes easier for the target-word encoder to predict, so both negative- and positive-lag correlations increase. Because the manipulated variable is only observation noise, the larger pre-onset score at high SNR cannot be interpreted as stronger anticipation; it is the same strictly causal source process observed with less measurement error.

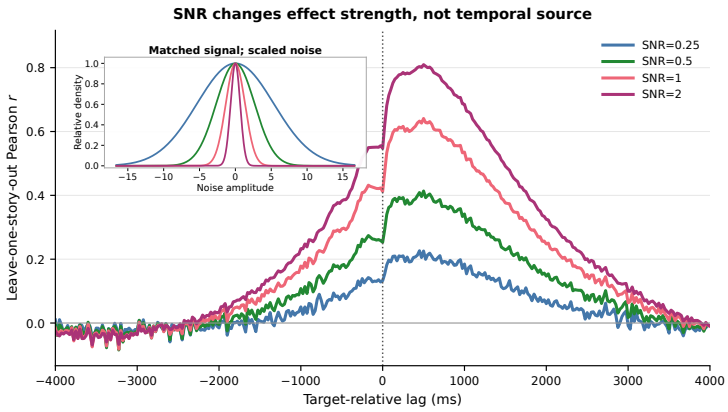


Figure 7: **SNR changes effect strength, not temporal source provenance.** The noiseless response and source sequence are fixed across conditions, and only the amplitude of a shared noise realization is changed. Higher SNR makes the source-mixture structure more visible throughout the lag–r curve.

B.5 Temporal-Axis Precision and Alignment

Additionally, we tested whether imprecision in the event–response time axis could itself explain the lag–r profile. We distinguished random timing imprecision from systematic alignment bias. In both analyses, the conventional target anchor remained the original lexical onset, while the response elicited by each source word was shifted by a per-token latency offset rounded to the 25 ms grid.

Zero-mean latency jitter. When offsets were sampled from $N(0, \sigma^2)$, increasing σ from 0 to 100 ms had little effect on the overall curve (Figure 8**A**). The pre-onset mean remained approximately $.073$ across all four conditions, the maximum negative-lag correlation remained within $.386$–$.418$, and the nonnegative peak declined only modestly from $.575$ to $.562$. The location of the maximum negative-lag point moved from -175 ms to -25 ms for the larger jitters, but the curves are nearly superposed, so this peak location should not be over-interpreted.

Random zero-mean jitter primarily blurs the source-relative-lag assignment of individual events. Because the response kernel is broad relative to 25–100 ms timing errors and the offsets cancel across many events, this blurring does not create a new source of pre-onset predictability.

Systematic latency offset. We then fixed the offset standard deviation at 50 ms and shifted the mean from -100 to $+100$ ms (Figure 8**B**). Unlike zero-mean jitter, a systematic offset changes where the same causal response appears in target-relative coordinates. Negative mean offsets move source responses earlier relative to the lexical anchor, increasing the pre-onset mean from $.0629$ at a $+100$ ms mean offset to $.0848$ at a -100 ms mean offset. The maximum negative-lag correlation

likewise increases from about .387 under a $+100$ ms offset to .491 under a -100 ms offset, while the nonnegative peak remains similar in magnitude but shifts from 575 ms to 450 ms.

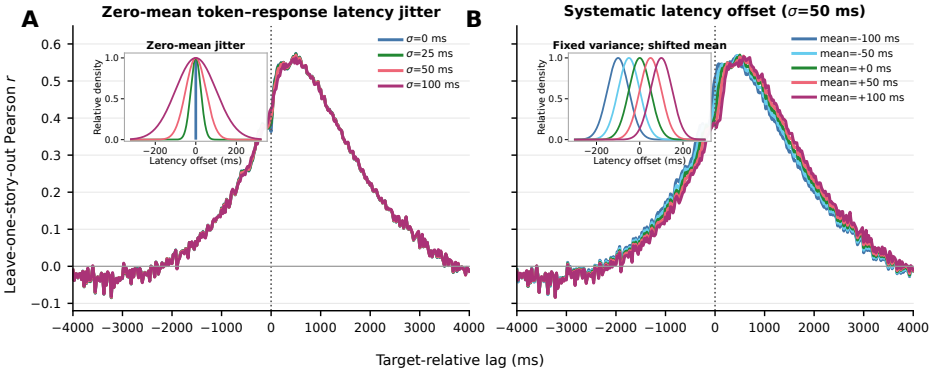


Figure 8: **Temporal-axis precision and alignment controls. A**, Zero-mean token–response latency jitter: per-token offsets are sampled from $N(0, \sigma^2)$ and rounded to the 25 ms grid; random timing imprecision has little effect under the broad default response kernel. **B**, Systematic latency offset: offset variance is fixed at (50 ms)^2, and changing only the mean shifts where the same strictly causal response appears on the target-relative lag axis. Insets show the offset distributions.

This result is a coordinate-system effect. If response timestamps are systematically advanced relative to word onsets, activity elicited by already-onset words is projected into more negative target-relative lags. If timestamps are delayed, the same activity shifts rightward. Temporal-axis accuracy therefore matters for interpreting peak location and pre-onset magnitude, but the offset manipulation still does not introduce any future-word contribution into the generator.

