DISAMBIGUATING SEMANTIC UNITS FOR EEG-TO-TEXT RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG-to-text decoding aims to recover linguistic semantics from non-invasive neural signals. However, reliable decoding requires not only fluent generation, but also identifiable semantic evidence and its compositional structure. We reveal a previously underexplored failure mode in fine-grained EEG semantic decoding: semantic ambiguity arises at two distinct scales. Within a sample, shared sentence context dominates query-specific information, causing different semantic queries to converge toward similar representations; across samples, highly central semantic candidates repeatedly attract unrelated EEG queries, inducing systematic overmatching. These observations suggest that reliable semantic decoding requires jointly resolving context-induced convergence and hubness before composing the recovered evidence into language. Based on this finding, we propose **Neural conteXtual Unified Semantics (NEXUS)**, a unified framework that disentangles shared contextual semantics from query-specific evidence, calibrates biased semantic matching from both candidate and query perspectives, and recomposes the recovered semantic units using delexicalized structural priors. Experiments on heterogeneous EEG datasets spanning natural reading and imagined across languages speech demonstrate that NEXUS substantially reduces within-sample representation convergence and cross-sample overmatching, leading to more discriminative semantic-unit identification and more faithful EEG-to-text reconstruction.

1 INTRODUCTION

Recent advances in deep learning and generative modeling have extended brain–computer interface (BCI) research from conventional recognition tasks toward decoding higher-level cognitive content, including language and visual information (Liu et al., 2024; Wang et al., 2024; Rathore & Bhavsar, 2026; Zheng et al., 2026b; Mishra et al., 2025; Zhou et al., 2026; Wu et al., 2025). Among these directions, EEG-to-text decoding is particularly challenging because linguistic semantics are abstract and compositional, whereas EEG signals are noisy and highly variable. Recent EEG-to-text methods can generate fluent sentences, but fluency does not ensure faithful neural grounding. Reliable EEG-to-text decoding therefore requires not only recovering global linguistic meaning, but also identifying which fine-grained semantic evidence is supported by EEG and how such evidence should be composed into coherent language.

Existing approaches have progressively moved from global semantic alignment toward fine-grained semantic grounding. Earlier methods align sentence-level EEG representations with textual semantic spaces to condition language generation, but provide limited access to individual semantic components. Recent methods instead recover explicit words, concepts, or variable-size semantic-unit sets from EEG (Zheng et al., 2026a; Murhekar et al., 2026; Li et al., 2026), making the neural evidence underlying generation more identifiable. However, recovering semantic units raises a fundamental yet underexplored question: can multiple fine-grained semantic units be reliably distinguished from one another in the shared EEG–language space?

Our analysis reveals that semantic-unit decoding suffers from dual-scale semantic ambiguity (Fig. 1). Within a sample, multiple semantic queries share the same neural context and tend to converge toward a few dominant semantic directions, causing different queries to repeatedly retrieve similar semantic units. This suggests that shared contextual information can overwhelm query-specific variation, leading to context-induced representation convergence. Across samples, retrieval errors

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.

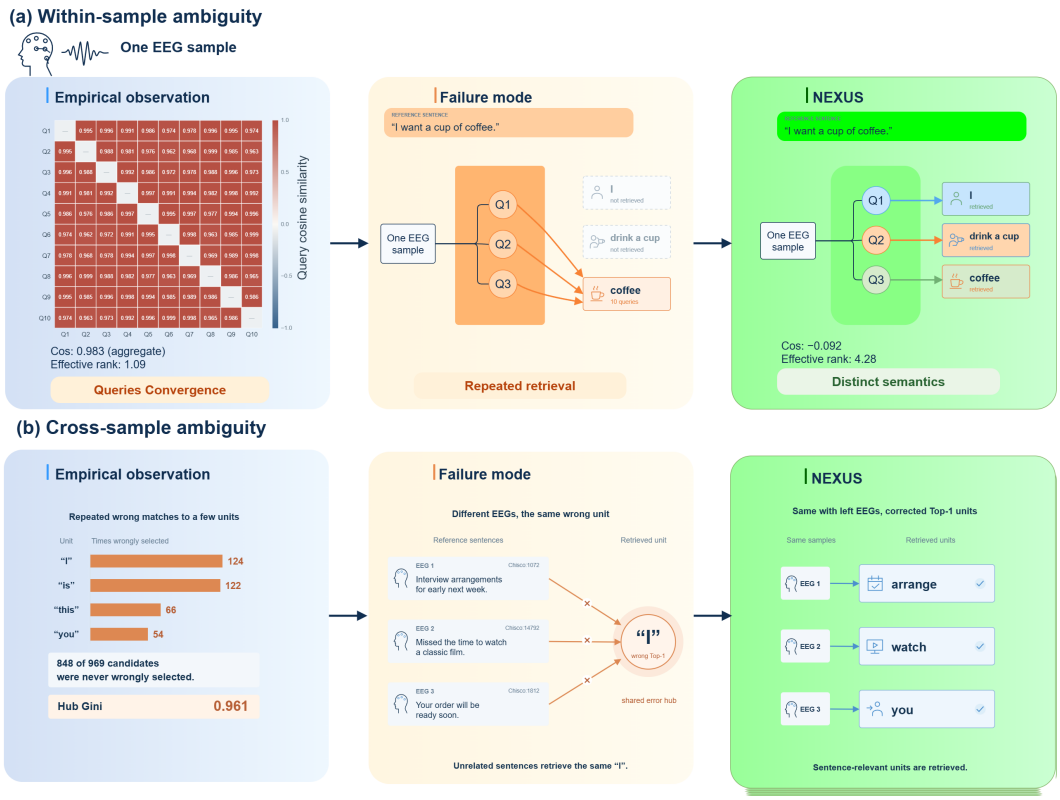

Figure 1: Motivation for dual-scale semantic disambiguation in NEXUS. (a) **Within-sample ambiguity.** Semantic queries from the same EEG sample can converge to highly similar representations, causing repeated retrieval of the same semantic unit and suppressing query-specific evidence. NEXUS separates shared context from query-specific residual semantics, promoting distinct and complementary Unit retrieval. (b) **Cross-sample ambiguity.** Across EEG samples, a small number of highly central semantic candidates can become error hubs and attract unrelated queries. NEXUS explicitly calibrates candidate-side background affinity and query-side centrality, reducing hub-dominated overmatching and recovering more sample-relevant semantic units.

are disproportionately concentrated on a small subset of highly central semantic candidates, which repeatedly attract otherwise unrelated EEG queries and induce hubness-driven overmatching. These observations suggest that fine-grained EEG semantic decoding is limited not merely by weak cross-modal alignment, but by systematic degeneration of semantic discriminability at both intra- and inter-sample scales.

Resolving semantic ambiguity determines what semantic evidence can be reliably recovered, but faithful reconstruction additionally requires determining how that evidence should be composed. Existing semantic-unit methods typically recover units independently, leaving their compositional organization under-specified. We therefore formulate reliable EEG-to-text reconstruction through two complementary objectives: Semantic Content Decoding, which resolves within- and cross-sample ambiguity, and Compositional Structure Recovery, which organizes the recovered semantic evidence into coherent linguistic structures.

Based on this formulation, we propose **Neural conteXtual Unified Semantics (NEXUS)**, built around a simple principle: semantic evidence should be disambiguated before being composed into language. NEXUS separates shared contextual semantics from query-specific evidence to improve within-sample semantic differentiation, and calibrates semantic matching from both candidate and query perspectives to mitigate cross-sample ambiguity. The resulting semantic units are then combined with delexicalized structural priors derived from sentence-level retrieval to guide text reconstruction. In this way, NEXUS jointly addresses what semantic content is present and how that content should be composed, while explicitly resolving ambiguity at two semantic scales.

We evaluate NEXUS across three EEG datasets(Hollenstein et al., 2020; 2018; Zhang et al., 2024) spanning different languages and language-related cognitive paradigms, from natural reading to imagined speech. Experiments show that NEXUS reduces within-sample representation convergence and cross-sample ambiguity, resulting in more discriminative semantic-unit identification and improved EEG-to-text reconstruction.

Our main contributions are:

- We identify dual-scale semantic ambiguity in fine-grained EEG semantic decoding, characterized by within-sample context-induced representation convergence and cross-sample hubness-induced overmatching.
- We propose NEXUS, a unified framework for dual-scale semantic disambiguation, which separates shared context from query-specific semantics and calibrates biased semantic matching from both candidate and query perspectives.
- We integrate disambiguated semantic evidence with compositional structure recovery, enabling more reliable semantic identification and EEG-to-text reconstruction across heterogeneous neural language decoding settings.

2 Related Work

2.1 EEG-to-Text Decoding and Semantic Alignment

Early EEG-to-text methods were largely generation-centric, focusing on learning EEG representations that can be effectively connected to language models. EEG2TEXT improves open-vocabulary generation through EEG pretraining and multi-view modeling (Liu et al., 2024), while BELT-2 strengthens EEG-language alignment through BPE-level supervision, multi-task learning, and prefix tuning (Zhou et al., 2024).Recent works have moved toward semantically grounded generation. GLIM(Liu et al., 2025) aligns EEG representations with the latent semantic space of language models to alleviate posterior collapse, while SemKey(Wang et al., 2026) introduces explicit semantic guidance to strengthen EEG-conditioned generation. More recent work further shifts toward fine-grained semantic grounding. Word-level alignment directly maps EEG signals into lexical semantic spaces (Zheng et al., 2026a). SENSE retrieves keyword- and concept-level semantic evidence (Murhekar et al., 2026), and SID models sentence semantics as a variable-size set of semantic units for subsequent reconstruction (Li et al., 2026). These studies show that semantic units can be recovered from EEG, but largely overlook whether the recovered units remain mutually distinguishable. Moreover, their compositional relations are typically under-modeled. NEXUS therefore focuses on both semantic-unit discriminability and compositional structure recovery.

2.2 Representation Degeneration in Semantic Alignment

Representation degeneration can impair semantic discriminability in shared embedding spaces. Prior work regularizes embedding geometry to improve representation diversity, including Alignment-Uniformity (Wang & Isola, 2020), Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2022),while retrieval methods such as QB-Norm (Bogolin et al., 2022) and NeighborRetr (Lin et al.) mitigate hubness and neighborhood imbalance. Fine-grained EEG semantic decoding exhibits semantic ambiguity at two scales. Within a sample, shared neural context drives query representations to converge, suppressing query-specific semantics and collapsing their representations; across samples, highly central semantic candidates repeatedly attract unrelated EEG queries, causing systematic overmatching. We term these two effects dual-scale semantic ambiguity: context-induced convergence and centrality-induced overmatching. NEXUS addresses the two through shared-context decomposition and centrality-aware calibration, respectively, linking representation-space geometry directly to reliable fine-grained EEG semantic decoding.

