USEFUL OR MERELY DIFFERENT? PAIRED-CONTROL EVALUATION FOR NEURAL CONDITIONING OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

Anonymous benchmark code (PairedControlBench)

ABSTRACT

Language models are typically conditioned on text alone, yet many continuous side signals can alter output distributions without the change itself being useful. We study brain-state conditioning as a control problem and introduce a **paired-control evaluation** principle: matched brain input should improve prediction over both removal controls (ZERO) and mismatch controls (SHUF), not just deform the next-token distribution. We quantify this distinction between generic perturbation and **useful deformation** with Jensen-Shannon divergence (JSD) and next-token negative log-likelihood (NLL). We evaluate this principle in two regimes. First, in controlled synthetic TVB experiments, brain conditioning strengthens with dimensionality. In the primary 136-D diagnostic, matched states improve NLL while shuffled and Gaussian-matched states harm it; at 544 dimensions and moderate input scaling, shuffled states can also improve over removal, but matched states improve more and retain the paired advantage. Freezing the brain pathway removes all effects. Second, using real MEG recordings from MEG-MASC, we evaluate a constrained target-word likelihood task with a T5 cross-attention model over post-onset MEG windows. Matched MEG improves target-word NLL over both zeroed and shuffled MEG on random, story-blocked, and subject-blocked splits, with ΔNLL gains of roughly 0.08-0.11. In two independently trained leave-one-subject-out sweeps, both contrasts and their per-subject confidence intervals remain below zero for all 11 held-out subjects. The effect is small and token-level: relative to ZERO, the preferred token changes on about 10% of held-out examples, but target top-1 accuracy changes by only -0.09 percentage points. This is not evidence of full neural decoding or improved standard autoregressive generation. Finally, applying the protocol to a released fMRI-to-text checkpoint changes the conclusion suggested by a mismatch-only test: matched fMRI beats locally mismatched fMRI but is substantially worse than removing fMRI, so it fails the full criterion. Together, these results demonstrate the practical consequence of paired controls and identify a modest, pairing-specific MEG likelihood effect.

1 INTRODUCTION

Large language models (LMs) generate text by predicting the next token from preceding context, thus treating language as a purely text-conditioned process. Biological cognition, in contrast, is shaped by continuously evolving neural states that integrate perception, memory, affect, and internal dynamics (Friston, 2010; Barrett, 2017). We study whether a latent neural state can serve as a control signal that modulates linguistic prediction.

Prior work in brain-computer interfaces (BCI) has established that neural activity contains decodable information about intended communication, including handwriting and speech, using invasive recordings (Anumanchipalli et al., 2019; Makin et al., 2020; Willett et al., 2021; Moses et al., 2021; Metzger et al., 2023). More recently, non-invasive recordings have been used to reconstruct aspects of the semantics of continuous language from fMRI (Tang et al., 2023). While these efforts primarily

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.

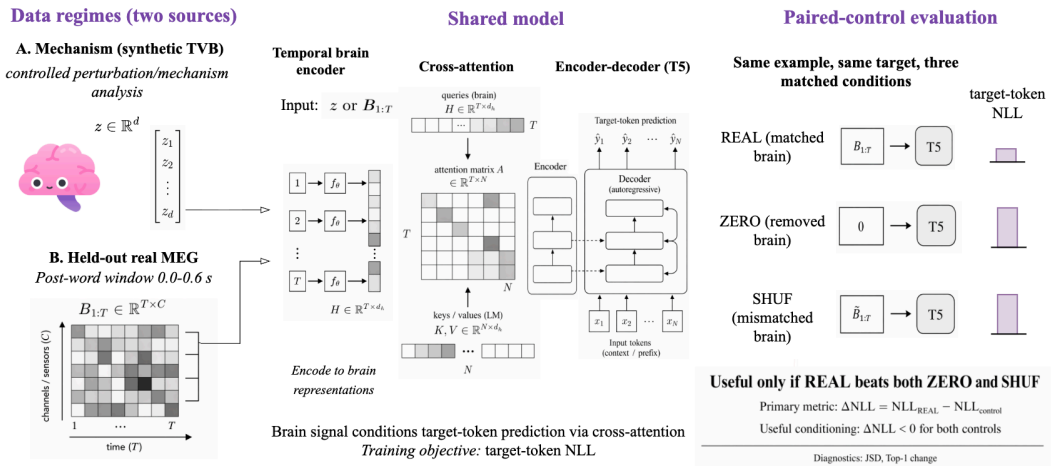

Figure 1: Overview of the pipeline. Left: two data regimes, a constructed synthetic TVB setting for mechanism and a held-out real post-onset MEG setting with windows $B_{1:T} \in \mathbb{R}^{T \times C}$. Middle: both regimes use the same temporal brain encoder and T5 cross-attention model with target-token NLL training. Right: paired-control evaluation compares matched brain input (REAL) against removed-brain (ZERO) and mismatched-brain (SHUF) controls, conditioning is useful only if REAL improves prediction over both.

emphasize decoding language from neural measurements, our focus is different: we treat a brain-state vector as an exogenous input to a language model and ask whether it produces measurable, structured control over the model's next-token distribution.

A related line of work studies controllable language generation, where auxiliary signals such as attributes, prompts, or learned control vectors are used to steer model outputs (Keskar et al., 2019; Dathathri et al., 2020). These methods show that relatively small conditioning signals can substantially alter a model's predictions. For paired-control evaluation, however, distributional change alone is insufficient: a neural signal may shift the next-token distribution without making the correct continuation more likely. We therefore evaluate brain conditioning along two axes: how much the signal changes the predictive distribution, and whether that change improves target-token likelihood.

We study brain-conditioned language modeling using paired text contexts and brain-state measurements. Given a context $x_{1:t}$ and brain input z, we compare next-token distributions under matched, removed, and mismatched brain conditions. This paired comparison separates two effects that are often falsely combined: deformation, the extent to which brain input changes the predictive distribution, and utility, whether that change increases probability assigned to the observed continuation. We quantify the deformation by Jensen-Shannon divergence (JSD) between $p(\cdot \mid x_{1:t}, z)$ and the zero-brain baseline $p(\cdot \mid x_{1:t}, 0)$, and the utility by next-token negative log-likelihood (NLL) on the ground-truth token. For real MEG-MASC, the resulting task is a constrained target-word likelihood evaluation rather than standard full-context autoregressive generation.

We evaluate the framework in two regimes. We first use TVB-generated synthetic brain signals in a controlled setting where the current token drives a token-specific stimulus pattern in a noisy network simulation and the resulting state is paired with that text prefix. This lets us study dimensionality scaling, token sensitivity, and low-rank control geometry without making claims about natural brain-language alignment. We then turn to real MEG-MASC recordings, where the brain input is a post-onset MEG window $B_{1:T} \in \mathbb{R}^{T \times C}$ and the key question is whether any pairing-specific effect survives held-out evaluation.

Across both regimes, paired controls distinguish useful deformation from generic distributional change. In synthetic TVB experiments, matched and mismatched signals can induce comparable deformation. Matched states retain the paired advantage, although at high dimension and moderate scaling even SHUF can improve over ZERO, illustrating why both controls are necessary. In

real MEG-MASC, post-onset MEG yields modest but consistent target-word NLL gains over both ZERO and SHUF across random, story-blocked, subject-blocked, and leave-one-subject-out evaluations. On a released fMRI-to-text checkpoint, the protocol also rejects a result that would appear positive against SHUF alone because ZERO performs best. Diagnostic failures with leakage-prone decoding, poor signal scaling, Gaussian streams, and peri-word windows delimit the claim. To make this evaluation protocol reusable beyond the present experiments, we package the paired-control metrics, confidence intervals, and mismatch-control variants in **PairedControlBench**, a toolkit for matched/removal/mismatch evaluation of auxiliary conditioning signals.

2 Brain-Conditioned Language Modeling and Evaluation

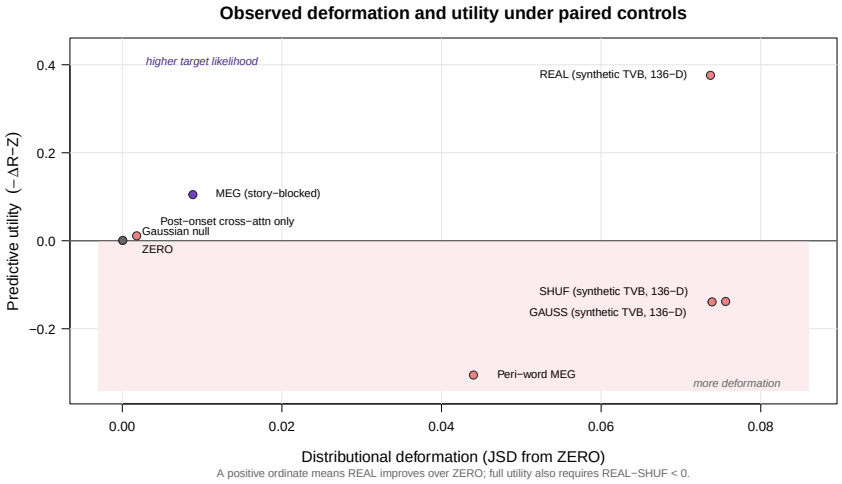

Figure 2: Quantitative summary using the exact values reported in the corresponding tables. The horizontal axis shows JSD from ZERO and the vertical axis shows $-\Delta R\text{-}Z$; the full paired-control criterion additionally requires REAL to outperform SHUF. The rejected BERT+MEG residual is omitted because it does not robustly beat the nuisance-preserving mismatch.