B.6 Positive Control with a Ground-Truth Predictive Signal

The preceding simulations are negative controls: their response kernels are strictly one-sided, so any apparent pre-onset prediction must arise without a future-word contribution. We complemented them with a positive control that asks what conventional encoding and temporal generalization recover when the generator actually contains a target-linked predictive signal. This experiment is also a ground-truth motivation for the Target Isolation analysis in Section 4.2: it tests whether a cross-onset temporal-generalization pattern is specific to genuine target-word prediction or can be reproduced by temporal source mixing.

Predictive-signal construction. We retained the generator, timing distribution, representations, causal response, and observation model from Appendix B.1. For each word j, we added a noncausal component driven by that word's own representation through the same channel weights used for its causal response. Let $w_e \in \mathbb{R}^6$ be the shared linear weights for channel e. The added contribution at absolute time t is

$$S_e^{\text{pred}}(t) = \sum_j x_j^\top w_e \, g(t - O_j), \qquad g(\tau) = \begin{cases} 0.75 \exp\left[-\frac{1}{2}\left(\frac{\tau + 300}{110}\right)^2\right], & -600 \le \tau \le 0 \text{ ms}, \\ 0, & \text{otherwise.} \end{cases} \tag{57}$$

Thus, the target representation begins contributing 600 ms before its own onset, reaches its maximum at -300 ms with amplitude 0.75, and is added to the causal responses of all words on the same absolute neural timeline (Figure 9**A**). Because the latency offset is zero in this experiment, the predictive kernel is anchored to the lexical onset. The on/off comparison is exactly paired: onset gaps, Gaussian innovations, representations, channel weights, causal responses, and observation noise are identical, and only $S_e^{\text{pred}}(t)$ is included or omitted. All conditions use seed 20260721, 12 stories, 180 tokens per story, six-dimensional representations, six response channels, and a 25 ms time step.

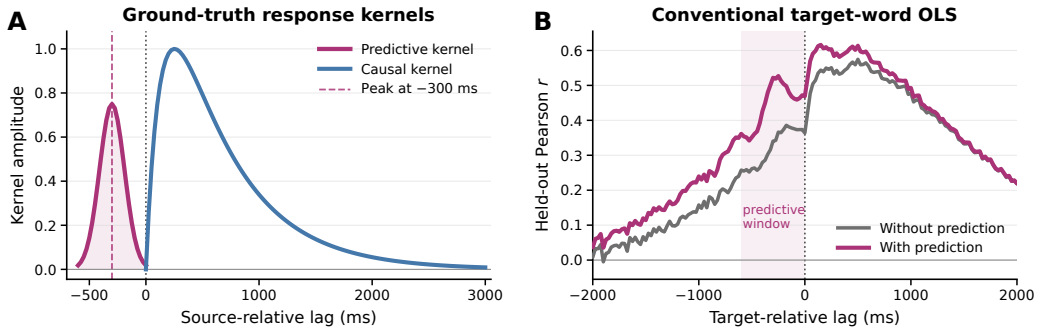


Figure 9: **Ground-truth Gaussian predictive-signal construction. A**, The target-linked predictive kernel $g(\tau)$ is supported from -600 to 0 ms and is added to the one-sided causal rise–decay kernel on the absolute response timeline. Both components are driven by the same token representation and channel weights. **B**, Conventional target-word OLS under the paired default condition, with the predictive interval shaded. Only the predictive component differs between curves.

Figure 9**B** first verifies that the manipulation enters the conventional target-word analysis under the default temporal-source-mixing condition ($\rho = .72$, full response superposition). Adding the predictive component increases the mean held-out r over -600 to 0 ms from .322 to .444, and increases the maximum negative-lag r from .386 at -175 ms to .527 at -250 ms. The prediction-free condition nevertheless retains substantial pre-onset performance, reiterating that a lag–r curve alone does not identify the source of the predictable activity.

Temporal-generalization positive control. We next repeated the factorial mechanism experiment using train-lag×test-lag temporal generalization. At each training lag, an intercept-bearing target-word OLS model was fitted on all but one complete simulated story, held fixed, and evaluated against the response at every testing lag in the omitted story. Out-of-fold tokens were concatenated, Pearson r was computed separately for each channel, and the six channel correlations were equally averaged. Training and testing lags both span -2000 to 2000 ms in 25 ms steps. The diagonal of every matrix reproduces the independently evaluated conventional lag–r curve to a maximum absolute discrepancy below 9.5×10^{-10}. Figure 10 uses one shared color scale for all eight matrices.

Each large panel crosses absent/present sequential dependence ($\rho = 0/.72$) with suppressed/retained response superposition. The top-left matrix is the ground-truth target-only contrast: independent representations prevent the target embedding from accessing other words, and truncation at the next onset prevents their causal responses from overlapping. Without the predictive component (Figure 10**A**), generalization is concentrated in the post-onset first quadrant. When the predictive component is enabled (Figure 10**B**), the same target representation and channel weights drive both the pre-onset Gaussian and post-onset causal responses. This creates the expected paired cross-onset blocks in the second quadrant (post-onset training, pre-onset testing) and fourth quadrant (pre-onset training, post-onset testing): their peak r values increase from $.034/.154$ without prediction to $.287/.355$ with prediction.