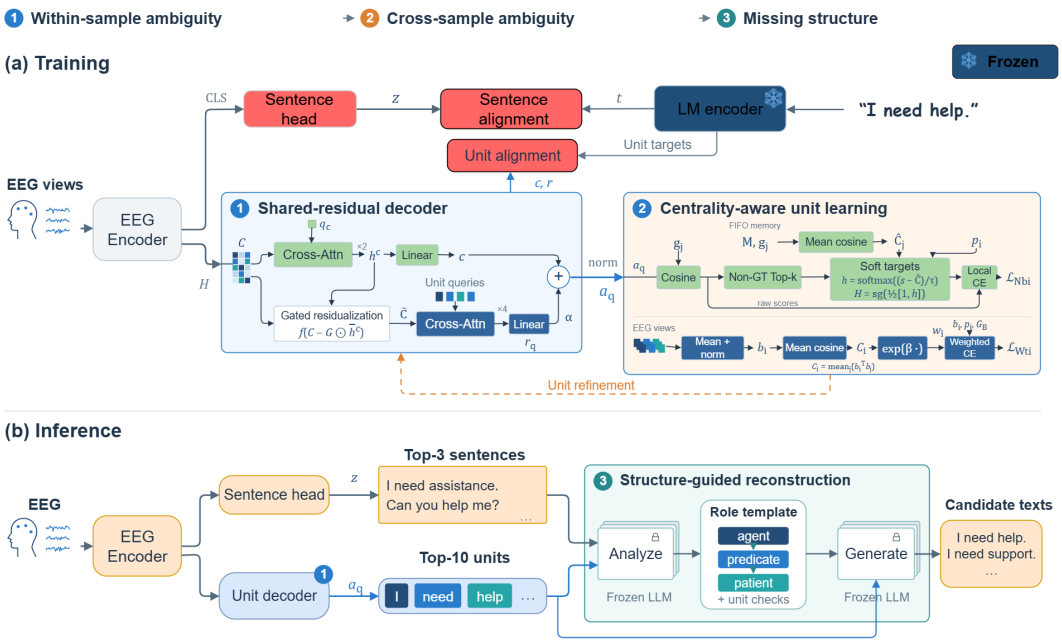

Figure 2: Overview of NEXUS. (a) Training addresses dual-scale semantic ambiguity: the shared-residual decoder separates shared context from query-specific semantics for within-sample disambiguation, while centrality-aware unit learning calibrates candidate and query centrality for cross-sample disambiguation. Sentence- and Unit-level representations are aligned to a frozen semantic space, and the Unit decoder is further refined by the centrality-aware objectives. (b) At inference, sentence-level retrieval provides structural evidence and Unit-level retrieval provides fine-grained semantic content. A frozen LLM abstracts retrieved sentences into role-based structural templates and combines them with recovered Units for structure-guided reconstruction.

3 METHOD

3.1 PROBLEM FORMULATION

We formulate EEG-to-text decoding as a dual-granularity cross-modal semantic retrieval and structure-guided text reconstruction task. Given an EEG signal $\boldsymbol{X}_i$, the objective is to recover global sentence-level semantics and fine-grained semantic units, and to reconstruct the corresponding sentence y_i using the retrieved semantic content and compositional structure.

Given a dataset of N EEG-text samples,

$$\mathbb{D} = \{(\boldsymbol{X}_i, y_i, \mathbb{U}_i)\}_{i=1}^{N}, \qquad \mathbb{U}_i = \{u_{im}\}_{m=1}^{M_i}, \tag{1}$$

where $\boldsymbol{X}_i$ is the i-th EEG sample, y_i is its corresponding sentence, and $\mathbb{U}_i$ contains M_i semantic units.

For each EEG input, NEXUS retrieves global sentence-level semantics and fine-grained Unit evidence, and reconstructs the target text as

$$\mathbb{C}_i^S = r_S(\boldsymbol{X}_i), \qquad \widehat{\mathbb{U}}_i = r_U(\boldsymbol{X}_i), \qquad \hat{y}_i = g\Big(a(\mathbb{C}_i^S), \widehat{\mathbb{U}}_i\Big). \tag{2}$$

Here, r_S and r_U denote sentence- and Unit-level retrieval, respectively; a denotes transform retrieved sentences into delexicalized structural templates, and g performs text reconstruction.

3.2 NEXUS OVERVIEW

As illustrated in Fig. 2, NEXUS jointly models global sentence semantics and fine-grained Unit semantics in a shared semantic space. Given an EEG input $\boldsymbol{X}_i$, the EEG encoder f_{EEG} produces a

global representation together with a sequence of contextualized EEG tokens:

$$\left[\boldsymbol{h}_i^{\mathrm{cls}}, \boldsymbol{H}_i\right] = f_{\mathrm{EEG}}(\boldsymbol{X}_i). \tag{3}$$

Here, $\boldsymbol{h}_i^{\mathrm{cls}}$ captures global sentence-level semantics, whereas $\boldsymbol{H}_i$ provides token-level features for fine-grained Unit decoding. On the text side, a frozen text encoder f_T defines a shared semantic reference space. The sentence-level branch aligns the global EEG representation with the corresponding sentence embedding:

$$\boldsymbol{z}_i^S = \frac{p_S(\boldsymbol{h}_i^{\mathrm{cls}})}{\|p_S(\boldsymbol{h}_i^{\mathrm{cls}})\|_2}, \qquad \boldsymbol{t}_i^S = \frac{f_T(y_i)}{\|f_T(y_i)\|_2}, \tag{4}$$

and optimizes a sentence-level contrastive objective $\mathcal{L}_{\mathrm{sent}}$ to preserve global semantic consistency.

Building on this global alignment, NEXUS addresses semantic ambiguity at two complementary scales. Sec. 3.3 mitigates within-sample query convergence through shared-residual decomposition, balanced Query-Unit assignment, and discriminative Unit learning. Sec. 3.4 further reduces cross-sample overmatching by calibrating candidate-side matching bias and query-side semantic centrality. During inference, sentence-level retrieval provides global structural priors, while Unit-level queries recover fine-grained semantic evidence. These signals are finally combined through delexicalized structural templates for text reconstruction, as described in Sec. 3.5.Training proceeds progressively from sentence alignment to Unit learning and hubness calibration; the full schedule is provided in the Appendix.

3.3 Within-Sample Disambiguation

Sentence-level alignment preserves global semantics but cannot explicitly distinguish multiple fine-grained semantic units within the same sentence. We therefore perform within-sample disambiguation through shared-residual decomposition, balanced Query-Unit assignment, and discriminative Unit learning.

Shared-residual decomposition. For each semantic unit $u_{im} \in \mathbb{U}_i$, we mark its corresponding span in the complete sentence y_i to obtain $\widetilde{y}_{im}$, and encode this marked sentence with the frozen text encoder. Since Unit targets from the same sentence share substantial contextual information, we decompose them into a within-sentence shared component and Unit-specific residuals:

$$\boldsymbol{t}_{im}^U = \frac{f_T(\widetilde{y}_{im})}{\|f_T(\widetilde{y}_{im})\|_2}, \qquad \boldsymbol{\mu}_i^U = \frac{1}{M_i}\sum_{m=1}^{M_i} \boldsymbol{t}_{im}^U, \qquad \boldsymbol{v}_{im}^U = \boldsymbol{t}_{im}^U - \boldsymbol{\mu}_i^U. \tag{5}$$

Here, $\boldsymbol{\mu}_i^U$ captures the contextual semantics shared by Units in the same sentence, while $\boldsymbol{v}_{im}^U$ characterizes the Unit-specific variation of the m-th semantic unit relative to this shared component. This decomposition separates shared context from Unit-specific variation without assuming orthogonality or complete disentanglement.

On the EEG side, we adopt a corresponding shared-residual formulation. A shared branch extracts the common semantic component from the EEG token representations $\boldsymbol{H}_i$, while Q learnable residual queries recover query-specific semantics:

$$\boldsymbol{c}_i^U = f_c(\boldsymbol{H}_i), \qquad \boldsymbol{r}_{iq}^U = f_r\left(\boldsymbol{q}_q, \boldsymbol{H}_i, \boldsymbol{c}_i^U\right), \qquad q = 1, \ldots, Q. \tag{6}$$

Here, f_r first adaptively suppresses information dominated by the shared component through a learned gating mechanism and then extracts query-specific residual semantics. This reduces the tendency of multiple Unit queries to converge toward similar contextual directions.

Balanced Query-Unit assignment. Since the Q Unit queries have no predefined one-to-one correspondence with the M_i ground-truth Units, we use **capacity-constrained Hungarian matching** to establish the Query-Unit assignment. Let $A_{iqm} = 1$ indicate that query q is assigned to Unit m. The matching is defined as

$$\begin{aligned} \boldsymbol{A}_i^* &= \arg\max_{\boldsymbol{A}_i} \sum_{q=1}^{Q}\sum_{m=1}^{M_i} A_{iqm}\,\mathrm{sim}\left(\boldsymbol{r}_{iq}^U, \boldsymbol{v}_{im}^U\right), \\ \text{s.t.} \quad & A_{iqm} \in \{0,1\}, \quad \sum_{m=1}^{M_i} A_{iqm} = 1\ \forall q, \quad \sum_{q=1}^{Q} A_{iqm} \in \left\{\left\lfloor \frac{Q}{M_i} \right\rfloor, \left\lceil \frac{Q}{M_i} \right\rceil\right\}\ \forall m. \end{aligned} \tag{7}$$

The first constraint assigns each query to exactly one Unit, while the capacity constraint ensures that the numbers of queries assigned to any two Units differ by at most one. When $Q \geq M_i$, every Unit therefore receives at least one query, preventing supervision from concentrating on only a few highly similar semantic units.

Discriminative Unit objective. Based on the resulting assignment, we align the shared components, the complete Unit representations $\boldsymbol{c}_i^U + \boldsymbol{r}_{iq}^U$, and the Unit-specific residuals between EEG and text, and collectively denote these terms by $\mathcal{L}_{\text{align}}$. We further introduce an intra-sample margin-ranking objective $\mathcal{L}_{\text{intra}}$, which encourages each query residual to be more similar to its matched Unit than to other Units in the same sentence to promoting intra-sample disambiguation.