Let $x_{1:t}$ denote a sequence of tokens and $z \in \mathbb{R}^d$ an auxiliary brain-state vector with d components. In the synthetic regime, we train an autoregressive language model parameterized by θ to predict the next token conditioned on both text and a single TVB state:

$$p_\theta(x_{t+1} \mid x_{1:t}, z) = \mathrm{softmax}(f_\theta(x_{1:t}, z)) \,. \tag{1}$$

The synthetic model is trained using standard next-token cross-entropy:

$$\mathcal{L}(\theta) = -\sum_t \log p_\theta(x_{t+1} \mid x_{1:t}, z). \tag{2}$$

Brain conditioning is implemented via a learned projection that injects z into the transformer hidden states. When $z = \mathbf{0}$, the model reduces to a standard text-only LM.

For real MEG-MASC, the brain input is temporal rather than a single snapshot. We use a post-onset MEG window $B_{1:T} \in \mathbb{R}^{T \times C}$, project it with a small temporal brain encoder, and feed the resulting sequence to a T5 encoder-decoder model via encoder-side continuous embeddings (see Figure 1). For a target word tokenized as $y_{1:m}$, the decoder receives only shifted target tokens rather than the preceding story context:

$$p_\theta^{\mathrm{MEG}}(y_{1:m} \mid B_{1:T}) = \prod_{j=1}^{m} p_\theta(y_j \mid y_{<j}, B_{1:T}). \tag{3}$$

The T5 decoder states supply cross-attention queries, while the encoded MEG sequence supplies keys and values. This architecture preserves temporal structure and lets us evaluate whether the correct MEG window improves prediction over zeroed or shuffled brain controls.

2.1 PAIRED-CONTROL EVALUATION OF BRAIN INFLUENCE

For each example, let c denote the decoder-side linguistic context allowed by the protocol, a the auxiliary input, and $y_{1:m}$ the supervised target. In the synthetic regime $c = x_{1:t}$ and $m = 1$; in the MEG regime c contains no preceding story tokens and consists only of the shifted target prefix $y_{<j}$. We evaluate the same trained model and target under three conditions:

$$p_{\mathrm{REAL}} = p_\theta(\cdot \mid c, a), \tag{4}$$

$$p_{\mathrm{ZERO}} = p_\theta(\cdot \mid c, \mathbf{0}), \tag{5}$$

$$p_{\mathrm{SHUF}} = p_\theta(\cdot \mid c, \tilde{a}), \tag{6}$$

where a is paired with the example, $\mathbf{0}$ removes the auxiliary signal, and $\tilde{a}$ is drawn from another example so that marginal signal statistics and specified nuisance structure are preserved while exact sample-wise pairing is broken.

We use two complementary metrics. First, we measure **deformation**, the degree to which a nonzero brain signal changes the model's predictive distribution relative to the zero-brain baseline. For a distribution p induced by a brain input and the zero-brain distribution $q = p_{\mathrm{ZERO}}$, we compute Jensen-Shannon divergence:

$$\mathrm{JSD}(p \parallel q) = \frac{1}{2}\mathrm{KL}(p \parallel m) + \frac{1}{2}\mathrm{KL}(q \parallel m), \quad m = \frac{1}{2}(p + q). \tag{7}$$

We also report top-1 agreement with ZERO, the fraction of examples for which $\arg\max p = \arg\max p_{\mathrm{ZERO}}$. For MEG, aggregate JSD and top-1 agreement use the final supervised target position; first-token probability and rank analyses are labeled separately.

Second, we measure **utility**, whether the induced deformation increases probability mass on the supervised target. For condition $C \in \{\mathrm{REAL}, \mathrm{ZERO}, \mathrm{SHUF}\}$, the MEG evaluation uses mean token NLL within each target word,

$$\ell_C(y_{1:m}) = -\frac{1}{m}\sum_{j=1}^{m} \log p_C(y_j \mid c, y_{<j}), \tag{8}$$

and then averages these word-example values. The synthetic next-token case is recovered with $m = 1$. We define

$$\Delta_{\mathrm{R-Z}} = \ell_{\mathrm{REAL}}(y) - \ell_{\mathrm{ZERO}}(y), \tag{9}$$

$$\Delta_{\mathrm{R-S}} = \ell_{\mathrm{REAL}}(y) - \ell_{\mathrm{SHUF}}(y). \tag{10}$$

Negative values indicate that REAL assigns higher probability to the target token than the corresponding control.

Definition 2.1 (Useful paired conditioning)**.** A conditioning signal is useful under paired controls if matched input improves target-token prediction over both removal and mismatch controls:

$$\Delta_{\mathrm{R-Z}} < 0 \quad \text{and} \quad \Delta_{\mathrm{R-S}} < 0. \tag{11}$$

A signal that increases JSD or changes top-1 predictions without satisfying these inequalities is treated as generic deformation rather than useful conditioning.

This definition separates three regimes that are otherwise easy to conflate: (i) no effective control, where the side signal produces little distributional change, (ii) generic perturbation, where the side signal changes the distribution but does not improve the target likelihood, and (iii) useful deformation, where the matched signal changes the distribution in a direction that improves prediction over both ZERO and SHUF. All synthetic and MEG experiments below use this paired-control criterion. Although our specific case study uses neural signals, the same protocol applies to any auxiliary conditioning channel for which matched, removal, and mismatch controls can be defined.

2.2 PAIREDCONTROLBENCH TOOLKIT

We package the paired-control protocol as **PairedControlBench**, a reusable evaluation toolkit for auxiliary-signal conditioning experiments. Given outputs under matched, removed, and mismatched inputs, it computes ΔR-Z, ΔR-S, JSD from ZERO, top-1 agreement, and paired or clustered uncertainty. It also provides global, within-group, local temporal, circular-shift, block-permutation, and phase-randomized mismatch controls.

2.3 DATA

We construct paired datasets of $(x_{1:t}, z_t, x_{t+1})$ triples by combining text from a fixed general-domain corpus derived from Wikipedia (Wiki-40B, following Raffel et al. (2020)) with simulated multi-region activity generated using The Virtual Brain (TVB) Schirner et al. (2022). TVB is a dynamical simulator, not a language-to-brain model, so the correspondence is explicitly constructed. Each tokenizer ID indexes a seed-fixed bipolar stimulus pattern over designated language nodes. For an example ending in token x_t, that pattern drives a noisy TVB trial that is reset for the example; we record the full-node state at the gated decoder readout, or the approximately 200-ms activity snapshot when no decoder readout is available, and pair it with the text prefix to predict x_{t+1}. Thus REAL follows a token-conditioned generative mapping rather than an arbitrary row-index assignment. SHUF permutes snapshots across examples, and GAUSS retains only feature-wise low-order statistics. The random 90/10 synthetic split holds out examples from this constructed stream, but adjacent prefixes can overlap and the same token-to-stimulus map appears in both partitions. We therefore use it only as a pipeline sanity check for whether paired controls recover a known correspondence, not as corpus-disjoint language generalization or biological evidence. We evaluate state widths $d \in \{68, 136, 272, 544\}$, each with approximately 100k examples and a separately trained LM. The retained readouts use generic labels and generated coordinates rather than a named anatomical parcellation; Appendix B gives the dimensionality interpretation.

For real-neural validation, we use MEG-MASC (Gwilliams et al., 2023), a natural-speech MEG dataset with word- and phoneme-level timing annotations. We preprocess KIT MEG recordings, build word-aligned examples, and use post-onset windows from 0.0 to 0.6 seconds after word onset. A broader peri-word window (-0.1 to $+0.8$ seconds) was also tested, but it performed worse than ZERO and is therefore treated as a negative diagnostic rather than the main result. The real-data experiment contains 200k word-aligned examples. To prevent decoder-context leakage, the decoder is trained in a target-only mode: it receives shifted target tokens, but not the preceding story context. The brain window is normalized per example before the temporal brain encoder. This makes the real-data experiment a controlled target-word likelihood modulation test under paired controls rather than a standard full-context autoregressive language-modeling setup. We report random train/test splits, story-blocked held-out evaluation to reduce story-level leakage, subject-blocked evaluation, and leave-one-subject-out (LOSO) subject generalization. Appendix Table 14 lists the exact model, preprocessing, optimization, and split settings used for the main MEG runs.

3 CONTROLLED SYNTHETIC VALIDATION OF PAIRED CONTROLS

The synthetic TVB regime is not meant to establish natural brain-language alignment. Instead, it provides a controlled sanity check in which sample-wise correspondence is known by construction, so matched input should retain an advantage over both removal and mismatch controls when that correspondence is preserved.