The bottom-right matrices expose the identifiability problem. When sequential dependence and response superposition are both present, the prediction-free generator already exhibits prominent red regions in the second and fourth quadrants, with peak $r = .396/.569$. Adding the true predictive component changes their magnitudes to $.526/.618$, but the qualitative cross-onset pattern remains the same. Thus, the raw presence of cross-onset red regions is not diagnostic of a predictive neural source: temporal source mixing can create the same signature that genuine target-linked prediction is expected to produce.

The informative comparison appears only after the target-word contribution is isolated. In the simulation, the ground-truth target-only condition makes the predictive on/off contrast visible directly. In the neural analysis, where sources cannot be switched off experimentally, CAST-FIR approximates this comparison by cross-fitted subtraction of modeled non-target contributions while leaving the target representation unchanged. This is the motivation and inferential logic of Section 4.2: conventional temporal generalization establishes predictive adequacy, whereas the comparison before and after

target isolation tests whether the cross-onset pattern has target-word provenance. The collapse observed after isolation in Figure 5**A** does not prove that the brain never predicts; it shows that the raw red quadrants are largely explained by temporal source mixing and therefore do not, on their own, provide source-specific evidence for a predictive signal.

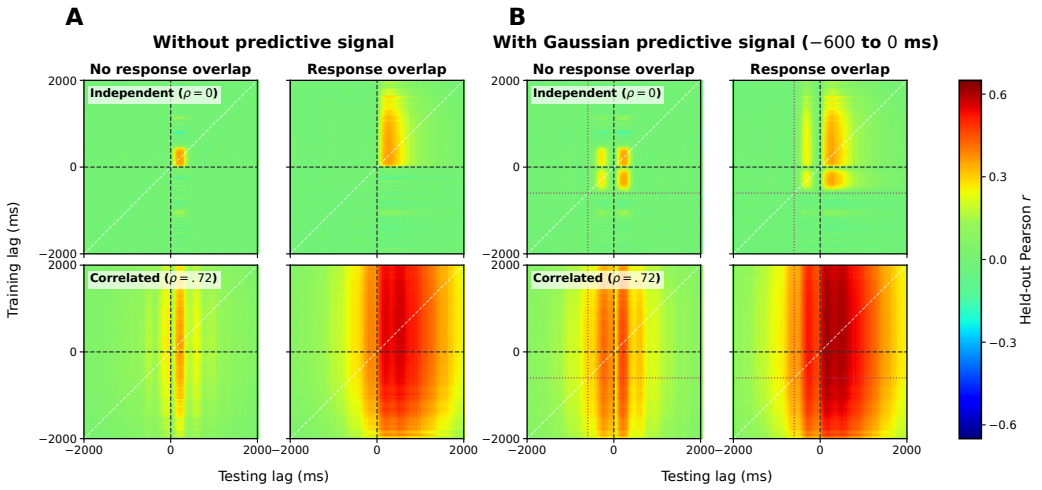


Figure 10: **Temporal generalization with and without a ground-truth predictive signal. A**, Prediction-free generator. **B**, The same paired generator with the target-linked Gaussian predictive component over -600 to 0 ms. Within each large panel, rows manipulate adjacent-representation dependence and columns manipulate causal-response overlap. Horizontal and vertical dashed lines mark target onset; magenta dotted lines in **B** mark the predictive support boundary. All matrices share the same color scale. In the target-only top-left condition, adding prediction produces the expected second- and fourth-quadrant blocks. In the temporal-source-mixing bottom-right condition, prominent cross-onset red regions occur both with and without true prediction.

C Data, Preprocessing, and Cross-Validation Details

Table 5: **Dataset composition, temporal structure, and validation design for the primary neural analyses.** Participant counts are reported as recruited / included, and contact counts give the electrodes retained in the primary cohort. Word counts denote the canonical Chinese token inventory and the fully aligned English events; analysis-specific filters may yield smaller denominators. Mean onset gap is the average interval between adjacent word onsets. $\mathrm{TOI}(W)$, defined in Equation 7, is the mean fraction of an assumed W-ms single-word response that remains when the next word begins; values are reported at $W = 200$, 500, and 1000 ms, with larger values indicating greater expected response superposition. The final row gives the outer cross-validation unit used for held-out evaluation.

	Chinese dataset	English dataset
Recording modality	SEEG	ECoG
Recruited / included participants	8 / 5	8 / 8
Included contacts	46	184
Stories	6	1
Words	2634	5013
Stimulus duration	27.3 min*	29.9 min
Mean onset gap	623.2 ms	352.1 ms
TOI(200 ms)	1.0%	15.9%
TOI(500 ms)	20.5%	47.2%
TOI(1000 ms)	45.3%	68.3%
Primary outer CV	leave-one-story-out	10 contiguous time folds

*Sum of onset spans across the six unique stories, not the duration of a single session.