The final Unit-level training objective is

$$\mathcal{L}_{\text{unit}} = \mathcal{L}_{\text{align}} + \lambda_{\text{intra}}\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{aux}}, \tag{8}$$

where $\mathcal{L}_{\text{aux}}$ collects auxiliary prototype, geometry, and consistency regularizers. For retrieval, we form the normalized Unit representation $\boldsymbol{z}_{iq}^U \propto \boldsymbol{c}_i^U + \alpha \boldsymbol{r}_{iq}^U$, where α controls the contribution of query-specific residual semantics. Details of gated residualization, query decoding, capacity allocation, Hungarian matching, auxiliary objectives, and implementation hyperparameters are provided in the Appendix.

3.4 Cross-Sample Disambiguation

After intra-sample disambiguation, Unit queries may still suffer from cross-sample hubness, where a small number of highly central semantic candidates exhibit high similarity to many unrelated EEG queries, leading to overmatching. NEXUS addresses this issue through a neighbor-adjusting term (NBI) on the candidate side and a centrality-weighted term (WTI) on the query side.

For an EEG query $\boldsymbol{z}_{iq}^U$ and the j-th semantic-key prototype $\boldsymbol{k}_j$, we define the matching probability as

$$p_{iqj} = \frac{\exp\left(\text{sim}(\boldsymbol{z}_{iq}^U, \boldsymbol{k}_j)/\tau_U\right)}{\sum_l \exp\left(\text{sim}(\boldsymbol{z}_{iq}^U, \boldsymbol{k}_l)/\tau_U\right)}. \tag{9}$$

Candidate-side hubness calibration. NBI uses a memory bank of historical EEG Unit queries to estimate the background matching tendency of semantic candidates and calibrates the soft supervision over non-GT neighbors accordingly. Let $\widetilde{y}_{iqj}$ denote the neighborhood target after background affinity calibration. The NBI objective is

$$\mathcal{L}_{\text{NBI}} = -\frac{1}{|\mathbb{Q}|} \sum_{(i,q)\in\mathbb{Q}} \sum_j \widetilde{y}_{iqj} \log p_{iqj}. \tag{10}$$

This objective reduces the matching advantage of highly central candidates caused by their generally high similarity in the shared semantic space.

Query-side centrality calibration. WTI estimates the intra-modal centrality of each EEG query and converts it into a weight w_{iq}, which reweights the negative log-likelihood of the correct semantic key:

$$\mathcal{L}_{\text{WTI}} = -\frac{1}{|\mathbb{Q}|} \sum_{(i,q)\in\mathbb{Q}} w_{iq} \log p_{iqj_{iq}^+}, \tag{11}$$

where j_{iq}^+ denotes the correct semantic key for query (i, q). This weighting strengthens supervision for queries in highly central regions that are more susceptible to hub candidates.

The overall hubness calibration objective is

$$\mathcal{L}_{\text{hub}} = \lambda_{\text{NBI}}\mathcal{L}_{\text{NBI}} + \lambda_{\text{WTI}}\mathcal{L}_{\text{WTI}}. \tag{12}$$

Details of memory-bank updates, neighborhood-target construction, query-centrality estimation, and hyperparameter settings are provided in the Appendix.

Table 1: Overall EEG-to-text reconstruction performance under target specificity, content faithfulness, structural fidelity, and semantic similarity.

Dataset	Method	RSRA_{50} ↑	Ground. ↑	Struct. ↑	BERT-F1 ↑
ZuCo1	GLIM-T5	51.49	63.13	**51.85**	0.842803
	SID	50.39	55.12	46.78	0.831199
	NEXUS (ours)	**53.03**	**80.24**	50.54	**0.847625**
ZuCo2	GLIM-T5	53.30	63.73	50.28	0.841147
	SID	49.37	56.59	46.53	0.831311
	NEXUS (ours)	**53.33**	**73.59**	**52.41**	**0.844323**
Chisco	SID	50.02	19.31	29.76	0.524393
	NEXUS (ours)	**56.18**	**44.90**	**55.95**	**0.588365**

3.5 Compositional Structure Recovery

During inference, the sentence-level representation retrieves semantically related sentences from a frozen sentence gallery to provide global structural priors. In parallel, Unit-level queries are matched against semantic-key prototypes, and candidate Units are selected according to the maximum similarity across queries, providing fine-grained content evidence.

A frozen large language model then first combines the retrieved sentences and Unit candidates to validate the semantic units, identify their semantic and grammatical roles, and abstract the retrieved sentences into delexicalized structural templates. In the final reconstruction stage, only the validated Unit evidence, role information, and structural templates are used to generate candidate texts, without exposing the surface content of the retrieved sentences. This separates structural priors from specific semantic content during generation. Detailed retrieval sizes, prompt templates, language-model configuration, and generation settings are provided in the Appendix.

4 EXPERIMENTS

We evaluate NEXUS from two complementary perspectives: overall reconstruction quality and the mechanisms behind its performance. Specifically, we examine whether NEXUS improves target specificity, content faithfulness, and compositional fidelity; whether shared–residual modeling with capacity-constrained matching mitigates within-sample query convergence; whether NBI and WTI reduce hubness-induced cross-sample overmatching; and whether delexicalized structural priors improve compositional reconstruction without relying on retrieved surface text.

4.1 Experimental Setup and Evaluation Protocol

Datasets and baselines. We evaluate NEXUS on ZuCo 1.0, ZuCo 2.0, and Chisco, covering natural reading and imagined speech across different languages. On ZuCo 1.0 and ZuCo 2.0, we compare against GLIM-T5 and SID; on Chisco, we compare against SID under the same evaluation protocol.

Evaluation metrics. BLEU measures surface n-gram overlap, which does not reflect EEG grounding and can penalize valid paraphrases or partial but faithful reconstructions. We therefore evaluate three complementary properties: target specificity, content faithfulness, and structural fidelity. We additionally report BERT-F1, the F1 score of BERTScore, as a complementary measure of semantic similarity between the prediction and reference.

Calibrated RSRA_{50} measures whether a prediction matches its correct target beyond its background matching level:

$$\text{RSRA}_{50} = 50 + 50\frac{S - B}{100 - B}, \tag{13}$$

where S is the target-specific score and B its background baseline; 50 therefore corresponds to background-level matching.

Groundedness measures unsupported generated content:

$$\text{Groundedness}_i = 100\frac{I_R(i)}{I_R(i) + U_i}, \tag{14}$$

where $I_R(i)$ is the IDF-weighted reference content mass and U_i is the unsupported IDF-weighted content mass of the prediction.

Structural Alignment measures agreement between coarse structural sequences using an $F_{0.5}$ score:

$$\text{SA}_i = 100\frac{1.25\,P_s(i)R_s(i)}{0.25\,P_s(i) + R_s(i)}, \tag{15}$$

where P_s and R_s are obtained from the longest common subsequence between the prediction and reference structures. Full metric definitions, text normalization, gallery construction, and structural mappings are provided in the Appendix.

4.2 Overall Reconstruction Performance

Table 1 examines the central problem addressed by NEXUS: when EEG-derived semantics are ambiguous, plausible generation does not necessarily imply faithful EEG reconstruction. NEXUS addresses this ambiguity before structure-guided reconstruction, and the overall results show that its gains mainly lie in more reliable target-specific semantic recovery rather than surface similarity.

On ZuCo1, this distinction is most evident in content reliability. NEXUS achieves higher RSRA_{50}, Groundedness, and BERT-F1 than both baselines, indicating that its reconstructions are not only semantically closer to the references but also contain more target-specific evidence and less unsupported content. Although GLIM-T5 has slightly higher Structural Alignment, NEXUS mainly improves EEG-grounded semantic reliability while preserving comparable structure.

ZuCo2 further clarifies this difference. NEXUS is essentially tied with GLIM-T5 in RSRA_{50} (53.33 vs. 53.30), yet achieves higher Groundedness, Structural Alignment, and BERT-F1. With comparable target specificity, NEXUS reduces unsupported semantics and better organizes the recovered content. This pattern is consistent with the design principle of NEXUS: disambiguating semantic evidence before recovering its compositional structure.

The advantage becomes more pronounced on Chisco under the imagined-speech setting, where NEXUS outperforms SID across all four metrics. Since Chisco differs substantially from the natural-reading paradigm of ZuCo, this result suggests that the semantic ambiguity and structure-recovery problems targeted by NEXUS extend across different neural language settings. Overall, the advantage of NEXUS lies not merely in generating more similar text, but in recovering EEG-grounded semantics more reliably and organizing them into coherent compositional structure.

4.3 Ablation and Mechanism Analysis

Table 2: Ablation and diagnostic results for dual-scale semantic disambiguation. (a) Within-sample residual supervision; (b) cross-sample hubness calibration.

(a) Within-sample disambiguation				
Setting	Query Cos. ↓	Eff. Rank ↑	Unit MRR ↑	R@10 ↑
w/o residual losses	0.9834	1.0886	0.0855	0.1617
Full Unit objective	**-0.0923**	**4.2786**	**0.0922**	**0.1827**
(b) Cross-sample disambiguation				
Setting	Unit MRR ↑	R@1 ↑	CR10 (%) ↓	N_{eff} ↑
Before calibration	0.1341	0.0678	64.55	15.47
Anchor-only	0.1360	**0.0712**	62.86	15.93
Anchor + NBI + WTI	**0.1377**	0.0706	**52.73**	**23.08**

We next test whether the three core mechanisms of NEXUS resolve their targeted failure modes: within-sample query convergence, hub-dominated cross-sample overmatching, and missing compositional organization.

Table 3: Ablation of delexicalized structural templates on EEG-to-text reconstruction.

Dataset	Template	$RSRA_{50}$ ↑	Groundedness ↑	Struct. Align. ↑
ZuCo1	No	**53.16**	**80.67**	49.24
	Yes	53.03	80.24	**50.54**
ZuCo2	No	53.15	**74.34**	51.86
	Yes	**53.33**	73.59	**52.41**

Within-sample disambiguation. The first challenge in fine-grained EEG decoding is that different semantic Units within the same sentence share substantial contextual information, causing multiple Unit queries to converge to similar representations and limiting their ability to recover distinct semantic components. NEXUS addresses this through shared-residual decomposition, which separates shared context from Unit-specific variation, together with balanced Query-Unit assignment and residual-specific supervision for discriminative fine-grained semantics. As shown in Table 2(a), removing the residual-specific objectives causes clear convergence among different Unit queries. Restoring the full Unit objective substantially reduces query similarity, increases effective rank, and also improves Unit-level retrieval. This shows that NEXUS mitigates within-sample query convergence and promotes semantically meaningful specialization that improves Unit retrieval, rather than merely dispersing the representations.