3.1 PAIRED CONTROLS IN A CONSTRUCTED SYNTHETIC REGIME

Unless noted otherwise, all synthetic metrics are computed on 5,000 held-out contexts using the paired-control criterion in Definition 2.1. The central synthetic question is whether paired-control evaluation distinguishes useful conditioning from generic perturbation when the text-brain correspondence is assigned sample-wise and therefore known by construction.

Table 1: Controls for the 136-D brain-conditioned model.

Condition	JS mean	Top-1	NLL mean	ΔNLL
REAL	0.0737	0.661	2.417	-0.376
ZERO	-	1.000	2.793	0.000
SHUF	0.0756	0.655	2.931	$+0.138$
GAUSS	0.0739	0.654	2.932	$+0.139$

Table 1 is the core synthetic sanity check. In this primary 136-D diagnostic, REAL, SHUF, and GAUSS all produce substantial distributional deformation relative to ZERO, but only REAL improves next-token likelihood. This is the main synthetic point: changing the distribution is easy, while useful deformation requires preserving the constructed pairing.

The supporting synthetic analyses are moved to the appendix. Across dimensionalities, control strength grows monotonically in JSD and top-1 change (Appendix Table 6). Alpha-scaling sweeps

show that deformation rises smoothly with control strength while predictive utility depends on whether the signal is paired correctly (Appendix Figures 4-6). Token sensitivity and PCA further show content-weighted, low-rank deformation without changing the central conclusion (Appendix Tables 7 and 8).

4 External Falsification Case: BrainDEC

To test the protocol outside our models, we re-evaluate the released stage-1 fMRI-to-text checkpoint of Hmamouche et al. (2024) without retraining. On its complete 250-example test split, REAL uses matched fMRI, ZERO removes it, and SHUF cyclically reassigns chunks within each subject and five-chunk conversation, preserving participant, recording block, visual stimulus, and local context.

Table 2: Paired-control reanalysis of the released BrainDEC stage-1 checkpoint. Lower teacher-forced target NLL is better; confidence intervals are paired over 250 test examples.

Comparison	NLL control	ΔREAL-control	95% CI
REAL vs. ZERO	3.4075	$+1.4994$	$[+1.2556, +1.7431]$
REAL vs. SHUF	5.1519	-0.2450	$[-0.4813, -0.0087]$

REAL has NLL 4.9069: it beats SHUF but is substantially worse than ZERO, so Definition 2.1 rejects it. Conversation-clustered intervals preserve both conclusions (R-Z $[+1.2159, +1.7959]$; R-S $[-0.4287, -0.0718]$). This is a scoped test of one released checkpoint and target-likelihood endpoint, not of BrainDEC's separate generation metrics, but it supplies an external case where the two-control standard changes the conclusion.

5 Real MEG-MASC: held-out paired-control evaluation

We next test whether the same REAL/ZERO/SHUF logic transfers to real neural recordings. Using MEG-MASC, we train a T5 encoder-decoder model in which a temporal brain encoder maps a post-onset MEG window (0.0-0.6 seconds after word onset) into T5 encoder embeddings, and the decoder predicts the target word tokens via cross-attention. The final model is trained with target-only decoder inputs to prevent decoder-context leakage, per-example brain normalization, and a mostly frozen T5 in which we unfreeze decoder cross-attention in the last six decoder blocks together with the final encoder and decoder blocks, their final layer norms, and the LM head. This experiment tests whether matched post-onset MEG improves target-token likelihood under paired controls.

This setup exposed two important negative results before producing a usable signal. First, when decoder-side text context was available under standard teacher forcing, REAL, ZERO, and SHUF controls were essentially indistinguishable: the language model could rely on text priors without using MEG. This failure is consistent with evidence that implicit teacher forcing and strong decoder priors can inflate EEG-to-text evaluations and make noise inputs appear competitive with neural inputs (Jo et al., 2024). Second, an initial z-scoring implementation added a fixed 10^{-6} constant to channel standard deviations. Because raw MEG values are stored in very small physical units, this collapsed the effective brain inputs toward zero, a diagnostic batch had mean absolute input magnitude $\approx 1.4 \times 10^{-7}$ and nearly identical logits for REAL and ZERO. After correcting these issues, a conservative cross-attention-only model produced only a tiny in-sample effect (reported in Appendix C). We therefore evaluate the stronger post-onset hybrid configuration on three held-out regimes: a standard random split, a story-blocked split that withholds entire stories at test time, and a subject-blocked split. We then test subject generalization more aggressively with leave-one-subject-out (LOSO) evaluation over all 11 subjects.

5.1 Held-out Paired-Control Results

Figure 3 shows that REAL beats ZERO/SHUF on random ($-0.1057/-0.1155$), story-blocked ($-0.1048/-0.0778$), and subject-blocked ($-0.0795/-0.0824$) tests, with all paired intervals below zero. The story-blocked R-Z effect is an $e^{0.1048} \approx 1.11\times$ target-probability multiplier, not open-vocabulary decoding.

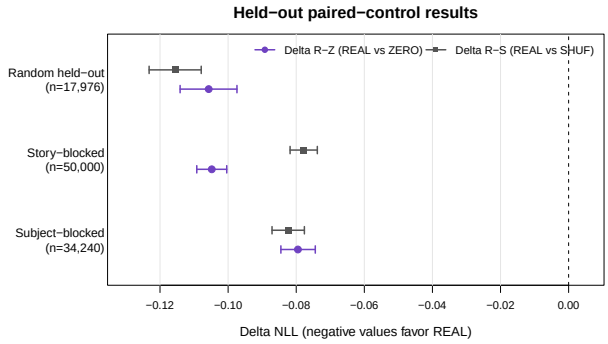

Figure 3: Held-out post-onset hybrid results for MEG-MASC/T5. Below-zero ΔNLL means improved NLL under REAL relative to the control.

Top-1 agreement with ZERO is 0.87-0.90: REAL changes the preferred token on a minority of examples, but this is not a 10-point accuracy gain. On the story-blocked split, target top-1 accuracy changes by -0.0009, while top-5, top-10, and mean reciprocal rank change by $+0.0060$, $+0.0081$, and $+0.0021$.

The story-blocked three-seed means are -0.1081 ± 0.0033 (R-Z) and -0.0879 ± 0.0136 (R-S); subject-blocked means are -0.0634 ± 0.0254 and -0.0686 ± 0.0168, with every seed negative. Subject-clustered bootstrap also preserves the story-blocked effect (95% CIs $[-0.1146, -0.0967]$ and $[-0.0870, -0.0701]$). Removing per-example normalization instead collapses deformation and utility (ΔR-Z/R-S ≈ 0), showing that signal scaling is necessary. A story-clustered interval is not identifiable because this test sample contains one story.

5.2 Robustness to Nuisance-Preserving Mismatch Controls

Replacing the original SHUF baseline with stricter mismatch controls-global-sample SHUF, within-subject SHUF, within-sound SHUF, a joint within-subject+sound local-time SHUF that keeps donors within a ± 32-token word-index neighborhood, circular time shift, local block permutation, and phase-randomized MEG-also preserves a negative story-blocked REAL-vs. SHUF gap in every case (ΔR-S from -0.0665 to -0.0889, all 95% CIs below zero).

In particular, the same-subject+same-sound local-time control remains strongly below zero (ΔR-S $= -0.0808$, 95% CI $[-0.0848, -0.0768]$), indicating that the held-out gain is not explained by subject identity, sound identity, or coarse local temporal proximity alone. Post-hoc covariate matching further reduces several nuisance explanations without removing the effect. Matching examples by subject, sound, word type, target length, story position, and training-frequency bin still gives a negative cell-balanced story-blocked effect for all tokens (ΔR-Z $= -0.1022$, 95% CI $[-0.1147, -0.0897]$; ΔR-S $= -0.0775$, 95% CI $[-0.0853, -0.0697]$), and the same remains true on a matched content-word-only subset (ΔR-Z $= -0.0452$, 95% CI $[-0.0549, -0.0355]$; ΔR-S $= -0.0560$, 95% CI $[-0.0625, -0.0495]$). A still stricter mismatch control that shuffles MEG only within the same subject and sound likewise preserves the REAL advantage (ΔR-S $= -0.0782$, 95% CI $[-0.0822, -0.0741]$).

LOSO evaluation strengthens this further. At seed 42, the 11 subject means are -0.08192 ± 0.01980 for REAL-ZERO and -0.07957 ± 0.02449 for REAL-SHUF; population-level tests give $t(10) = -13.72$ and -10.78 ($p < 10^{-6}$ for both), with exact two-sided Wilcoxon $p = 0.00098$ for each. A second independently trained LOSO sweep (seed 7) gives -0.07296 ± 0.01716 and -0.07313 ± 0.01537. Both contrasts and their per-subject paired 95% CIs are below zero for every subject at both seeds (44/44 subject-by-contrast intervals). This establishes subject transfer under shared narrative stimuli, but not simultaneous generalization to unseen subjects and unseen stories.