C.1 ENGLISH ECOG

The English dataset comprises intracranial recordings from eight participants (ages 20–48 years in the original study) who listened to the same approximately 30 min podcast. The official word clock contains 5136 rows, of which 5013 events have complete word order, timing, embeddings, and neural responses. Following the original study, the analysis uses the 184 selected contacts, with participant-specific counts of $15/43/19/8/40/12/16/31$ (Zada et al., 2025a; Bhattacharjee et al., 2026). The naturalistic-listening design follows the broader use of continuous speech to study language processing under ecologically rich temporal structure (Hamilton & Huth, 2020).

Following the preprocessing pipeline of the original study (Goldstein et al., 2022), English neural activity is represented as 70–200 Hz Morlet-wavelet log power. Frequencies near the $120/180$ Hz line-noise harmonics are excluded; power is standardized within frequency and then averaged, smoothed with a 50 ms Hamming window, and spikes are replaced by cubic interpolation. Depending on each participant's noise profile, the data are referenced using either common average referencing (CAR) or independent component analysis (ICA). The target-aligned response is a centered 200 ms window. For the present source-time analysis, however, the original continuous signals were not uniformly available in the current analysis assets; the absolute-time model therefore projects the center of each $O_i + \tau$ window to the nearest 25 ms bin and averages duplicate positions, covering $69533/71368$ bins. Thus, the resulting timeline is reconstructed from windows rather than derived from the raw continuous trace.

For GPT-2 XL L48 (Radford et al., 2019), we use the author-released 5013-row features; L24 is aligned by monotonic word order. For the English conventional and downstream target-word encoding analyses, including temporal generalization, we follow the released implementation of Bhattacharjee et al. (2026): embeddings are centered across all 5013 words, reduced to 50 principal components, and normalized rowwise to unit L_2 norm before tenfold cross-validation over contiguous word segments. This preprocessing uses held-out predictor values, but not neural responses, and is therefore not strictly training-fold-only. For the English source-time FIR models, the scaler and PCA are instead fitted exclusively on non-held-out, non-purged training tokens and then applied to held-out tokens. A 3100-ms purge is applied around each absolute-time test fold to prevent overlap between the temporal support of training and held-out samples.

C.2 CHINESE SEEG

Eight participants were recruited for the Chinese dataset; the primary cohort includes five participants and 46 valid contacts, with participant-specific counts of $5/2/9/20/10$. Four participants completed two presentations of all six stories, and one completed two presentations of five stories. The stimuli comprise 2634 canonical token spans and were presented at $0.45\times$ the speed of the original audio.

Chinese SEEG was sampled at 2 kHz. Session-specific manually identified bad and non-signal channels were removed before referencing. Continuous recordings were notch-filtered at $50/100/150/200$ Hz, when below Nyquist, rereferenced using CAR, band-pass filtered at 70–200 Hz, transformed to analytic-amplitude envelopes using the Hilbert transform, and z-scored within each session$\times$contact. Filtering used MNE-Python's default zero-phase FIR implementation, with automatically selected filter lengths and transition bandwidths. Word-centered epochs spanned $[-2, 2]$ s, and epochs extending beyond the recording boundaries were discarded. Neural responses were computed using centered 200 ms windows.

For Qwen2.5-1.5B (Yang et al., 2024), we use contextual representations from L14/L27. To avoid Qwen's anomalously large activation at the first token, each story is prepended with the model's fixed BOS token (ID 151643); the prefix participates in causal contextualization, but its hidden state is discarded before alignment. Model-token hidden states are mapped to story characters using tokenizer offset mappings and averaged within each canonical timestamp span. The primary cross-validation scheme is leave-one-story-out, with all repeated presentations of the held-out story assigned to the same test fold. To prevent data leakage, the StandardScaler, PCA50, and all encoding and source-time models are fitted within each fold using only the remaining complete stories.

D Acoustic and Representation Controls

D.1 Acoustic Benchmark Setup

We construct a non-neural response from the audio presented in the Chinese and English experiments. Chinese words are aligned to the six original story waveforms using their onset/offset timestamp files and joined to Qwen2.5-1.5B L27 representations by story and token index. English words use the official podcast clock; GPT-2 XL L48 subtoken states are mean-pooled within each word.

The word-level acoustic operator follows the released word-averaged acoustic-feature procedure of Schönmann et al. (2026). Mono waveforms are resampled to $18{,}000$ Hz for Chinese and $9{,}216$ Hz for English. We compute a centered 2048-sample periodic-Hann short-time Fourier transform with an 18-sample hop, project the power spectrum onto eight Slaney-normalized Mel bands, and convert power to decibels with an 80-dB range. For each word, each Mel band is averaged over frames whose timestamps fall strictly inside its onset–offset interval. The ninth coordinate is the variance over all Mel values in that interval. We use polyphase resampling as implemented in SciPy; the resulting features are therefore operator-matched rather than a bitwise reproduction of the public cache.