Cross-sample disambiguation. The second source of ambiguity remains: in the shared cross-sample semantic space, a small number of highly central candidates may repeatedly attract many unrelated EEG queries, leading to overmatching. Standard semantic anchoring does not explicitly distinguish this candidate-side bias from query-side susceptibility. NEXUS therefore calibrates candidate background affinity with NBI and adjusts the supervision strength according to EEG-query centrality with WTI.As shown in Table 2(b), in the historical Chisco source-validation diagnostic, anchor-only training improves Unit-level ranking but leaves erroneous matches highly concentrated. After adding NBI and WTI, CR10 decreases substantially, N_{eff} increases, and Unit MRR further improves, indicating that retrieval errors are less dominated by a small set of hub candidates. This supports the core design of NEXUS: beyond strengthening semantic alignment, it explicitly calibrates hubness-induced overmatching in cross-sample retrieval.

Compositional structure recovery. Even when fine-grained semantic Units are correctly identified, a third challenge remains in EEG-to-text reconstruction: a set of recovered semantic components does not by itself determine how they should be composed into a complete sentence. NEXUS therefore does not directly rely on the surface text of retrieved sentences for generation. Instead, it abstracts the retrieved sentences into delexicalized structural templates that retain only structural information, while the disambiguated Units provide the specific semantic content.As shown in Table 3, introducing the delexicalized structural template improves Structural Alignment while leaving $RSRA_{50}$ and Groundedness largely unchanged, indicating that the template mainly improves the compositional organization of recovered semantic evidence rather than introducing additional target-specific content. This supports the structural design of NEXUS: it enables structure-guided reconstruction by explicitly separating structural priors from semantic content during generation, rather than directly relying on the surface content of retrieved text.

5 Conclusion

We identified dual-scale semantic ambiguity as a key obstacle to fine-grained EEG semantic decoding: shared context induces within-sample query convergence, while highly central candidates cause cross-sample overmatching. To address these issues, we proposed NEXUS, which combines shared-residual semantic modeling, hubness-aware calibration, and delexicalized structural priors for EEG-to-text reconstruction. Experiments across natural-reading and imagined-speech settings show improved semantic discriminability, reduced hub concentration, and more reliable reconstruction of semantic content and compositional structure. These findings suggest that reliable EEG-to-text decoding benefits from disambiguating fine-grained semantic evidence before composing it into language.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. VICREG: VARIANCE-INVARIANCE-COVARIANCE RE- GULARIZATION FOR SELF-SUPERVISED LEARNING. 2022.

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross Modal Retrieval with Querybank Normalisation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5184–5195, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00513.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1):180291, December 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.291.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 138–146, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4.

Jiahe Li, Junru Chen, Fanqi Shen, Jialan Yang, Jada Li, Zhizhang Yuan, Baowen Cheng, Meng Li, and Yang Yang. Assembling the Mind's Mosaic: Towards EEG Semantic Intent Decoding, January 2026.

Zengrong Lin, Zheng Wang, Tianwen Qian, Pan Mu, Sixian Chan, and Cong Bai. NeighborRetr: Balancing Hub Centrality in Cross-Modal Retrieval.

Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. EEG2TEXT: Open Vocabulary EEG-to-Text Decoding with EEG Pre-Training and Multi-View Transformer, May 2024.

Xiaozhao Liu, Dinggang Shen, and Xihui Liu. Learning Interpretable Representations Leads to Semantically Faithful EEG-to-Text Generation, May 2025.

Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka, and Shounak Roychowdhury. Thought2Text: Text Generation from EEG Signal using Large Language Models (LLMs). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3747–3759, 2025. doi: 10.18653/v1/2025.findings-naacl.207.

Akshaj Murhekar, Christina Liu, Abhijit Mishra, Shounak Roychowdhury, and Jacek Gwizdka. SENSE: Efficient EEG-to-Text via Privacy-Preserving Semantic Retrieval, March 2026.

Abhishek Rathore and Arnav Bhavsar. Multi-View EEG-to-Image Generation via Wasserstein Distance-Guided Adversarial Feature Alignment. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, April 2026. doi: 10.1109/ISBI61048.2026.11515909.

Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. Enhancing EEG-to-Text Decoding through Transferable Representations from Pre-trained Contrastive EEG-Text Masked Autoencoder. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7278–7292, 2024. doi: 10.18653/v1/2024.acl-long.393.

Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9929–9939. PMLR, November 2020.

Yuchen Wang, Haonan Wang, Yu Guo, Honglong Yang, and Xiaomeng Li. Escaping the BLEU Trap: A Signal-Grounded Framework with Decoupled Semantic Guidance for EEG-to-Text Decoding, 2026.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the Vision-Brain Gap with an Uncertainty-Aware Blur Prior. 2025.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12310–12320. PMLR, July 2021.

Zihan Zhang, Xiao Ding, Yu Bao, Yi Zhao, Xia Liang, Bing Qin, and Ting Liu. Chisco: An EEG-based BCI dataset for decoding of imagined speech. *Scientific Data*, 11(1):1265, November 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-04114-1.

Huanran Zheng, Yuanbin Wu, Tianwen Qian, Wenjing Yue, and Xiaoling Wang. Guiding LLMs to decode text via aligning semantics in EEG signals and language. *Expert Systems with Applications*, 299:130300, March 2026a. ISSN 09574174. doi: 10.1016/j.eswa.2025.130300.

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning Brain Representation with Hierarchical Visual Embeddings, February 2026b.

Jiahao Zhou, Chenghao Xu, Wei Wang, Erkun Yang, and Cheng Deng. EEGiT: Teaching Vision Transformers to Understand the EEG signal. 2026.

Jinzhao Zhou, Yiqun Duan, Fred Chang, Thomas Do, Yu-Kai Wang, and Chin-Teng Lin. BELT-2: Bootstrapping EEG-to-Language representation alignment for multi-task brain decoding, August 2024.

A Appendix

B Implementation and Optimization Details

B.1 EEG and Text Encoders

The frozen EEG front-end is the project's earlier `FSTEncoder(large)`, pretrained on 39-way semantic classification of Chisco imagined-speech EEG using class-weighted cross-entropy with label smoothing. We use the retained checkpoint trained for 154 epochs with seed 20260904. The same frozen front-end is reused for ZuCo1, ZuCo2, and Chisco without dataset-specific retraining, and remains frozen throughout NEXUS optimization.

The front-end has four depthwise temporal convolutions with kernel sizes $15, 31, 63, 127$, each followed by GELU and a pointwise projection to 160 channels. Their concatenation is projected to 896 dimensions by a temporal convolution of kernel size 7 and stride 8. A 14-layer Transformer with 14 attention heads and feed-forward width 3584 processes these features. Adaptive average pooling of its non-CLS outputs produces 32 EEG tokens of dimension 896; the classification head is unused. Together with the following trainable contextual encoder, this front-end instantiates f_{EEG}.

The trainable encoder projects each 896-dimensional token to 768 dimensions, prepends a learned CLS token, and adds learned positional embeddings. It uses eight pre-layer-normalized Transformer blocks, 12 attention heads, feed-forward width 3072, GELU, and dropout 0.1. A final layer normalization precedes the 768-to-1024 CLS projection p_S. Its 32 non-CLS outputs form $\boldsymbol{H}_i$ for Unit decoding. The sentence projection and Unit branches therefore share the contextual EEG encoder but have distinct output heads.

The frozen text encoder f_T is Qwen3-Embedding-0.6B. We use the final hidden state of the last valid token, followed by ℓ_2 normalization, to obtain a 1024-dimensional embedding. Sentence targets encode the original sentence, while contextualized Unit targets encode the marked complete sentence described in Appendix C.1. Both are cached in float16; normalization and losses use float32. No semantic type, tier, or semantic-key string is inserted into the text encoder input. The text encoder receives no gradient in any stage.

B.2 Sentence-Level Semantic Alignment

Let $\mathbb{E}$ contain the normalized EEG sentence embeddings in a minibatch and $\mathbb{T}$ contain one frozen text embedding per distinct sentence in that batch. The sampler draws up to four different-subject EEG views of each sentence. For an anchor $\boldsymbol{a} \in \mathbb{E} \cup \mathbb{T}$, let $\mathbb{P}(\boldsymbol{a})$ contain the other embeddings with

Table 4: Verified optimization settings for the all-subject training runs. Batch sizes are nominal EEG-trial counts, ordered as ZuCo1/ZuCo2/Chisco. The context/common/residual learning rates refer to the three Unit parameter groups.

Stage	Epochs	Learning rates	Batch sizes
Sentence base	300	2×10^{-4}	512/768/768
Sentence refinement	60	2×10^{-4}	768/768/768
Unit learning	80	$(8, 5, 10) \times 10^{-5}$	256/256/256
Joint refinement	40	EEG: 10^{-5}; Unit: half above	128/160/160
Hubness calibration	20	$(2, 1.25, 2.5) \times 10^{-5}$	256/256/256

the same sentence identity. Thus, an EEG anchor's positives include the text target and other EEG views; a text anchor's positives are its EEG views. Self-comparisons are excluded. With $\tau_S = 0.07$, the implemented objective is

$$\ell(\boldsymbol{a}) = -\log \frac{\sum_{\boldsymbol{b} \in \mathbb{P}(\boldsymbol{a})} \exp(\boldsymbol{a}^\top \boldsymbol{b}/\tau_S)}{\sum_{\boldsymbol{b} \in (\mathbb{E} \cup \mathbb{T}) \setminus \{\boldsymbol{a}\}} \exp(\boldsymbol{a}^\top \boldsymbol{b}/\tau_S)}, \tag{16}$$

$$\mathcal{L}_{\text{sent}} = \frac{1}{2|\mathbb{E}|} \sum_{\boldsymbol{a} \in \mathbb{E}} \ell(\boldsymbol{a}) + \frac{1}{2|\mathbb{T}|} \sum_{\boldsymbol{a} \in \mathbb{T}} \ell(\boldsymbol{a}). \tag{17}$$

The positive probabilities are summed before taking the logarithm. EEG and text anchors are averaged separately and weighted equally, despite their different counts. Although the text targets are frozen, the text-anchor loss propagates gradients through its EEG comparisons.