The gain is not confined to function words. Across LOSO subjects, content-word means remain negative for REAL-ZERO (-0.0300, 95% CI $[-0.0422, -0.0177]$) and REAL-SHUF (-0.0623, $[-0.0758, -0.0487]$), with both contrasts negative for all 11 subjects. Function-word effects are

Table 3: Auxiliary-stream null and positive controls on the story-blocked split. Gaussian streams fail to produce a pairing-specific gain, whereas a context-only BERT side channel is strongly exploited by the same pathway. Each row is a separately trained model, so absolute NLL scales should be interpreted within rows through the paired contrasts rather than compared across auxiliary inputs.

Auxiliary input	n	NLL REAL	NLL ZERO	NLL SHUF	ΔR-Z	95% CI	ΔR-S	95% CI	Top-1 R/Z
Gaussian iid stream	50,000	7.2048	7.2057	7.2048	-0.0008	$[-0.0010, -0.0006]$	$+0.0000$	$[-0.0002, +0.0002]$	0.9994
Gaussian global stream	50,000	7.1820	7.1832	7.1821	-0.0012	$[-0.0013, -0.0010]$	-0.0000	$[-0.0001, +0.0001]$	0.9990
Context-only BERT	50,000	14.1450	17.0439	15.6755	-2.8989	$[-2.9606, -2.8372]$	-1.5305	$[-1.5900, -1.4710]$	0.0493

larger ($-0.1887/-0.1151$), so the result cannot be interpreted specifically as semantic recovery; stimulus-locked acoustic or phonological information, syntax, timing, prosody, and lexical predictability remain viable contributors.

5.3 Null and Positive Controls for the Auxiliary Pathway

The paired-control gains above depend on the model using a continuous auxiliary stream through the brain-conditioning pathway. We therefore run two additional sanity checks on the story-blocked split. First, we ask whether arbitrary continuous inputs can spuriously satisfy the paired-control criterion. To test this, we replace each MEG window with matched-shape Gaussian control streams and retrain the same architecture. Second, we ask whether the auxiliary pathway is capable of carrying useful information when such information is present. For this positive control, we replace MEG with fixed context-only BERT Devlin et al. (2019) embeddings and pass them through the same conditioning pathway. This BERT stream is not a brain signal - it is an intentionally informative non-brain side channel used to verify that the architecture can exploit meaningful auxiliary information.

Both Gaussian variants fail to produce pairing-specific utility (ΔR-S ≈ 0; negligible JSD; top-1 agreement ≈ 0.999). Subject-level JZS tests give $BF_{01} = 3.32$ for IID Gaussian input and 2.82 for global Gaussian input at $r = 1/\sqrt{2}$. Conversely, context-only BERT is strongly exploited (ΔR-Z $= -2.899$, ΔR-S $= -1.530$). An exploratory frozen-BERT+MEG model that appeared positive under global SHUF fails the strict same-subject, same-sound, local-time test: across three donor seeds, ΔR-S is only $-0.000236/-0.000237/-0.000336$, and only one interval excludes zero. We exclude it from positive evidence; Appendix Table 17 gives full results.

Table 4: Story-blocked temporal ablation for the post-onset hybrid MEG-MASC/T5 model. The full 0.0-0.6 s window is strongest; the early half fails the paired criterion, the late half is weaker, and a narrow late slice fails.

Window	Bins	ΔR-Z	ΔR-S	Top-1
0.0-0.6s	120	-0.1048	-0.0778	0.8997
0.0-0.3s	60	-0.0001	-0.0000	1.0000
0.3-0.6s	60	-0.0227	-0.0206	0.9976
0.45-0.6s	30	$+0.0005$	$+0.0001$	0.9999

5.4 Temporal Localization of the MEG Effect

Each temporal window uses a separately trained model, not an inference-time truncation. The early half (0.0-0.3 s) has a tiny REAL-ZERO shift ($\approx -7 \times 10^{-5}$) but no paired gain (ΔR-S ≈ 0; $BF_{01} = 3.25$); the late half (0.3-0.6 s) is weaker ($-0.0227/-0.0206$), and the narrow 0.45-0.6 s model fails ($+0.0005/\approx 0$). Utility therefore requires the broader post-onset window and remains far from fluent generation or robust word decoding.

6 Discussion

The constructed TVB experiment tests controllability, not natural brain-language alignment. Deformation grows with dimensionality: in the 136-D diagnostic SHUF and GAUSS worsen prediction, whereas at 544-D SHUF improves over ZERO at moderate scales ($-0.238/-0.337/-0.280$) but REAL improves more ($-0.321/-0.488/-0.577$). Thus SHUF is not categorically harmful, and improvement over removal alone is insufficient.

For MEG-MASC, the post-onset T5 model beats both controls on random, story-blocked, subject-blocked, and LOSO tests. This is modest target-word likelihood modulation, not standard autoregres-

sive generation, but it survives independent unseen-story and unseen-subject evaluations and is not reproduced by Gaussian streams. The mechanistic evidence must remain separated by regime. In the constructed TVB diagnostic, token sensitivity and PCA show that deformation is content-weighted and low-rank. In real MEG, separately trained temporal-window models show that the paired gain requires the broader post-onset interval, while similar REAL and SHUF attention profiles caution against interpreting a late attention peak as neural localization.

6.1 Minimum Reporting Standard

These results motivate a simple reporting standard for auxiliary-signal conditioning. Because distributional change alone can be mistaken for useful conditioning, studies should report five items: (i) explicit definitions of REAL, ZERO, and SHUF, including what nuisance structure the mismatch control preserves; (ii) the decoder linguistic-context regime—full preceding context, target-only teacher forcing, or free-running generation—together with the exact target aggregation; (iii) both utility metrics, such as ΔR-Z and ΔR-S, and deformation metrics, such as JSD from ZERO; (iv) at least one held-out split and one nuisance-preserving mismatch control with paired or clustered uncertainty estimates; and (v) an explicit falsification criterion, for example, failure of REAL to beat both ZERO and SHUF, or collapse under held-out or nuisance-preserving controls. The BrainDEC reanalysis in Section 4 shows why both controls matter empirically: matched fMRI beats a strict mismatch but remains harmful relative to removal.

7 Related Work

Neural decoding reconstructs handwriting, speech, or semantics from invasive and non-invasive recordings (Anumanchipalli et al., 2019; Makin et al., 2020; Willett et al., 2021; Moses et al., 2021; Metzger et al., 2023; Tang et al., 2023). Non-invasive approaches include contrastive speech retrieval (Défossez et al., 2023), discrete EEG tokenization (Duan et al., 2023), multimodal/LM decoders (Huang et al., 2025; Hmamouche et al., 2024), and reusable EEG/MEG representations (Xiao et al., 2025; Jayalath & Parker Jones, 2026). Their heterogeneous retrieval, translation, and generation metrics are not directly comparable to target-only NLL. Complementarily, we test whether the same model improves over removal and mismatch; the BrainDEC case shows that passing mismatch alone is insufficient.

EEG-to-text audits find that implicit teacher forcing can inflate metrics and that noise may approach EEG performance (Jo et al., 2024); hence our reporting standard includes decoder context and noise controls, without claiming transfer to free generation. Controllable generation and parameter-efficient steering likewise show that continuous signals can alter models (Keskar et al., 2019; Dathathri et al., 2020; Perez et al., 2018; Houlsby et al., 2019; Hu et al., 2021; Li & Liang, 2021; Lester et al., 2021; Scalena et al., 2024); PairedControlBench tests whether such deformation is useful. Appendix Table 5 provides a task-level comparison.

8 Conclusion

We formulate neural conditioning as a paired-control problem. Constructed TVB pairs validate the criterion; real MEG gives a modest, repeatable likelihood gain; and a released fMRI checkpoint shows that mismatch alone can mislead. Neural conditioning should therefore be judged by removal and nuisance-preserving mismatch controls, population-aware uncertainty, and explicit falsification—not output change alone.

AI use statement

Generative AI tools were used to assist with software development and debugging, experimental-analysis planning and interpretation checks, literature discovery and summarization, figure preparation, and editing and formatting the manuscript. All AI-assisted code and analyses were executed and checked by the authors against the underlying outputs, cited literature was manually verified, and the authors take responsibility for the final content of the submission.

Ethics statement

This work uses previously released synthetic data and public MEG recordings rather than collecting new human-subject data. The principal societal risks are privacy-invasive or surveillance-oriented uses of neural-language technology, uneven generalization across participants, and overstatement of modest likelihood effects as practical mind-reading. We mitigate these risks through narrow claims, held-out paired controls, explicit failure cases, and release of an evaluation toolkit rather than a high-capability neural decoder.

Reproducibility statement

The anonymous release provides PairedControlBench, evaluation scripts, split definitions, mismatch-control variants, and statistical summaries. Appendix C reports the model, training, evaluation, compute, asset, and license details needed to reproduce the experiments; third-party data and pretrained weights are obtained from their original hosts under their respective terms.

References

Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 2019.

Lisa Feldman Barrett. *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt, 2017.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *International Conference on Learning Representations (ICLR)*, 2020.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5: 1097–1107, 2023. doi: 10.1038/s42256-023-00714-5.