Absolute-time modeling requires one response value at each query time. We therefore define a physical-onset zero-order hold: after a word begins, its nine-dimensional vector remains active until the next physical word onset; times before the first word are not future-filled. Target epochs sample this process at $O_i + \tau$, with $\tau = -1950, -1925, \ldots, 1950$ ms, and targets without a complete epoch are excluded. This gives 2615 Chinese and 5124 English targets before history matching. The five-method comparison in Figure 4 uses the common subset with eight exact same-story or same-transcript predecessors, yielding 2583 Chinese and 5122 English targets.

The conventional model applies training-fold feature standardization and PCA50 to the current-word representation and fits an intercept-bearing, unregularized multi-output OLS to the raw acoustic epoch. Target Isolation instead fits a separate constrained-availability source-time FIR on a 25-ms absolute-time response grid. Source representations receive fold-local scaling and PCA50, and a multi-output ridge model ($\alpha = 1$, unpenalized intercept) estimates 25-ms source-age bins over a 3000-ms support. At query time t, only sources with onset no later than t enter the design. For target i, the subtracted reconstruction is further restricted to sources with onset strictly earlier than O_i; the target word, simultaneous-onset words, target-relative future words, and FIR intercept are not subtracted. The residual response is then predicted from the unchanged current-word representation using the same fold-local PCA50–OLS pipeline.

The three predictor-side controls all predict the original raw acoustic response. Previous-word residualization first transforms the current and immediately preceding word with the conventional fold-local PCA50, fits a training-fold multivariate OLS from the previous score to the current score, and standardizes the resulting residual using training-fold statistics before downstream OLS. Pairwise projection is performed separately for each token in the original embedding space,

$$x_i^{\mathrm{pair}} = x_i - x_{i-1} \frac{x_i^\top x_{i-1}}{x_{i-1}^\top x_{i-1}}.$$

Previous-eight projection removes the component of x_i in the within-token span of $x_{i-8}, \ldots, x_{i-1}$ using least squares without an intercept. Each geometric residual is subsequently given its own training-fold feature standardization, PCA50, and intercept-bearing OLS.

Chinese evaluation uses six leave-one-story-out folds, so neither response nor history crosses a story boundary. English evaluation uses ten contiguous physical-time folds. A 3100-ms purge surrounds every held-out boundary. The FIR and downstream masks also embargo the complete onset–offset interval of each held-out acoustic-response owner and prohibit reuse of its word row as a training response. Every scaler, PCA, residualizer, FIR, and encoding model is fitted only on the corresponding outer-training data, and every target receives exactly one out-of-fold prediction.

At each lag, Pearson correlation is computed separately for the nine acoustic dimensions within each outer test fold. Correlations are Fisher-transformed, averaged across dimensions within fold and then across folds, and transformed back to r; the ribbons show the back-transformed mean ± 1.96 across-fold Fisher-z standard errors. The pre-onset values reported in the main text are descriptive arithmetic means of these curves over all plotted $\tau < 0$ lag centers. Figure 4**C** additionally summarizes

dependence between each method's 50-dimensional control-adjusted representation z_i, before the final encoding OLS, and the unmodified previous-word embedding x_{i-1}. The additional invertible numerical scaler used after previous-word residualization is omitted from this diagnostic. Within each test fold, we center both matrices, compute $\Gamma_f = Z_f^\top X_f^- / (n_f - 1)$, and report $\|\Gamma_f\|_F / \sqrt{d_z d_x}$; fold values are combined by a quadratic mean weighted by $n_f - 1$.

D.2 Representational-Content Audit

Table 6: **Representation benchmarks on the common previous-eight suffix (scores $\times 100$).** Columns correspond to the five methods in Figure 4, with the conventional OLS baseline labeled Raw. Target Isolation changes the response but not the predictor representation, so its scores equal Raw by construction. Relative to Raw, ↑, ↓, and = denote numerically higher, lower, and unchanged scores, respectively; these symbols do not indicate statistical significance. 1-Proj, 8-Proj, and Linear Res denote the Goldstein pairwise projection, Azizpour previous-eight projection, and Schönmann linear residualization controls. STS-B reports Spearman correlation, PAWS-X cosine average precision, and Amazon 8-shot-per-class linear-probe accuracy.

Language	Benchmark	Raw	Target Isolation	1-Proj	8-Proj	Linear Res
EN	STS-B	34.04	34.04 =	30.72 ↓	31.95 ↓	32.26 ↓
EN	PAWS-X	55.57	55.57 =	54.73 ↓	58.14 ↑	54.62 ↓
EN	Amazon	30.00	30.00 =	25.52 ↓	22.99 ↓	27.48 ↓
CN	STS-B	32.10	32.10 =	21.14 ↓	30.11 ↓	27.79 ↓
CN	PAWS-X	56.14	56.14 =	56.02 ↓	59.83 ↑	56.00 ↓
CN	Amazon	33.31	33.31 =	25.49 ↓	21.99 ↓	29.38 ↓

Predictor-side controls change the representation being evaluated, so attenuation of encoding performance can reflect both temporal control and altered representational content, a distinction that motivates explicit computational controls in natural-language encoding (Toneva et al., 2022). We therefore evaluate the frozen representations on STS-B, PAWS-X, and Multilingual Amazon Reviews (Cer et al., 2017; Yang et al., 2019; Keung et al., 2020; Muennighoff et al., 2022). The protocol uses the common previous-eight suffix subset; PCA, scaling, and residualization parameters are inherited from the corresponding acoustic outer fold and are not refitted on these benchmarks.