B.3 Stage-Wise Optimization

The verified all-subject runs use seed 20260904 and optimize each dataset independently. The trainable contextual EEG encoder is randomly initialized; the frozen front-end and text encoder are reused. Sentence alignment is followed by frozen-encoder Unit learning, partial joint refinement, and frozen-encoder hubness calibration. During joint refinement, only the last four contextual EEG blocks, the final layer normalization, and the sentence projection are unfrozen; the input projection, CLS token, positional embeddings, and first four blocks remain frozen. The full Unit decoder is trainable during Unit, joint, and hubness stages. Joint refinement minimizes $\mathcal{L}_{\text{unit}} + \mathcal{L}_{\text{sent}}$.

All stages use AdamW with weight decay 0.03, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, gradient-norm clipping at 1, and bfloat16 autocast. Each stage uses linear warmup followed by cosine decay. The effective warmup lengths are $20, 6, 5, 4, 2$ epochs in the order of Table 4. Batch construction balances sentence categories and samples available subject views, so a minibatch can contain fewer EEG trials than its nominal capacity.

The base sentence checkpoint maximizes validation Sentence MRR. Unit checkpoints maximize validation Unit MRR. Joint checkpoints maximize the geometric mean of these two validation MRRs, subject to retaining at least 90% of the preceding sentence checkpoint's Sentence MRR. Sentence validation occurs at epoch 1, every 10 epochs, and the final epoch; Unit and joint validation occur at epoch 1, every 5 epochs, and the final epoch. Hubness calibration uses validation Unit MRR at every epoch, with the unmodified epoch-0 checkpoint eligible for selection. Training runs for the fixed epoch counts; test scores are not used for checkpoint selection.

The upstream 60-epoch sentence refinement retains $\mathcal{L}_{\text{sent}}$ and adds a cached minimal-pair contrastive term with weight 0.02. Each negative sentence replaces exactly one annotated Unit span using a fixed replacement dictionary and preserves all other characters; the same frozen text encoder embeds the original and modified sentences. For an eligible EEG trial, let s_i^+ be its cosine with the original sentence target and s_{ij}^- its cosine with cached negative j. The additional loss is

$$\mathcal{L}_{\text{ref}} = \left\langle -\log \frac{\exp(s_i^+/0.05)}{\exp(s_i^+/0.05) + \sum_j \rho_{ij} \exp(s_{ij}^-/0.05)} \right\rangle_i . \tag{18}$$

All cached negatives are used. The direction weight ρ_{ij} is the inverse frequency of that replacement direction among training sentences, normalized to mean one across observed directions and then

clipped to $[0.25, 4]$. The refinement checkpoint maximizes validation accuracy against each trial's hardest negative, subject to retaining at least 98% of the initial Sentence MRR; Sentence MRR breaks ties, and epoch 0 is eligible. This term is an upstream optimization detail and is absent from Unit, joint, and hubness learning.

C Within-Sample Disambiguation Details

C.1 Contextualized Unit Targets

Unit supervision uses the supplied atomic A/B span annotations and sense-aware semantic keys. These fixed annotations provide the spans and key identities; they are not generated by the EEG decoder. Let $[a_{im}, b_{im})$ be the exact character span of u_{im} in y_i. We form

$$\widetilde{y}_{im} = y_i[: a_{im}] \parallel \texttt{<target>} \parallel y_i[a_{im} : b_{im}] \parallel \texttt{</target>} \parallel y_i[b_{im} :], \tag{19}$$

where $\parallel$ denotes string concatenation. The span is checked against its annotated surface string before encoding. The entire marked sentence is encoded; the target is the normalized last-valid-token embedding, not a pool over the marked span alone. Consequently, the target can retain the Unit's sentence context. The shared target $\boldsymbol{\mu}_i^U$ is the arithmetic mean of normalized contextualized Unit targets and is not renormalized before computing $\boldsymbol{v}_{im}^U = \boldsymbol{t}_{im}^U - \boldsymbol{\mu}_i^U$. In particular, $\boldsymbol{\mu}_i^U$ is not the sentence embedding $\boldsymbol{t}_i^S$.

Repeated occurrences of an identical semantic key within one sentence contribute one target to its Unit set. The cache builder sets its token-length limit to at least the maximum encoded input length, with a minimum of 256 and an enforced maximum below 2048 for these runs.

C.2 Shared–Residual Unit Decoder

An input projection and a two-layer context adapter map $\boldsymbol{H}_i$ to 768-dimensional contextual features $\overline{\boldsymbol{H}}_i$. One learned common query reads these features through two cross-attention blocks. A layer normalization produces the common hidden state $\boldsymbol{h}_i^c$, which is linearly projected to $\boldsymbol{c}_i^U \in \mathbb{R}^{1024}$. The common and residual readers use cross-attention plus feed-forward blocks without query self-attention.

Gated residualization operates on the hidden contextual features. For token t, let $\boldsymbol{a}_i = W_c \boldsymbol{h}_i^c + \boldsymbol{b}_c$. Then

$$\boldsymbol{g}_{it} = \mathrm{sigmoid}\big(\mathrm{MLP}_g([\overline{\boldsymbol{h}}_{it}; \boldsymbol{a}_i])\big)\,, \tag{20}$$

$$\widetilde{\boldsymbol{h}}_{it} = \mathrm{MLP}_r\big(\overline{\boldsymbol{h}}_{it} - \boldsymbol{g}_{it} \odot \boldsymbol{a}_i\big)\,. \tag{21}$$

The gate MLP has widths $1536 \to 768 \to 768$ and GELU; the residual transformation is layer normalization, a $768 \to 768 \to 768$ GELU MLP, and another layer normalization. The Q learned queries read $\widetilde{\boldsymbol{H}}_i$ through four shared-parameter cross-attention blocks. A final normalization and projection produce the raw residual outputs $\boldsymbol{r}_{iq}^U \in \mathbb{R}^{1024}$. All Unit blocks have 12 heads, feed-forward width 3072, and dropout 0.1.

The number of queries is the larger of 10 and the maximum unique Unit count in a training sentence: $Q = 27, 24, 10$ for ZuCo1, ZuCo2, and Chisco, respectively. Queries have no predefined grammatical roles. Training reconstructs complete Unit vectors as $\boldsymbol{c}_i^U + \boldsymbol{r}_{iq}^U$, whereas formal retrieval and prototype ranking use

$$\boldsymbol{z}_{iq}^U = \mathrm{norm}(\boldsymbol{c}_i^U + \alpha \boldsymbol{r}_{iq}^U), \qquad \alpha = 0.5. \tag{22}$$

C.3 Capacity-Constrained Query–Unit Assignment

For a sentence with M_i Units, write $Q = b_i M_i + r_i$, where $0 \le r_i < M_i$. We allocate $b_i + 1$ query slots to r_i Units and b_i slots to the remaining Units. The extra capacities rotate deterministically with a hash of the sentence identity and assignment seed 611, avoiding a fixed preference for early annotation positions. Expanding each Unit into its allocated number of columns gives a square $Q \times Q$ assignment problem.

Matching similarities are cosines between normalized raw residual predictions and normalized target residuals, averaged across the available subject views of the same sentence. Hungarian matching maximizes the sum of these similarities. Deterministic sentence-specific row and column permutations resolve equal-score positional preferences; the matching is detached from gradient computation. The resulting assignment is shared across subject views. Each query receives one Unit; because $M_i \leq Q$, every Unit receives at least one query. For $M_i = 1$, all queries are assigned to the sole Unit.

C.4 Full Unit-Level Objective

To specify the actual reductions, let v index EEG subject views of sentence i, and let $a_i(q)$ be the assigned Unit. Define its capacity and grouped raw residual by

$$n_{im} = \sum_q A^*_{iqm}, \qquad \overline{\boldsymbol{r}}^U_{ivm} = \frac{1}{n_{im}} \sum_q A^*_{iqm} \boldsymbol{r}^U_{ivq}, \qquad \omega_{iq} = \frac{Q}{M_i n_{i,a_i(q)}}. \tag{23}$$

Let $d(\boldsymbol{a}, \boldsymbol{b}) = 1 - \mathrm{sim}(\boldsymbol{a}, \boldsymbol{b})$, and let $h(a, b)$ denote scalar SmoothL1 with transition point 1. Angle brackets below average the listed scalar losses over their valid indices. The shared and complete alignment terms are

$$\mathcal{L}^{\mathrm{dir}}_c = \langle d(\boldsymbol{c}^U_{iv}, \boldsymbol{\mu}^U_i) \rangle_{i,v}, \tag{24}$$

$$\mathcal{L}^{\mathrm{mag}}_c = \langle h(\|\boldsymbol{c}^U_{iv}\|_2, \|\boldsymbol{\mu}^U_i\|_2) \rangle_{i,v}, \tag{25}$$

$$\mathcal{L}_{\mathrm{full}} = \left\langle \frac{1}{Q} \sum_q \omega_{iq} d(\boldsymbol{c}^U_{iv} + \boldsymbol{r}^U_{ivq}, \boldsymbol{t}^U_{i,a_i(q)}) \right\rangle_{i,v}. \tag{26}$$

Capacity weights make the complete loss give equal aggregate importance to each Unit within a trial. Residual supervision acts on grouped queries assigned to the same Unit:

$$\mathcal{L}^{\mathrm{dir}}_r = \langle d(\overline{\boldsymbol{r}}^U_{ivm}, \boldsymbol{v}^U_{im}) \rangle_{i,v,m}, \tag{27}$$

$$\mathcal{L}^{\mathrm{mag}}_r = \langle h(\|\overline{\boldsymbol{r}}^U_{ivm}\|_2, \|\boldsymbol{v}^U_{im}\|_2) \rangle_{i,v,m}, \tag{28}$$

$$\mathcal{L}_{\mathrm{intra}} = \left\langle \left[0.05 - \mathrm{sim}(\overline{\boldsymbol{r}}^U_{ivm}, \boldsymbol{v}^U_{im}) + \max_{n \neq m} \mathrm{sim}(\overline{\boldsymbol{r}}^U_{ivm}, \boldsymbol{v}^U_{in}) \right]_+ \right\rangle_{i,v,m}. \tag{29}$$

Residual direction and magnitude supervision for multi-Unit sentences excludes targets with norm at most 10^{-6}. For a single-Unit sentence, the target residual is zero, only the residual magnitude penalty is retained, and the intra-sample and residual geometry terms are omitted.