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

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. DeWave: Discrete EEG waves encoding for brain dynamics to text translation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. Introducing MEG-MASC: a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific Data*, 10(1):862, 2023. doi: 10.1038/s41597-023-02752-5.

Youssef Hmamouche, Ismail Chihab, Lahoucine Kdouri, and Amal El Fallah Seghrouchni. A multimodal LLM for the non-invasive decoding of spoken text from brain recordings. *arXiv preprint arXiv:2409.19710*, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Wanqiu Huang, Ke Ma, Tingyu Xie, and Hongwei Wang. BrainChat: Interactive semantic information decoding from fMRI using large-scale vision-language pretrained models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889434.

Dulhan Jayalath and Oiwi Parker Jones. MEG-XL: Data-efficient brain-to-text via long-context pre-training. *arXiv preprint arXiv:2602.02494*, 2026.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are EEG-to-text models working? *arXiv preprint arXiv:2405.06459*, 2024.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

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

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

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei Sun, Josh Chartier, Mary E Dougherty, Cheoljun Liu, Gerard M Abrams, and Edward F Chang. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 2021.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 577–603, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.34. URL https://aclanthology.org/2024.blackboxnlp-1.34/.

Michael Schirner, Lia Domide, Dionysios Perdikis, Paul Triebkorn, Leon Stefanovski, Roopa Pai, Paula Prodan, Bogdan Valean, Jessica Palmer, Chloë Langford, et al. Brain simulation as a cloud service: The virtual brain on ebrains. *NeuroImage*, 251:118973, 2022.

Jerry Tang, Adam LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 2023.

Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 2021.

Qinfan Xiao, Ziyun Cui, Chi Zhang, Siqi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified eeg and meg signals. *arXiv preprint arXiv:2505.18185*, 2025.

APPENDICES

A Discussions 13
A.1 Limitations 13
A.2 Future Directions 13
A.3 Broader Implications 13

B Additional Dose-Response Figures 14

C Diagnostic Ablations for MEG-Conditioned T5 14
C.1 Additional Statistical Robustness Analyses 20
C.2 Qualitative Paired-Control Examples 21
C.3 Cross-Attention Timing in the Story-Blocked Model 22

A Discussions

A.1 Limitations

This work should be interpreted primarily as an evaluation and conditioning study, not as a claim of full neural decoding. The central empirical result is a paired target-token likelihood effect: matched neural input assigns slightly higher probability to the correct token than either zeroed or mismatched input. It does not imply that MEG alone can support fluent language generation, open-vocabulary brain-to-text decoding, or reliable recovery of intended speech.

The synthetic TVB experiments are deliberately controlled and should not be read as evidence that simulated TVB trajectories naturally encode language. The text-TVB correspondence is constructed sample-wise, so the synthetic regime tests whether paired-control evaluation can distinguish aligned from misaligned continuous side signals under known conditions. All evidence concerning real neural recordings therefore comes from the MEG-MASC experiments, not from TVB.

The real MEG effect is statistically consistent but modest. Across held-out settings, matched MEG improves target-token NLL over ZERO and SHUF, but the gains are small and often change probability mass rather than the top-ranked prediction. Thus, the result is best understood as weak but reliable token-level alignment, not as a practically useful communication interface. Characterization analyses also show larger gains for function words than for content words, so the safest interpretation is that the present MEG effect reflects timing, syntax, or lexical predictability more strongly than high-level semantic alignment alone. The story- and subject-blocked results have three-seed checks, and LOSO has two independently trained seeds per fold, but temporal, sensor, and most auxiliary ablations remain single-training-run diagnostics.

The real-data task also differs from standard autoregressive language modeling. To avoid decoder-context leakage, the T5 decoder receives target-only inputs rather than the full preceding story context. This makes the paired-control test cleaner, but it also narrows the interpretation: the model is evaluated on whether the MEG window improves target-word likelihood under controlled supervision, not on whether neural input improves ordinary long-context generation.

Our controls reduce but do not eliminate all possible confounds. ZERO tests whether the model uses any brain input at all, while the strictest SHUF preserves subject, sound, and local story time while breaking exact pairing; clustered and population-level tests address dependence across examples. The loudest remaining alternative is that post-onset MEG contributes stimulus-locked acoustic or phonological information, prosody, syntax, or lexical predictability rather than semantic content. We balance word type, length, story position, and frequency, but do not simultaneously match trial-level acoustic envelope, phoneme identity and timing, prosody, and all lexical predictors.

A.2 Future Directions

The current analyses support pairing-specific token-level modulation, but they do not yet identify a precise neural mechanism. The temporal ablations suggest that useful information is distributed across the post-onset response rather than concentrated in a single narrow late slice, and the attention analysis shows a strong late-window positional preference that is not itself pairing-specific. Future work should replace the present inference-time sensor masks with retrained or energy-matched sensor-subset models, source-localized regional analyses, frequency-band controls, and finer time-resolved ablations to determine which neural components contribute to the paired-control gain.

Several methodological extensions would also sharpen the real-data claim. A simultaneous unseen-subject and unseen-story sweep and replication on a second neural dataset would test whether the result generalizes beyond shared narratives and MEG-MASC. Trial-level acoustic-envelope, phoneme-timing, prosodic, and lexical-surprisal controls would help determine whether the MEG effect is semantic, syntactic, temporal, acoustic, or prosodic.

A.3 Broader Implications

Methodologically, the paired-control protocol may help future work evaluate auxiliary neural or physiological signals more rigorously by distinguishing useful conditioning from generic perturbation. At the same time, any improvement in neural-language modeling raises privacy, surveillance,

Table 5: Comparison to representative neural language modeling and brain-to-text work. The cited groups correspond to semantic fMRI reconstruction, MEG-MASC, invasive BCI decoding, and recent brain-conditioned LLMs.

Work	Neural data	Task	Held-out stories	Held-out subjects	Mismatch control	Deformation vs. utility
fMRI reconstruction	fMRI	semantic reconstruction	yes	limited	not central	not central
MEG-MASC	MEG	dataset / encoding benchmark	yes	yes	not central	not central
Invasive BCI decoding	ECoG / intracortical	speech or handwriting decoding	task-dependent	limited	not central	not central
Brain-conditioned LLMs	fMRI / EEG	captioning or spoken-text decoding	varies	varies	varies	not central
BrainDEC reanalysis	fMRI	target likelihood	original split	21 subjects	same-conversation	central
This work	TVB + MEG	paired target-word likelihood (target-only MEG decoding)	separate split	separate split	central	central

and overclaiming risks if weak token-level effects are misrepresented as practical mind-reading or if models generalize unevenly across subjects or populations. We therefore frame the result narrowly, release only an anonymized evaluation toolkit rather than a high-capability decoder, and emphasize held-out controls, failure cases, and subject-level caution in interpretation.

B Additional Dose-Response Figures

This appendix contains the alpha-scaling figures removed from the main text. They support the dose-response claim in Section 3.1: increasing the brain-signal scale produces smooth distributional deformation, while predictive utility depends on whether the signal is correctly paired.

The reported TVB dimensionalities denote retained synthetic node readouts, not named anatomical subsets. No empirical atlas or structural connectome is loaded: the simulator uses generic labels, generated coordinates, and sparse symmetric coupling, and a separate LM is trained for each retained width $d \in \{68, 136, 272, 544\}$. The doubling series tests how the conditioning pathway and paired-control conclusion scale with side-channel dimension; it must not be interpreted as selecting particular cortical regions.

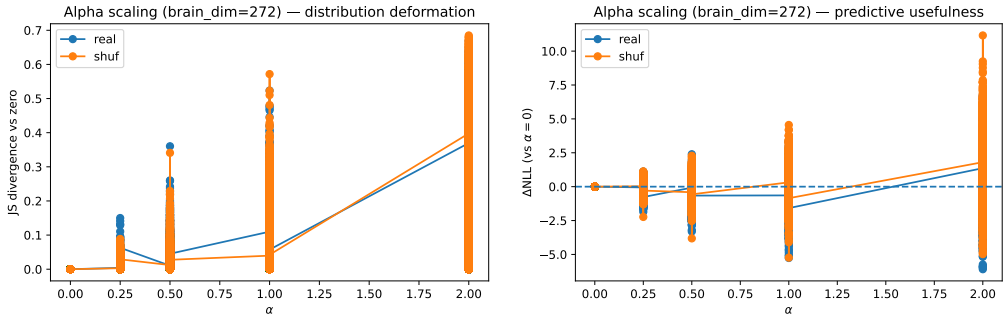

Figure 4: Alpha scaling for the 272-D brain-conditioned LM. (Left) JS divergence vs. α relative to the $\alpha = 0$ (zero-brain) baseline. (Right) ΔNLL vs. $\alpha = 0$. Real brain improves next-token prediction for moderate α (below-zero ΔNLL), while shuffled brain degrades prediction despite comparable distributional deformation.