Across the 18 predictor-side control comparisons, 16 scores decrease relative to Raw; the only increases occur for PAWS-X under 8-Proj in English and Chinese. Thus, in the large majority of cases, predictor-side controls impair the semantic expressivity of the representations, making any accompanying reduction in neural encoding performance ambiguous.

D.3 Joint Predictor/Response Controls

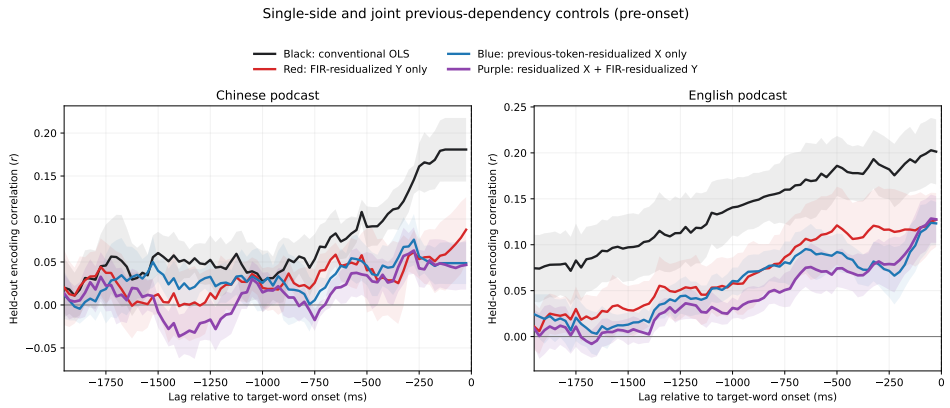


Figure 11: **Joint acoustic benchmark combining predictor- and response-side controls.** The left panel shows the Chinese dataset, and the right panel shows the English dataset.

To test whether the two control axes provide complementary information, we jointly apply previous-word residualization and previous-source subtraction in the acoustic benchmark.

Across all evaluated $\tau < 0$ lag centers, mean held-out r for Chinese/English was $.0727/.1375$ for Raw (black), $.0300/.0690$ for Target Isolation applied only to Y (red), $.0321/.0555$ for Linear Res applied only to X (blue), and $.0146/.0431$ when both controls were combined (purple). Combining predictor- and response-side controls therefore reduced negative-lag predictability beyond either control alone, but the additional decrease was limited and a residual signal remained in both languages.

E Supplementary Neural Source-Audit Results

E.1 Cross-LLM Robustness Audit

Table 7: **Cross-LLM pre-onset encoding before and after Target Isolation.** Values are participant-equally weighted mean held-out r over -1500 to -100 ms under leave-one-story-out validation.

	Middle representation			Last representation		
Model	Raw	Target Isolation	Reduction	Raw	Target Isolation	Reduction
InternLM2-1.8B	.0631	.0103	.0528	.0459	-.0053	.0512
Hunyuan-1.8B-2Bit	.0768	.0008	.0761	.0534	.0011	.0523
LFM2.5-230M	.0804	.0197	.0607	.0460	.0006	.0455
Phi-4-mini-instruct	.0651	.0099	.0552	.0568	.0069	.0499
DeepSeek-VL2-Tiny	.0651	.0078	.0574	.0565	.0005	.0560
GLM-Edge-V-2B	.0697	.0049	.0648	.0574	.0058	.0516