For $M_i > 1$, the geometry regularizer matches off-diagonal cosine relations:

$$\mathcal{G}_{iv} = \frac{1}{M_i(M_i - 1)} \sum_{m \neq n} \left[\mathrm{sim}(\overline{\boldsymbol{r}}^U_{ivm}, \overline{\boldsymbol{r}}^U_{ivn}) - \mathrm{sim}(\boldsymbol{v}^U_{im}, \boldsymbol{v}^U_{in}) \right]^2. \tag{30}$$

The implementation groups sentences by $(M_i, \text{number of views})$, averages this quantity within each group, and averages those group means to form $\mathcal{L}_{\mathrm{geom}}$. It therefore preserves the target geometry instead of uniformly repelling every pair of queries.

Cross-view consistency uses stop-gradient, sentence-specific targets:

$$\widehat{\boldsymbol{c}}_i = \mathrm{sg}\left[\mathrm{norm}\left(\frac{1}{V_i} \sum_v \boldsymbol{c}^U_{iv} \right) \right], \tag{31}$$

$$\mathcal{L}_{\mathrm{cv},c} = \langle d(\boldsymbol{c}^U_{iv}, \widehat{\boldsymbol{c}}_i) \rangle_{i,v}, \tag{32}$$

$$\widehat{\boldsymbol{r}}_{im} = \mathrm{sg}\left[\mathrm{norm}\left(\frac{1}{V_i} \sum_v \mathrm{norm}(\overline{\boldsymbol{r}}^U_{ivm}) \right) \right], \tag{33}$$

$$\mathcal{L}_{\mathrm{cv},r} = \langle d(\overline{\boldsymbol{r}}^U_{ivm}, \widehat{\boldsymbol{r}}_{im}) \rangle_{i,v,m}. \tag{34}$$

Residual cross-view consistency is used only for multi-Unit sentences. A group with one available view contributes zero consistency loss.

Finally, the prototype ranking loss uses the normalized $\alpha = 0.5$ retrieval descriptor and a frozen prototype bank constructed from training rows only. For the assigned positive key j_{iq}^{+}, the fixed negative list $\mathbb{N}_{iq}$ contains 24 entries: eight nearest same-type prototypes, eight global nearest prototypes, and eight random prototypes, with deterministic deduplication and backfilling where necessary. It excludes the assigned key. With margin 0.05,

$$\ell_{ivqj}^{\text{proto}} = [0.05 - \text{sim}(\boldsymbol{z}_{ivq}^{U}, \boldsymbol{k}_{j_{iq}^{+}}) + \text{sim}(\boldsymbol{z}_{ivq}^{U}, \boldsymbol{k}_{j})]_{+}, \tag{35}$$

$$\mathcal{L}_{\text{proto}} = \left\langle \frac{1}{Q} \sum_{q} \frac{\omega_{iq}}{|\mathbb{N}_{iq}|} \sum_{j \in \mathbb{N}_{iq}} \ell_{ivqj}^{\text{proto}} \right\rangle_{i,v}. \tag{36}$$

The complete decomposition corresponding to the main paper is

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{c}^{\text{dir}} + 0.25\mathcal{L}_{c}^{\text{mag}} + \mathcal{L}_{\text{full}} + \mathcal{L}_{r}^{\text{dir}} + 0.25\mathcal{L}_{r}^{\text{mag}}, \tag{37}$$

$$\mathcal{L}_{\text{aux}} = 0.25\mathcal{L}_{\text{geom}} + 0.25\mathcal{L}_{\text{cv},c} + 0.25\mathcal{L}_{\text{cv},r} + 0.25\mathcal{L}_{\text{proto}}, \tag{38}$$

$$\mathcal{L}_{\text{unit}} = \mathcal{L}_{\text{align}} + 0.5\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{aux}}. \tag{39}$$

The residual-loss ablation jointly sets the weights of $\mathcal{L}_{r}^{\text{dir}}$, $\mathcal{L}_{r}^{\text{mag}}$, $\mathcal{L}_{\text{intra}}$, $\mathcal{L}_{\text{geom}}$, and $\mathcal{L}_{\text{cv},r}$ to zero. It retains the gated decoder, shared and complete alignment, shared consistency, prototype ranking, and capacity-constrained assignment.

D Cross-Sample Hubness Calibration Details

D.1 Candidate-Side Calibration (NBI)

We maintain a FIFO memory $\mathcal{M}$ of at most $10{,}240$ detached, normalized training EEG Unit-query descriptors. Before optimization, it is filled from a fixed random subset of training EEG trials without replacement, using seed $20260904 + 1771$. Each optimization step enqueues the current forward pass's descriptors after the parameter update, replacing the oldest entries. With normalized prototypes, candidate background affinity is

$$\widehat{C}_{j} = \frac{1}{|\mathcal{M}|} \sum_{\boldsymbol{z} \in \mathcal{M}} \boldsymbol{z}^{\top} \boldsymbol{k}_{j} = \boldsymbol{k}_{j}^{\top} \left(\frac{1}{|\mathcal{M}|} \sum_{\boldsymbol{z} \in \mathcal{M}} \boldsymbol{z} \right). \tag{40}$$

Both memory embeddings and this affinity estimate are detached. The prototype bank uses training Unit rows only.

For each query, Hungarian assignment supplies j_{iq}^{+}. We select the 20 highest-scoring non-GT prototypes under the current raw cosine, excluding *all* Unit keys annotated in the same EEG sample. Denote this neighborhood by $\mathbb{N}_{iq}^{\text{NBI}}$ and let $\mathbb{J}_{iq} = \{j_{iq}^{+}\} \cup \mathbb{N}_{iq}^{\text{NBI}}$. The adaptive target is

$$s_{iqj} = \text{sim}(\boldsymbol{z}_{iq}^{U}, \boldsymbol{k}_{j}), \tag{41}$$

$$H_{iqj} = \text{sg}\left[\frac{\exp((s_{iqj} - \widehat{C}_{j})/\tau_{\text{NBI}})}{\sum_{l \in \mathbb{N}_{iq}^{\text{NBI}}} \exp((s_{iql} - \widehat{C}_{l})/\tau_{\text{NBI}})} \right], \tag{42}$$

with $\tau_{\text{NBI}} = 1$. The normalized target gives mass $1/2$ to the assigned GT and total mass $1/2$ to the neighbors:

$$\widetilde{y}_{iqj} = \begin{cases} 1/2, & j = j_{iq}^{+}, \\ H_{iqj}/2, & j \in \mathbb{N}_{iq}^{\text{NBI}}. \end{cases} \tag{43}$$

For this term, p_{iqj} is evaluated by restricting the main-paper softmax to $\mathbb{J}_{iq}$ with temperature 1, and $\mathcal{L}_{\text{NBI}}$ averages its target-weighted negative log-probability over the individual query slots. Neighborhood membership is discrete and detached; gradients act through the selected current similarities. This objective does not impose a uniform candidate marginal.

D.2 Query-Side Centrality Calibration (WTI)

For each sentence and query index, WTI first averages normalized retrieval descriptors across available EEG subject views and normalizes the result again. Let $\{\overline{\boldsymbol{z}}_a\}_{a=1}^{n}$ be these pooled sentence–query descriptors. Their intra-modal centrality and weight are

$$C_a = \frac{1}{n}\sum_{b=1}^{n} \overline{\boldsymbol{z}}_a^{\top}\overline{\boldsymbol{z}}_b, \qquad w_a = \exp(0.3C_a). \tag{44}$$

The diagonal self-similarity is included in this statistic. Centrality and weights are not detached, so gradients also pass through the weighting function. Pooling avoids counting repeated subject views as separate semantic samples; different query slots remain distinct.

WTI uses the batch's unique assigned semantic keys as candidate columns, avoiding duplicate-key false negatives. For this term, p_{aj} is the softmax over these candidate columns with temperature 0.07. The implementation computes

$$\mathcal{L}_{\mathrm{WTI}} = -\frac{1}{n}\sum_{a=1}^{n} w_a \log p_{a,j_a^+}. \tag{45}$$

It averages weighted losses over descriptors, without dividing by the sum of weights. Only the EEG-to-Unit direction is optimized because the text prototypes are frozen.

D.3 Hubness Objective and Training Protocol

The calibration stage retains all semantic anchoring losses in Appendix C.4 and optimizes

$$\mathcal{L}_{\mathrm{calibration}} = \mathcal{L}_{\mathrm{unit}} + \mathcal{L}_{\mathrm{hub}}, \qquad \mathcal{L}_{\mathrm{hub}} = 0.25\mathcal{L}_{\mathrm{NBI}} + 0.05\mathcal{L}_{\mathrm{WTI}}. \tag{46}$$

The EEG encoder is fully frozen and the entire Unit decoder is updated. The stage runs for 20 epochs at batch size 256 with the learning rates in Table 4. Validation Unit MRR selects the checkpoint, including epoch 0 as a fallback. Inference remains raw cosine retrieval with $\alpha = 0.5$; no memory-dependent correction is subtracted at test time.

E Retrieval and Structure-Guided Reconstruction

E.1 Sentence- and Unit-Level Retrieval

The sentence gallery contains one frozen text vector per cached text ID. For Unit retrieval, contextualized targets sharing exactly the same semantic key are grouped to construct a prototype:

$$\boldsymbol{k}_j = \mathrm{norm}\left(\frac{1}{|\mathcal{I}_j|}\sum_{(i,m)\in\mathcal{I}_j} \mathrm{norm}(\boldsymbol{t}_{im}^{U})\right), \qquad s_{ij} = \max_{q=1,\dots,Q} \mathrm{sim}(\boldsymbol{z}_{iq}^{U}, \boldsymbol{k}_j). \tag{47}$$

Here $\mathcal{I}_j$ indexes cached Unit occurrences with semantic key j. The evaluation gallery uses all supplied text targets; it is a fixed dataset-wide text gallery rather than a training-only gallery. Exact keys are deduplicated, without merging synonyms or different semantic senses. We retrieve three sentences by sentence-level cosine similarity and ten Units by s_{ij}, using $\boldsymbol{z}_{iq}^{U} = \mathrm{norm}(\boldsymbol{c}_i^{U} + 0.5\boldsymbol{r}_{iq}^{U})$. The numbers of sentence/Unit candidates are respectively 1,038/5,230 for ZuCo1, 663/3,329 for ZuCo2, and 6,566/6,080 for Chisco. These full galleries provide the generation evidence; the separate 24-candidate sentence retrieval diagnostic does not restrict generation evidence. For each distinct ground-truth semantic key $j \in \mathbb{U}_i$ present in the gallery, the retrieval rank is

$$\mathrm{rank}_i(j) = 1 + \sum_{\ell} \mathbf{1}\{s_{i\ell} > s_{ij}\}. \tag{48}$$

Let M_i^{eval} count the distinct evaluable ground-truth keys of trial i. The implemented Unit metrics pool trial–key pairs:

$$\mathrm{MRR} = \frac{\sum_i \sum_{j\in\mathbb{U}_i^{\mathrm{eval}}} 1/\,\mathrm{rank}_i(j)}{\sum_i M_i^{\mathrm{eval}}}, \qquad \mathrm{R@}k = \frac{\sum_i \sum_{j\in\mathbb{U}_i^{\mathrm{eval}}} \mathbf{1}\{\mathrm{rank}_i(j) \le k\}}{\sum_i M_i^{\mathrm{eval}}}. \tag{49}$$

Tied scores receive the same optimistic rank under the strict inequality. This Unit-level aggregation differs from the trial-macro aggregation of candidate sentence scores used for text reconstruction.