C Diagnostic Ablations for MEG-Conditioned T5

This appendix reports additional MEG-MASC/T5 diagnostic runs used to interpret the real-neural result. We include these results as ablations rather than as final benchmark claims. All runs use target-only decoding, per-example brain normalization, and REAL/ZERO/SHUF evaluation. Table 9 and Table 10 keep the in-sample and peri-word diagnostics. Table 11, Table 12, and Table 13 then summarize stricter mismatch controls, probability/rank characterization, and coarse sensor ablations

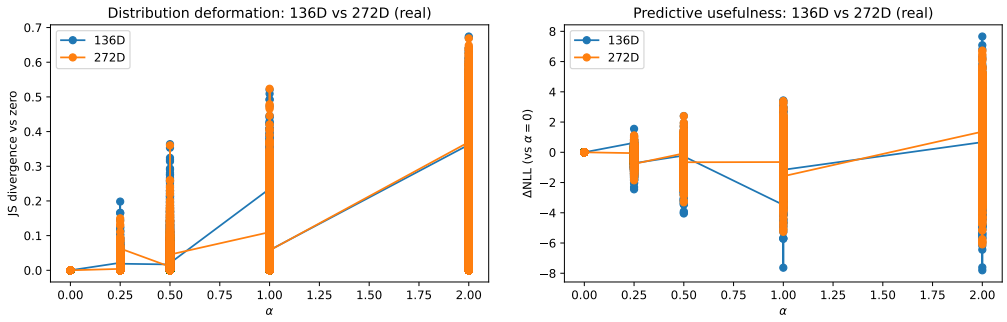

Figure 5: Direct comparison of alpha scaling between 136-D and 272-D models under the REAL condition. Overlaying curves on identical axes reveals differences in deformation strength (JS) and predictive usefulness (ΔNLL) that can be visually subtle in separate plots.

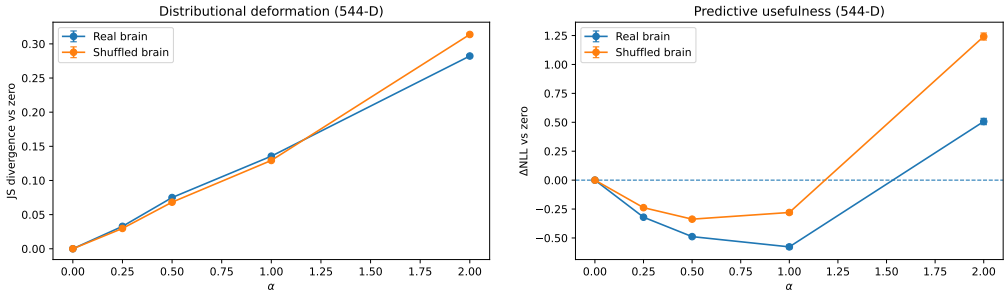

Figure 6: Alpha scaling for the 544-D brain-conditioned LM. (Left) JS divergence vs. α relative to the $\alpha = 0$ (zero-brain) baseline. (Right) ΔNLL vs. $\alpha = 0$. At $\alpha = 0.25/0.5/1.0$, both REAL and SHUF improve over ZERO, but REAL improves more ($-0.321/-0.488/-0.577$ versus $-0.238/-0.337/-0.280$); at $\alpha = 2$, both become harmful and SHUF degrades more. Thus a mismatch can act as a generic regularizer, while the REAL-SHUF comparison identifies the additional pairing-specific gain.

for the successful story-blocked model. Table 14 lists the exact model, preprocessing, optimization, and split settings used for the main MEG runs. Appendix C.1 reports multi-seed, population-level, Bayesian, and strict residual-control analyses added after the initial experiments. The training script also saves the full CLI configuration next to each checkpoint.

Table 6: Brain influence increases with brain dimensionality. JSD is computed between next-token distributions with real brain input vs. zeroed brain.

Brain dim	JSD mean	JSD median	JSD std	Top-1 agree
68	0.0657	0.0511	0.0604	0.681
136	0.0736	0.0540	0.0696	0.660
272	0.0857	0.0718	0.0710	0.620
544	0.1359	0.1162	0.1073	0.577

Table 7: Token-level sensitivity to brain conditioning in the 136-D model. We report mean absolute logit change $|\Delta\ell|$ for three token categories under real, shuffled, and Gaussian brain inputs. The two ratio columns are means of per-example ratios and therefore need not equal ratios of the displayed marginal means.

Condition	Punctuation	Stopwords	Content	Content / Stop	Content / Punct
REAL	0.891	0.731	0.815	1.115	0.961
SHUF	0.900	0.752	0.834	1.109	0.971
GAUSS	0.900	0.750	0.835	1.115	0.971

Table 8: PCA of logit deformations $\Delta\ell$ in the 272-D model under independently filtered, non-exhaustive context subsets. "Helpful" means $\Delta\mathrm{NLL}_{\mathrm{REAL}} < -0.1$; "harmful" means $\Delta\mathrm{NLL}_{\mathrm{SHUF}} > 0.1$; SHUF-on-helpful evaluates SHUF deformations on the REAL-helpful rows. The subsets can overlap and their counts are not intended to partition 5,000 examples. All conditions exhibit a low-dimensional control subspace (effective rank ≈ 8).

Condition	Samples	cumvar@1	cumvar@2	cumvar@10	eff. rank
REAL (helpful)	2791	0.3118	0.3638	0.5275	8.28
SHUF (on helpful)	2791	0.3206	0.3768	0.5443	8.13
SHUF (harmful)	2108	0.3181	0.3754	0.5473	8.29

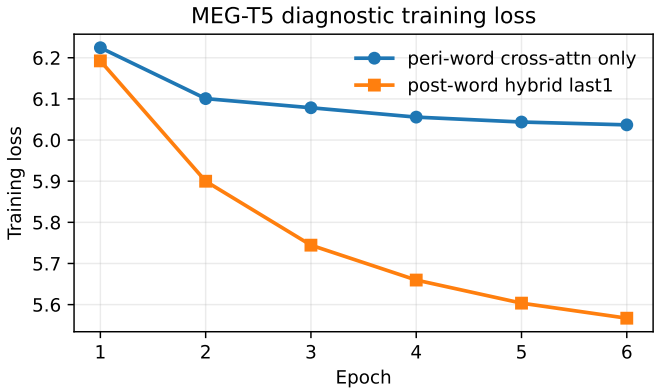

Figure 7: Training loss for two overnight MEG-MASC/T5 diagnostic runs. The post-onset hybrid model reaches lower training loss than the peri-word cross-attention-only model, but training loss alone is not an alignment metric; the REAL/ZERO/SHUF controls in Table 9 determine whether the learned dependence is useful and pairing-specific.

Table 9: MEG-MASC/T5 diagnostic ablations. Negative ΔNLL means REAL is better than the control. The cross-attention-only post-onset row is the earlier conservative baseline. The peri-word run shows that stronger use of brain input can be harmful relative to ZERO. The hybrid row is a strong in-sample result and motivates the stricter held-out analyses in the main text.

Setup	Window	Final train loss	NLL REAL	NLL ZERO	NLL SHUF	ΔR-Z	ΔR-S
Post-onset cross-attn only	0.0-0.6s	6.0947	6.8936	6.9048	6.9011	−0.0112	−0.0075
Peri-word cross-attn only	−0.1-0.8s	6.0369	7.4154	7.1102	7.4818	+0.3052	−0.0664
Post-onset hybrid last1	0.0-0.6s	5.5670	5.9504	6.0718	6.0579	−0.1214	−0.1074

Table 10: Paired-control details for the MEG-MASC/T5 diagnostic ablations. The hybrid model changes the predictive distribution much more strongly than the conservative cross-attention-only model, as reflected by lower top-1 agreement with ZERO.

Setup	ΔR-Z 95% CI	ΔR-S 95% CI	Win R-Z	Win R-S	JS REAL	Top-1 R/Z
Post-onset cross-attn only	[−0.0126, −0.0098]	[−0.0091, −0.0060]	0.5039	0.4937	0.0018	0.9997
Peri-word cross-attn only	[+0.2934, +0.3170]	[−0.0831, −0.0497]	0.4185	0.5001	0.0440	0.9517
Post-onset hybrid last1	[−0.1265, −0.1163]	[−0.1118, −0.1030]	0.5503	0.5588	0.0140	0.9146

Table 11: Stricter SHUF controls for the story-blocked post-onset hybrid MEG-MASC/T5 model. Matched MEG continues to beat every tested mismatch variant, including within-subject, within-sound, joint within-subject+sound, temporal, and spectral controls.