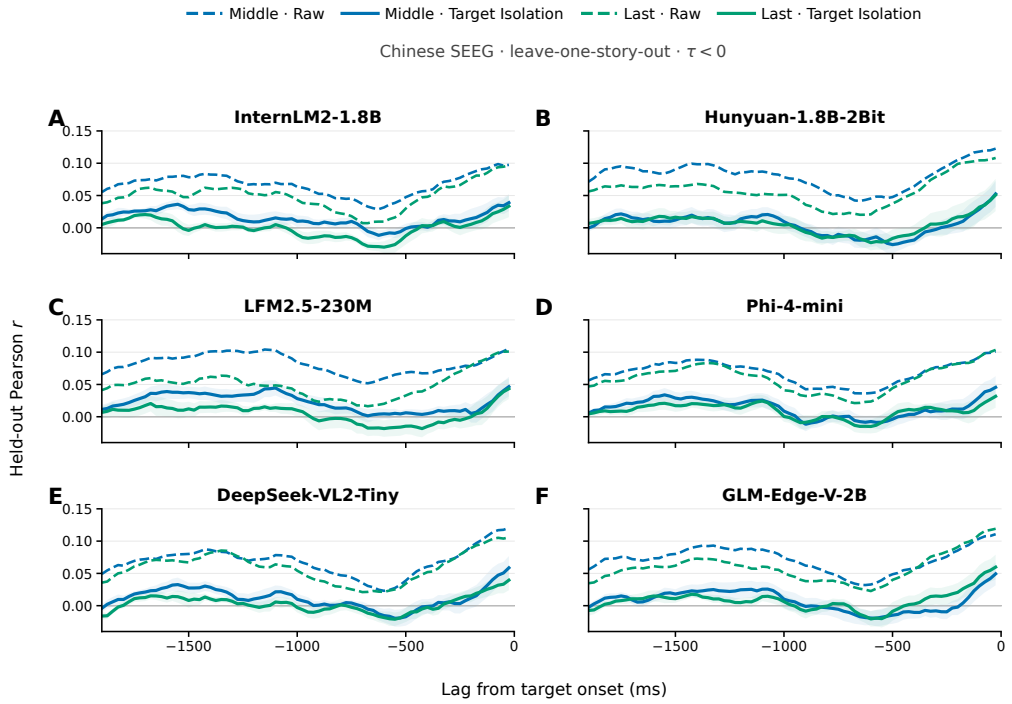


Figure 12: **Cross-LLM robustness of pre-onset encoding and Target Isolation under leave-one-story-out validation. A–F**, complete held-out lag–r curves at $\tau < 0$ for InternLM2-1.8B, Hunyuan-1.8B-2Bit, LFM2.5-230M, Phi-4-mini-instruct, DeepSeek-VL2-Tiny, and GLM-Edge-V-2B. Blue and green denote middle and last representations; dashed and solid lines denote Raw and Target Isolation. Pale bands show participant SEM ($n = 5$) around the Target Isolation curves. Raw window means are positive in all 12 model–layer combinations and decrease after Target Isolation.

Because the Qwen2.5-1.5B result is shown in the main analysis, this audit examines six additional LLM families: InternLM2-1.8B, Hunyuan-1.8B-2Bit, LFM2.5-230M, Phi-4-mini-instruct, the language branch of DeepSeek-VL2-Tiny, and the language branch of GLM-Edge-V-2B. We extract middle and last representations from each model and evaluate them using the same validation pipeline.

Across all 12 model–layer combinations, the participant-equally weighted Raw mean held-out r over -1500 to -100 ms is positive (Table 7). Target Isolation reduces mean r by .0455–.0761 in every combination and confines every post-isolation mean to $|r| \leq .0197$. This cross-model audit shows that both pre-onset encoding and its attenuation after Target Isolation are robust across LLM families and layers.

E.2 Absolute-Time Source-Support Sensitivity

Across the three support lengths ($W = 1000/2000/3000$ ms) and both model layers, all six English contextual profiles have their dominant peak at approximately 75 ms, whereas all six Chinese profiles peak at approximately 225 ms. Apart from modest tail upturns that vary with support length and may reflect boundary effects, the profiles retain matched rise–decay shapes across conditions. Each profile is derived from the L_2 norm of the 50-dimensional coefficient block for each contact, normalized within participant and then averaged across participants with equal weights. The dominant post-onset peak and kernel shape are robust across support lengths and model layers in both datasets.