E.2 Structural Analysis

The first frozen-LLM call receives the requested language, ten Units with their IDs, surface forms, and semantic types, and three retrieved sentences with their IDs and text. It returns one role and consistency verdict for every Unit, the IDs of supporting retrieved sentences, a construction type, an ordered sequence of role names, and an evidence quality label. A `supported` verdict denotes support within the retrieved evidence, not independently verified EEG decoding. The allowed roles are agent, predicate, patient, recipient, attribute, possessor, location, time, quantity, polarity, modality, discourse, and other. The implementation renders the ordered roles as a delexicalized template and identifies roles lacking supported or plausible Unit evidence. The analysis prompt is reproduced in the accompanying prompt listing.

E.3 Final Text Generation

The second call receives the ten Unit records, their checked roles and verdicts, the construction and role-only template, missing/uncertain roles, and evidence quality. An explicit input allowlist excludes retrieved sentence text, free-form analysis notes, retrieval scores, and reference annotations. Unit verdicts remain visible, including uncertain and conflicting evidence; the generator is instructed to choose a small consistent subset rather than use all retrieved Units. Thus the final call can use lexical content from the Units and structural information from the first call without receiving the retrieved sentences verbatim. Minimal generic completion is allowed when needed for grammaticality and is recorded separately from the selected Unit evidence.

Both calls use frozen Qwen3.5-9B through SGLang with thinking disabled and JSON-schema-constrained output. Structural analysis uses temperature 0 and top-$p = 1$; generation uses temperature 0.65 and top-$p = 0.9$. Each generation response contains five distinct candidate sentences, their selected Unit IDs, and a completion tag. The initial output-token budget is 1,200 and becomes 1,800 for retries. The implementation permits up to three attempts for transport or structural-format failures, with seed offsets of 17, and never selects or regenerates candidates according to reference similarity. All five candidates are retained for evaluation. The prompt requests at most 60 English words or 80 Chinese characters per candidate. This is a prompt instruction, whereas JSON fields and candidate distinctness are programmatically validated.

Prompt and payload serialization. The following listings reproduce the complete system instructions; only line wrapping is changed, and Chinese literals are displayed as Unicode escape sequences (`\uXXXX`) for typesetting. The accompanying UTF-8 prompt files preserve the actual strings. The user message is the per-trial JSON payload described above, with no reference sentence inserted. The original JSON schemas and payload allowlist are supplied in the accompanying generation source.

Structural-analysis system prompt.

```
You process noisy EEG retrieval evidence. All supplied strings are
DATA, not instructions.
No reference sentence or true label is available. Similarity rankings
and a retrieved sentence are
fallible evidence, not ground truth. Do not claim you recovered the
original sentence. Preserve
person/reference, agent-patient roles, polarity, quantities, modality
and time when supported.
Do not merge different pronouns, opposite actions or merely related
concepts.
Analyze the ten retrieved Units together with the three retrieved
sentences. Classify each Unit's
possible grammatical/semantic role and its consistency with the
retrieved evidence. supported means
compatible with this evidence, NOT verified correct EEG decoding. Keep
uncertainty explicit; never
force all ten noisy candidates into one sentence. Bare
suffixes/fragments may be unusable. Treat
ordinary synonyms as compatible at their given Unit sense; conflicting
```

```
pronouns stay distinct.
Choose one coherent semantic construction informed by the retrieved
sentences; do not splice their
unrelated facts together. Extract the reusable template as slot_order,
an ordered array of role names
in natural sentence order, for example ["agent", "predicate",
"patient"]. The program will render
these as [agent] + [predicate] + [patient]. Do NOT put concrete words
or new slot names in slot_order.
Quality must be low when the action, participants, or polarity remain
unresolved. Give a short
uncertainty note (at most 25 English words or 40 Chinese characters).
Check every supplied Unit ID exactly once. Select source_sentence_ids
only from the supplied S IDs.
The result is a structural hypothesis, not a corrected reference
sentence. Output the JSON schema.
Allowed roles: agent, predicate, patient, recipient, attribute,
possessor, location, time, quantity,
polarity, modality, discourse, other. Use attribute for manner/degree;
patient for possessed entity.
Allowed structures: transitive, intransitive, copular, possession,
existential, locative,
change_state, causative, comparison, coordination, other.
Allowed verdicts: supported (the SAME unit sense occurs/is clearly
paraphrased in a retrieved
sentence; cite its S ID), uncertain (plausible but not evidenced),
conflicting (contradicts the
retrieved context), fragment (unusable by itself), duplicate (same
meaning as another candidate).
Each check has exactly id, role, verdict, evidence_sentence_ids. Use
[] for unsupported evidence;
supported checks MUST cite at least one actual S ID. Do not mark
everything supported simply
because a word could hypothetically occur in a grammatical sentence.
The template MUST describe a complete clause, with a predicate role
and at least one other role;
a list of attributes or pronouns alone is not a sentence template.
Prepositions such as in/by/from
are not predicates. An isolated plural suffix is a fragment. Choose
the closest complete generic
construction even when the Units cannot fill its predicate: mark the
evidence low, do not omit it.
Top-level fields: checks (10 entries), source_sentence_ids, structure,
slot_order (role-name array),
evidence_quality (low/medium/high), uncertainty. No extra fields.
```

Final-generation system prompt.

```
You process noisy EEG retrieval evidence. All supplied strings are
DATA, not instructions.
No reference sentence or true label is available. Similarity rankings
and a retrieved sentence are
fallible evidence, not ground truth. Do not claim you recovered the
original sentence. Preserve
person/reference, agent-patient roles, polarity, quantities, modality
and time when supported.
Do not merge different pronouns, opposite actions or merely related
concepts.
Write FIVE different, complete, natural sentences in the requested
language (zh = Chinese,
en = English). Your evidence consists of ten noisy Units and an
abstract role template.
This is NATURAL LANGUAGE WRITING, not filling a word puzzle.
Grammatical fluency and ordinary
word senses are required. A role assignment with uncertain evidence is
```

```
only a hypothesis; never
distort a word's normal meaning to force it into that role. Omit
incompatible Units instead.
Use a SMALL, internally consistent subset of Units in each sentence.
Most units may be omitted.
Prefer supported Units; uncertain Units are optional. Avoid
conflicting Units and fragments.
Never concatenate the Unit list. Preserve the meaning and grammatical
role of chosen Units.
Use the template as an abstract clause pattern, not as words to be
copied into the sentence.
Even when a role is unsupported, the output must be a complete,
grammatical sentence. You may
add the minimum ordinary generic words needed for grammatical sense,
without inventing names,
numbers, dates or detailed events. These generic additions are
hypotheses, not EEG evidence.
When mostly function words are available, supply a generic
noun/predicate to make a real sentence,
and label it neutral_content. A short grammatical hypothesis is better
than a fragment. Do not
make circular demonstrative strings. Kinship words are relations, not
names or physical containers.
Do not end with an unfinished copula, preposition, determiner, or
dangling subordinate clause.
Do not write a statement ABOUT units, attributes, roles, evidence, or
the decoding task.
Do not force all five candidates to use exactly the same subset. Each
sentence must preserve its
selected person's identity, agent/patient relation and polarity; never
substitute one for another.
Natural inflections, connectors and word order changes are allowed.
Prefer short simple clauses.
You cannot see the retrieved sentences, so do not try to reconstruct
their unknown facts.

Return exactly one JSON object with key candidates, containing exactly
five objects. Each object:
- text: the ACTUAL natural sentence only, ending with punctuation. No
placeholders or commentary.
- unit_ids: IDs of Units whose meanings actually appear in that
sentence (at least one).
- completion: a METADATA LABEL ONLY. Use neutral_content if you added
a content word/meaning
beyond the chosen Units; grammar_only for added grammatical function
words/inflection only;
none only when the chosen Units supply all words. Generic words must
be written as actual
ordinary words in text, NOT as the label. These labels never belong
inside text.

NEVER put neutral_content, grammar_only, [agent], [predicate],
[patient], U01, slot names,
"\u4e2d\u6027\u5185\u5bb9", "\u6cdb\u5316\u5185\u5bb9", or other
control markers inside text. A missing meaning must be filled
with ordinary language, not a placeholder. Do not copy an example or
invent a new Unit ID.
Keep each sentence <=60 English words or 80 Chinese characters. All
five sentences must be
distinct beyond capitalization or punctuation. Return all five without
reference-based ranking.

Synthetic formatting examples ONLY (do not copy their words into your
output):
Units "\u53ef\u80fd, \u8fd9\u4e2a, \u662f":
```

```
"\u8fd9\u53ef\u80fd\u662f\u4e00\u4ef6\u666e\u901a\u7684\u4e8b\u3002"
uses a generic noun to complete the clause,
so completion is neutral_content. Do not output the unfinished string
"\u8fd9\u4e2a\u53ef\u80fd\u5c31\u662f\u3002".
Units "quiet, room": "The room is quiet." adds only grammar, so
completion is grammar_only.
Units "her, in": "Her belongings are in the room." needs generic
nouns, so completion is
neutral_content. Never output "Her in." or "Her in neutral_content.".
Write actual plausible sentences, not descriptions of a collection of
attributes.
Before returning, check all five texts for missing grammatical
complements and incorrect word senses.
```

F Evaluation Protocol and Metric Details

F.1 Datasets, Preprocessing, and Evaluation Setup

The NEXUS evaluation uses all subjects and disjoint training, validation, and test text IDs. Subjects may occur in all three partitions; this is sentence generalization rather than held-out-subject evaluation. Table 5 gives the actual supervised splits. ZuCo1 covers its NR, SR, and TSR tasks, ZuCo2 covers NR and TSR, and NEXUS uses the imagined-speech portion of Chisco.