SHUF variant	NLL SHUF	ΔR-S	95% CI	JS SHUF	Top-1 SHUF/ZERO
Original batch SHUF	7.2663	−0.0778	[−0.0818, −0.0738]	0.00876	0.90008
Global-sample SHUF	7.2679	−0.0794	[−0.0835, −0.0753]	0.00878	0.90034
Within-subject SHUF	7.2668	−0.0783	[−0.0823, −0.0742]	0.00883	0.90006
Within-sound SHUF	7.2692	−0.0807	[−0.0847, −0.0766]	0.00873	0.90074
Within-subject+sound SHUF	7.2667	−0.0782	[−0.0822, −0.0741]	0.00878	0.89974
Within-subject+sound local-time SHUF	7.2693	−0.0808	[−0.0848, −0.0768]	0.00884	0.90170
Circular time shift	7.2551	−0.0665	[−0.0702, −0.0629]	0.00755	0.87576
Block permute (10 bins)	7.2635	−0.0749	[−0.0787, −0.0712]	0.00477	0.91044
Phase-randomized MEG	7.2774	−0.0889	[−0.0930, −0.0848]	0.00727	0.85794

Table 12: Characterization of the story-blocked post-onset hybrid MEG-MASC/T5 effect. NLL contrasts are mean-token NLL differences for each target-word example; probability and rank columns use the first supervised target position. Aggregate JSD and top-1 agreement elsewhere use the final supervised position. The main effect is largely a probability/rank shift rather than a top-1 accuracy gain, and it is strongest for higher-surprisal targets.

Slice	n	ΔR-Z	ΔR-S	Prob. up	Rank up	Notes
All target-word examples	50,000	−0.1048	−0.0778	0.5526	0.4159	Δtop-1/top-5/top-10 = −0.0009/ + 0.0060/ + 0.0081; ΔMRR = +0.0021
Last-position argmax unchanged	44,986	-	-	0.5504	0.4010	Conditional fractions within the 89.972% REAL/ZERO-agreement subset
Content words	32,787	−0.0328	−0.0521	-	-	-
Function words	17,213	−0.2420	−0.1267	-	-	-
Low surprisal	16,600	−0.0067	−0.0229	-	-	-
Mid surprisal	16,707	−0.1332	−0.0816	-	-	-
High surprisal	16,693	−0.1739	−0.1286	-	-	-

Table 13: Coarse sensor-space ablations for the story-blocked post-onset hybrid MEG-MASC/T5 model. These are inference-time zero-masking diagnostics on a frozen checkpoint, not separately retrained sensor-subset models; the independent batch-SHUF pass gives the displayed baseline ($-0.1053/-0.0794$). Left-temporal and left-hemisphere masks weaken the paired-control gain more than frontal, occipital, or matched-size random masks, but zero-masking changes input energy and therefore does not establish anatomical localization.

Ablation	Channels	ΔR-Z	ΔR-S	Loss vs. baseline (R-Z)	Loss vs. baseline (R-S)
Baseline (no ablation)	0	-0.1053	-0.0794	0.0000	0.0000
Left temporal	23	-0.0411	-0.0239	0.0642	0.0555
Right temporal	22	-0.0656	-0.0477	0.0398	0.0317
Frontal	52	-0.1027	-0.0781	0.0027	0.0013
Occipital	52	-0.1057	-0.0784	-0.0004	0.0011
Left hemisphere	105	-0.0321	-0.0193	0.0732	0.0602
Right hemisphere	103	-0.0638	-0.0477	0.0415	0.0317
Random match 1	23	-0.1007	-0.0774	0.0047	0.0021
Random match 2	23	-0.0993	-0.0727	0.0060	0.0067
Random match 3	23	-0.1019	-0.0767	0.0035	0.0027

Table 14: Exact reproducibility details for the main post-onset hybrid MEG-MASC/T5 runs. Unless an appendix ablation table states otherwise, the held-out MEG results in the paper use this configuration.

Setting	Value
Base language model	T5-small encoder-decoder
Brain encoder	2-layer Transformer encoder, 4 heads, dropout 0.1
Training objective	target-only decoder supervision / target-token NLL
Decoder linguistic context	no preceding story tokens; teacher forcing uses only the shifted target-word subword prefix
NLL aggregation	mean over supervised subwords within each target word, then mean over word examples
Distribution diagnostics	JSD and top-1 agreement at the final supervised target position; first-token probability/rank analyses are labeled separately
Optimizer	AdamW
Learning rate	5×10^{-5}
Epochs	6
Batch size	32
Precision	bfloat16
Brain normalization	per-example for the main reported MEG runs
Trainable scope	freeze T5, unfreeze decoder cross-attention in last 6 decoder blocks, unfreeze final encoder+decoder block
Maximum target length	8 tokens
Maximum brain length	120 bins for the main post-onset model
Post-onset MEG window	0.0-0.6 s after word onset
Random seeds	42 for primary runs; seeds 7/123 for story- and subject-blocked checks; seed 7 for the second complete LOSO sweep
Story-blocked multi-seed check	per-example normalization, seeds 42/7/123: mean ΔR-Z $= -0.1081 \pm 0.0033$, mean ΔR-S $= -0.0879 \pm 0.0136$
Subject-blocked multi-seed check	seeds 42/7/123: mean ΔR-Z $= -0.0634 \pm 0.0254$, mean ΔR-S $= -0.0686 \pm 0.0168$
LOSO multi-seed check	two complete 11-fold sweeps (seeds 42/7), with all 44 subject-by-contrast paired intervals below zero
Normalization sensitivity	removing per-example normalization collapses the story-blocked effect (ΔR-Z ≈ 0, ΔR-S ≈ 0, JSD ≈ 0, top-1 R/Z $= 1.0$)
Held-out evaluations	random split, story-blocked, subject-blocked, and LOSO
Story-blocked test narrative	*Easy Money* (the `easy_money_0.wav`–`easy_money_7.wav` segments); 50k examples are sampled for the reported evaluation
Primary scripts	`build_meg_word_aligned_dataset.py`, `split_sharded_meg_dataset.py`, `train_t5_brain_crossattn.py`, `eval_brain_controls.py`
Compute worker	single NVIDIA RTX A6000 (48 GB VRAM) on an internal Linux workstation, no distributed or multi-GPU training
CPU / threading	8 CPU threads for the reported GPU runs, `num_workers`=0 in the main training and evaluation commands
Approx. preprocessing time	word-aligned manifest construction and split generation are CPU-bound and take on the order of tens of minutes per split
Approx. training time	reported T5-small runs use 6 epochs on one A6000 and take roughly 40-60 minutes per run, depending on the auxiliary stream and split
Approx. evaluation time	one 50k-example held-out evaluation pass takes roughly 3-10 minutes on one A6000, clustered/bootstrap summaries are additional short post-processing steps
Approx. reported compute	the original suite is on the order of 10-20 A6000 hours; the additional second 11-fold LOSO sweep contributes roughly 7-11 single-GPU hours
Compute beyond reported tables	the broader project used additional compute for discarded peri-word, leakage-prone, and frozen-pathway diagnostic runs that are described qualitatively but not all tabulated as main results
External assets and licenses	T5-small (`google-t5/t5-small`, snapshot `df1b051`), Apache-2.0; Hugging Face Transformers, Apache-2.0; The Virtual Brain / `tvb-library>=2.7.0`, GPL-3.0-or-later; MEG-MASC / MASC-MEG OSF release (DOI `10.17605/OSF.IO/AG3KJ`), CC0 1.0 Universal; Wiki-40B English via TensorFlow Datasets v1.3.0, CC-BY-SA terms inherited from Wikipedia; MASC story texts (corpus release `577-2`) are documented by Gwilliams et al. as distributed without license or other restrictions.

Table 14 continued from the previous page

Setting	Value
Redistribution practice	The anonymous release contains our code, manifests, and instructions, but not third-party raw MEG, TVB code, pretrained T5 weights, or BrainDEC checkpoints/data; users obtain these from their original hosts.
Anonymous code release	`https://anonymous.4open.science/r/neural-cond-llms/README.md`
Toolkit contents	anonymous release includes the protocol note, REAL/ZERO/SHUF evaluator, clustered and Bayesian summaries, stricter SHUF variants, and the pinned BrainDEC adapter

C.1 Additional Statistical Robustness Analyses

Table 15 summarizes the training-seed checks added after the initial study. Standard deviations for story- and subject-blocked rows are across independently trained seeds; LOSO standard deviations are across the 11 held-out subjects within each independently trained sweep.

Table 15: Multi-seed robustness of the primary held-out MEG evaluations. Every reported contrast remains negative. For LOSO, each of the 22 subject-by-contrast paired intervals lies below zero at each seed (44/44 total).

Evaluation	Training seeds	ΔR-Z mean ± SD	ΔR-S mean ± SD	Consistency
Story-blocked	42, 7, 123	-0.1081 ± 0.0033	-0.0879 ± 0.0136	3/3 negative
Subject-blocked	42, 7, 123	-0.0634 ± 0.0254	-0.0686 ± 0.0168	3/3 negative
LOSO sweep	42	-0.08192 ± 0.01980	-0.07957 ± 0.02449	22/22 CIs < 0
LOSO sweep	7	-0.07296 ± 0.01716	-0.07313 ± 0.01537	22/22 CIs < 0

Treating the 11 seed-42 LOSO subject means as independent population units gives REAL-ZERO $t(10) = -13.72$, $p = 8.20 \times 10^{-8}$, 95% CI $[-0.09522, -0.06862]$, and REAL-SHUF $t(10) = -10.78$, $p = 7.98 \times 10^{-7}$, 95% CI $[-0.09602, -0.06311]$. Exact two-sided Wilcoxon signed-rank and sign tests give $p = 0.00098$ for each contrast (11/11 negative). Seed 7 independently gives $t(10) = -14.10$ and -15.78 ($p = 6.31 \times 10^{-8}$ and 2.15×10^{-8}).