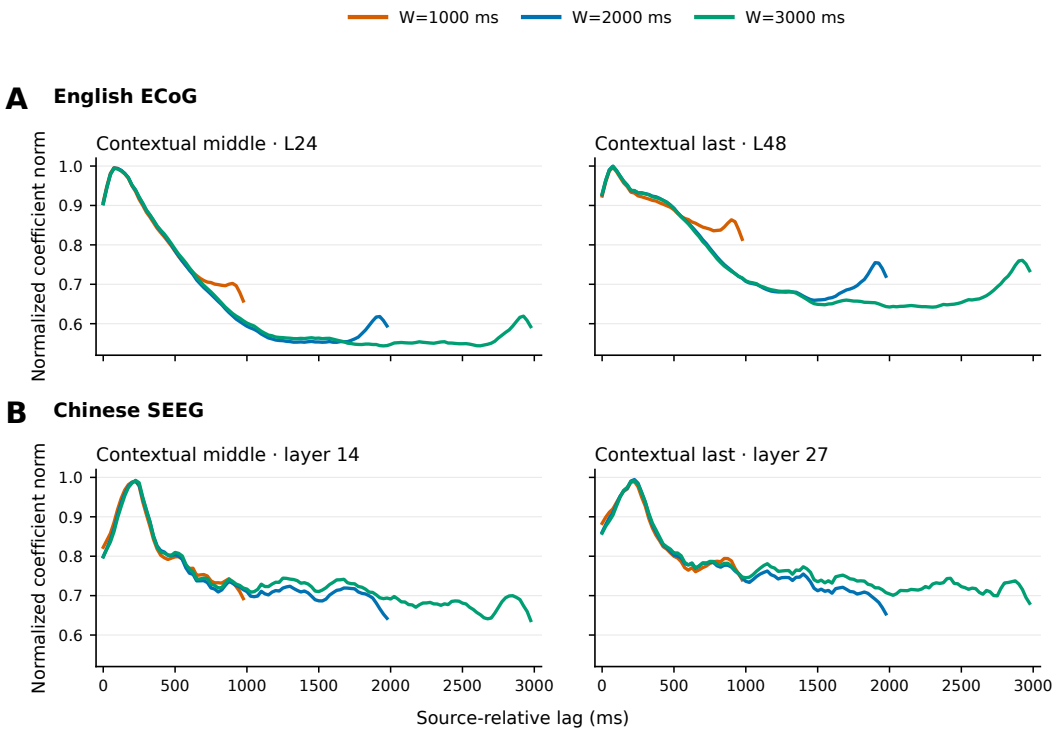


Figure 13: **Coefficient-norm profiles across source supports. A**, English ECoG; **B**, Chinese SEEG. The middle and last representations are shown for $W = 1000/2000/3000$ ms.

E.3 Frozen Source-Set Ablations

The purpose of this ablation is to make the buildup of the target-aligned curve visible as progressively more modeled source contributions are included. We froze the same last-layer, $W = 3000$ ms all-history kernel and changed only the sources summed in each held-out reconstruction. The historical series retains the target and the K most recent preceding sources, whereas the future series retains the target and the next K anchor-relative future sources, for $K = 1, 2, 4, 8, \text{all}$. Scaling, PCA, and the source-time kernel are identical across K; these curves are therefore nested reconstructions from

one fitted model, not separately optimized models. To isolate the relevant buildup, Figure 14 shows historical-source reconstructions only at $\tau < 0$ and future-source reconstructions only at $\tau > 0$.

In English, adding historical sources progressively builds the negative-lag curve, with the reconstruction approaching the direction-specific all-source profile as K increases. The future curves coincide near onset, when few later words are yet available, and separate progressively at later positive lags as additional words reach onset. Chinese reconstructions approach their all-source profiles after fewer sources, consistent with the lower temporal overlap of the slower stimulus. These nested curves visualize how the fitted model assembles its prediction.

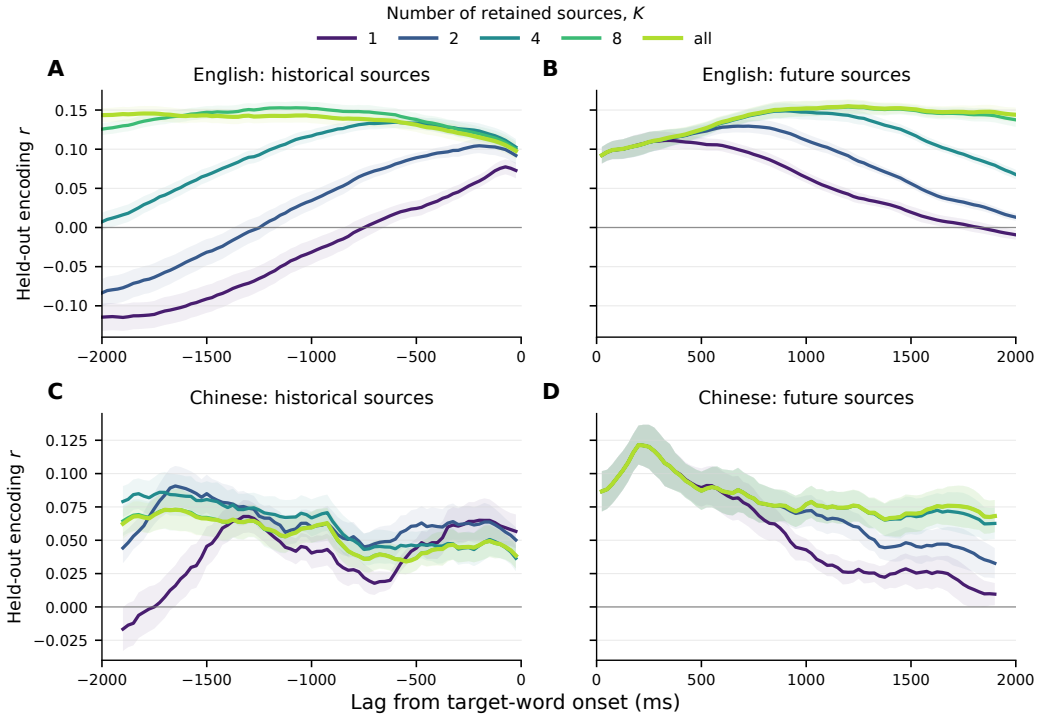


Figure 14: **Buildup of the target-aligned curve under nested source sets.** Top row, English ECoG; bottom row, Chinese SEEG. Left, the target plus the K most recent historical sources, evaluated only before target onset. Right, the target plus the next K anchor-relative future sources, evaluated only after target onset. Lines show participant-equally weighted held-out r, and ribbons show participant SEM (English $n = 8$, Chinese $n = 5$). Every curve uses the same frozen last-layer, $W = 3000$ ms all-history kernel and differs only in the source contributions summed at test time.