Table 16 reports subject-level JZS Bayes factors for diagnostics previously described informally as null. The evidence supports a narrower statement: these streams fail the paired-control criterion. In particular, the early window has a detectable but practically negligible REAL-ZERO shift and no pairing-specific REAL-SHUF gain; no-normalization REAL-ZERO is numerically negligible but Bayes-factor inconclusive.

Table 16: Subject-level Bayesian checks for diagnostic failures ($N = 11$ subjects; two-sided JZS test with Cauchy scale $r = 1/\sqrt{2}$). $BF_{01} > 1$ favors no effect; $BF_{01} < 1$ favors a nonzero effect.

Diagnostic	Contrast	Subject mean ± SD	BF_{01}	Interpretation
IID Gaussian	R-S	$1.49 \times 10^{-5} \pm 2.95 \times 10^{-4}$	3.32	moderate null evidence
Global Gaussian	R-S	$-3.41 \times 10^{-5} \pm 1.75 \times 10^{-4}$	2.82	suggestive null evidence
Early 0–300 ms	R-S	$-5.00 \times 10^{-6} \pm 6.02 \times 10^{-5}$	3.25	moderate null evidence
Early 0–300 ms	R-Z	$-7.30 \times 10^{-5} \pm 5.12 \times 10^{-5}$	0.020	tiny nonzero shift
No normalization	R-S	$-7.11 \times 10^{-6} \pm 2.86 \times 10^{-5}$	2.52	suggestive null evidence
No normalization	R-Z	$-2.28 \times 10^{-8} \pm 3.19 \times 10^{-8}$	0.485	inconclusive

The frozen-BERT residual illustrates sensitivity to mismatch construction. Keeping the 50,000 examples and BERT side channel fixed, each SHUF donor is drawn from the same subject and sound within ± 32 word positions. All examples have eligible nonidentity donors with no fallback. As Table 17 shows, the residual direction is negative but tiny and only one of three donor-seed intervals excludes zero; it is therefore not positive evidence that MEG adds robust pairing-specific information beyond BERT.

Table 17: Frozen-BERT+MEG residual under nuisance-preserving local SHUF. REAL-ZERO is fixed at -0.02739 (95% CI $[-0.04151, -0.01328]$); only the MEG donor assignment varies.

Donor seed	ΔR-S	95% CI	CI excludes zero
42	-0.000236	$[-0.000517, +0.000045]$	no
7	-0.000237	$[-0.000521, +0.000047]$	no
123	-0.000336	$[-0.000618, -0.000053]$	yes

C.2 Qualitative Paired-Control Examples

To complement the aggregate control tables, we also prepare a qualitative appendix of individual story-blocked examples. Below we show ten curated real-MEG examples from the story-blocked split, chosen from a larger single-token content-word pool. For each example we report the text context, the target token, and the first-target probability/rank under ZERO, REAL, and SHUF, together with the absolute ZERO$\rightarrow$REAL target-probability gain. These rows are illustrative only, the main evidence remains the aggregate paired-control tables. This presentation is intentionally token-level rather than free-form generation, because the question in this paper is whether matched MEG slightly improves the probability of the correct target, not whether it produces superficially different continuations.

Table 18: Ten curated story-blocked real-MEG examples. Each control cell is shown as probability / rank at the first supervised target position; Δp is the absolute REAL-minus-ZERO target-probability gain.

Context	Target	ZERO p/r	REAL p/r	SHUF p/r	Δp
... ugh I let the G word slip Was this crazy treasure hunt Or maybe this guy was just a wacko who wanted to take me out on a	dark	0.0012 / 125	0.0016 / 86	0.0011 / 147	+0.0004
... e and feet I could see same design from little journal and mound carved into rock I in shock blood pounded in head and I	stood	0.0012 / 127	0.0015 / 89	0.0012 / 124	+0.0004
... n paid mound anybody best door start sort management Deep Storage Fort Knox it read Fine Troy Ounces Book Value I had to	read	0.0020 / 62	0.0024 / 49	0.0016 / 83	+0.0004
... ll part of what I sell He drained last of his cup tossed bills on table I I I I through my life taking kind chance I put	away	0.0016 / 85	0.0019 / 74	0.0012 / 114	+0.0003
... dred ounces start Four Hundred I pulled out my palm phone looked up number lot very small part of what I sell He drained	last	0.0021 / 56	0.0024 / 46	0.0020 / 63	+0.0003

Table 19: Ten curated story-blocked real-MEG examples (continued).

Context	Target	ZERO p/r	REAL p/r	SHUF p/r	Δp
... your um business Fine I waived down waitress ordered cup decaf slice pie I did noticed just how tired looked around the	eyes	0.0024 / 47	0.0027 / 37	0.0021 / 57	+0.0003
... p Storage Fort Knox I sucked in a whole lot of air Anywhere Anywhere he walked around the and came back wrestling with a	big	0.0011 / 143	0.0014 / 102	0.0008 / 183	+0.0003
... hill on You could step three feet to the right and do the same thing and you would end up buried alive in a nearby rock	face	0.0033 / 31	0.0035 / 29	0.0012 / 108	+0.0002
... was a thirty foot tall black face with staring blue and white eyes and it wore a ridiculous little hat For some reason I	thought	0.0018 / 72	0.0020 / 69	0.0013 / 106	+0.0002
... said I I I expected him pull out a gadget blinking lights and was wondering which way duck Trick gesture like sleight of	hand	0.0024 / 49	0.0024 / 45	0.0018 / 70	+0.0001

C.3 Cross-Attention Timing in the Story-Blocked Model

To inspect where the successful story-blocked post-onset model reads the MEG window, we export decoder cross-attention for the final prediction step over 5,000 held-out examples and average over decoder layers and heads. Figure 8 shows a pronounced late-window preference. The last quarter of the post-onset window (450-600 ms after word onset) carries 50.9% of the mean attention mass, compared with 15.5%, 13.4%, and 20.2% for the first three quarters. The mean peak occurs at bin 107 out of 120, corresponding to approximately 535 ms after word onset, and the 10th, median, and 90th percentiles of the peak bin are all 107. We do not interpret this as localization of useful physiology at 535 ms. The temporal brain encoder is bidirectional, so a late encoder state can integrate the entire window; the REAL and SHUF timing profiles are nearly identical; and the separately trained 450-600 ms model has no paired utility. The peak is therefore best treated as a non-pairing-specific sequence-position preference over contextualized encoder states, while the temporal ablations indicate that useful information is distributed across the broader 0.0-0.6 s response.

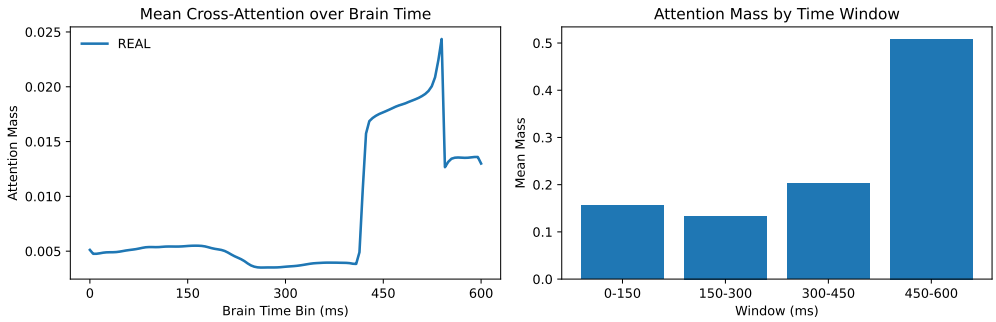

Figure 8: Mean decoder cross-attention over the post-onset MEG window for the story-blocked post-onset hybrid model, averaged across 5,000 held-out examples. Left: mean attention curve for the final decoder prediction step. Right: attention mass aggregated over four equal-width time windows. The model concentrates roughly half of its attention mass in the final quarter of the 0.0-0.6 s post-onset window, with a consistent peak around 535 ms after onset. Because states are bidirectionally contextualized and the same positional preference appears under SHUF, this is an architectural timing diagnostic rather than evidence of neural localization.

These ablations support three practical conclusions. First, loss curves are necessary but insufficient: the peri-word run optimizes, yet REAL is substantially worse than ZERO. Second, temporal structure matters within the successful post-onset window itself: the early half fails the paired criterion, the late half is only weakly useful, and the narrow late slice fails, so the main effect is distributed across the broader 0.0-0.6 s response rather than reducible to a single late peak. Third, increasing trainable capacity by unfreezing the last T5 block can dramatically improve controls, and the new held-out tables show that this gain is not purely in-sample. Random held-out, story-blocked, subject-blocked, and LOSO evaluation all preserve a clear REAL advantage over ZERO and SHUF, while the peri-word variant remains a cautionary example that stronger brain dependence can still be harmful.