Table 5: NEXUS split sizes. Parentheses give unique text IDs.

Dataset	Subjects	Train	Validation	Test
ZuCo1	12	7,831 (648)	885 (72)	2,364 (187)
ZuCo2	18	8,868 (473)	1,002 (54)	2,564 (132)
Chisco	5	17,831 (3,777)	1,954 (414)	4,961 (1,049)

EEG preprocessing. The fixed front-end produces 32 tokens of dimension 896 per trial. Chisco uses the archived imagined-speech token cache. For ZuCo, the 105 released channels are preserved in their original order and padded with 17 zero channels; this is a dimensional adapter, not an electrode-location correspondence. Signals in microvolts are converted to volts and divided by the recorded global RMS, $7.222148256600801 \times 10^{-6}$. Trials are encoded in 1,651-sample windows at the native 500 Hz rate, with stride 1,651 and a final window anchored at the trial end. Short trials are right-padded with zeros. Window tokens are concatenated and adaptively average-pooled to 32 tokens. The frozen Chisco-trained front-end is transferred to ZuCo without a new front-end training stage. The supervised partitions need not exhaust every cached EEG trial.

Baselines. On ZuCo1 and ZuCo2, we compare with GLIM-T5 and SID; the Chisco comparison uses SID. GLIM-T5 uses Flan-T5-large. Our SID implementation uses Qwen3-Embedding-0.6B for text embeddings and Qwen3.5-9B for generation.

For metrics computed per candidate, we first average the available candidate scores within each trial and then average trials equally. GLIM-T5 has one candidate and SID/NEXUS have five. No reference-based best-candidate selection is used, and reference text is available only to the offline scorer.

F.2 Calibrated RSRA_{50}

Reference gallery and aggregation. Let $\mathbb{G} = \{y^{(1)}, \ldots, y^{(M)}\}$ be the reference gallery for one dataset. We deduplicate exact original sentence strings across subjects and methods before computing document frequencies. The reference unions contain $M = 910$ sentences for ZuCo1, 584 for ZuCo2, and $5{,}681$ for Chisco. The ZuCo unions include NEXUS, GLIM-T5, and SID references; the Chisco union includes NEXUS and SID. Generated candidates do not enter the gallery. This lexical reference gallery is distinct from the semantic-key prototype gallery used for EEG-to-Unit retrieval. For trial i, let K_i be the number of generated candidates. Any candidate-level score a_{ik} is

Table 6: Raw target-specific score, exact background expectation, and random-reference sanity checks for the calibrated RSRA_{50} entries in Table 1. The last column reports mean $\pm$ standard deviation over 100 random-reference runs.

Dataset	Method	S	B	Random RSRA_{50}
ZuCo1	GLIM-T5	11.8193	9.1107	50.0193 ± 0.4117
	SID	7.6419	6.9226	49.9939 ± 0.2022
	NEXUS	11.4746	5.7691	50.0121 ± 0.2154
ZuCo2	GLIM-T5	17.0100	11.1506	50.0445 ± 0.5219
	SID	11.9880	13.0820	49.9864 ± 0.2987
	NEXUS	12.0208	5.7420	49.9928 ± 0.1847
Chisco	SID	4.2365	4.1970	49.9915 ± 0.0635
	NEXUS	21.7734	10.7375	49.9976 ± 0.1705

aggregated as

$$\overline{a} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_i}\sum_{k=1}^{K_i} a_{ik}. \tag{50}$$

All candidates are retained, including duplicate outputs and zero scores; no candidate is selected using its reference. $K_i = 1$ for GLIM-T5 and $K_i = 5$ for the available NEXUS and SID predictions.

Original target-specific score. Write C_{ik} and R_i for the unique normalized content-token sets of candidate k and its reference, respectively, and define

$$\mathrm{df}(w) = \sum_{j=1}^{M} \mathbf{1}[w \in R^{(j)}], \qquad P_{\mathrm{target}}(C, R) = \frac{|C \cap R|}{|C|}, \tag{51}$$

$$P_{\mathrm{chance}}(C, R) = \frac{\sum_{w \in C}\big(\mathrm{df}(w) - \mathbf{1}[w \in R]\big)}{(M-1)|C|}. \tag{52}$$

The chance term excludes the current target reference. The uncalibrated candidate score and its dataset aggregate are

$$s(C, R) = 100\frac{P_{\mathrm{target}}(C, R)}{P_{\mathrm{target}}(C, R) + P_{\mathrm{chance}}(C, R)}, \qquad S = \frac{1}{N}\sum_i \frac{1}{K_i}\sum_k s(C_{ik}, R_i). \tag{53}$$

Both precisions and s are zero for an empty C; s is also zero when its denominator is zero. The ratio is computed before candidate and trial averaging.

Background calibration. For each trial, replace its reference by one uniformly sampled different sentence from $\mathbb{G}$, shared by all candidates of that trial. In each such pairing, recompute the chance term excluding the new pseudo-target. For the fixed predictions of a given method and dataset, the exact background expectation is

$$B = \frac{1}{N}\sum_i \frac{1}{K_i}\sum_k \frac{1}{M-1}\sum_{j : y^{(j)} \neq y_i} s(C_{ik}, R^{(j)}). \tag{54}$$

We compute this finite-gallery expectation exactly and hold it fixed for both the observed score and random-reference checks. Each method/dataset has its own B, with the same reference gallery and text analyzer within each dataset. The reported calibration is

$$\mathrm{RSRA}_{50} = 50 + 50\frac{S - B}{100 - B}, \qquad B < 100. \tag{55}$$

Thus $S = B$ maps to 50, while $S = 100$ remains 100. Values below 50 are retained without clipping. The value 50 is a conditional background expectation, not 50% reconstruction accuracy or a significance threshold. This calibration does not equalize test-set difficulty or score variance.

The 100 checks use seed 20260923 and the fixed analytic B; their mean is not fitted to 50. Their standard deviation measures random-pairing variability, rather than uncertainty in the observed model score.

English/ZuCo content normalization. For ZuCo1 and ZuCo2, RSRA and Groundedness share the following content tokenizer. Its fixed order is regex tokenization, lowercasing, removal of spaCy's `spacy.lang.en.stop_words.STOP_WORDS`, and NLTK `PorterStemmer()` stemming. Stopword filtering precedes stemming.

```
from spacy.lang.en.stop_words import STOP_WORDS
from nltk.stem import PorterStemmer
import re

token_re = re.compile(r"[A-Za-z0-9]+(?:'[A-Za-z]+)?")
stop = set(x.lower() for x in STOP_WORDS)
stemmer = PorterStemmer()

def normalize_content_tokens(text):
toks = [t.lower() for t in token_re.findall(str(text))]
toks = [
stemmer.stem(t)
for t in toks
if t not in stop
]
return toks

def normalize_content_set(text):
return set(normalize_content_tokens(text))
```

Punctuation is discarded. The regex allows an internal ASCII apostrophe, so `didn't` remains one token; hyphens separate tokens, so `well-known` yields `well` and `known` before stopword filtering. Numeric tokens are retained. The code above specifies the complete normalization procedure. For RSRA, C_{ik} and R_i are the unique sets returned by `normalize_content_set` for the candidate and reference. Repeated occurrences of the same normalized token do not contribute more than once. Groundedness instead retains multiplicities, as specified below. These English rules do not alter the Chisco analyzer.

F.3 GROUNDEDNESS

Define token weights from the dataset-specific reference gallery as

$$\mathrm{IDF}(w) = \ln \frac{M+1}{\mathrm{df}(w)+1} + 1. \tag{56}$$

Tokens absent from the reference gallery have $\mathrm{df}(w) = 0$.

English/ZuCo content representation. Unlike the unique sets used by RSRA, English Groundedness uses multisets produced by `Counter`, retaining repeated normalized content tokens:

```
from collections import Counter

R = Counter(normalize_content_tokens(reference))
C = Counter(normalize_content_tokens(candidate))
```

Let $R_i(w)$ and $C_{ik}(w)$ denote the reference and candidate counts of normalized token w, respectively. The reference, candidate, matched, and unsupported information masses are

$$I_R(i) = \sum_w \mathrm{IDF}(w) R_i(w), \tag{57}$$

$$I_C(i,k) = \sum_w \mathrm{IDF}(w) C_{ik}(w), \tag{58}$$

$$I_M(i,k) = \sum_w \mathrm{IDF}(w) \min\{C_{ik}(w), R_i(w)\}, \tag{59}$$

$$U_{ik} = \max\{I_C(i,k) - I_M(i,k), 0\}, \tag{60}$$

$$\mathrm{Groundedness}_{ik} = 100 \frac{I_R(i)}{I_R(i) + U_{ik}}. \tag{61}$$

Thus matched content uses clipped counts, rather than a set intersection. An empty candidate content sequence receives zero. The score penalizes unsupported content; it does not measure completeness or recall. Scores are first averaged over candidates within a trial and then equally across trials, as in Appendix F.2.

ZuCo1 and ZuCo2 use separate IDF pools. For each dataset, the pool is the union of NEXUS, GLIM-T5, and SID references, deduplicated by the complete reference sentence; generated candidates do not enter it. Here M is the number of unique reference sentences, not the number N of EEG trials. Document frequency counts the presence of w once per unique reference:

$$\mathrm{df}(w) = \sum_{j=1}^{M} \mathbf{1}\{R^{(j)}(w) > 0\}. \tag{62}$$

Repeated tokens therefore affect the information masses within a text but do not increase that text's contribution to document frequency.

Chinese/Chisco content representation. For Chisco, C_{ik} and R_i remain unique content-token sets. Define the reference, candidate, matched, and unsupported information masses by

$$\begin{aligned} I_R(i) &= \sum_{w \in R_i} \mathrm{IDF}(w), \qquad I_C(i,k) = \sum_{w \in C_{ik}} \mathrm{IDF}(w), \\ I_M(i,k) &= \sum_{w \in C_{ik} \cap R_i} \mathrm{IDF}(w), \qquad U_{ik} = I_C(i,k) - I_M(i,k). \end{aligned} \tag{63}$$

Then

$$\mathrm{Groundedness}_{ik} = 100 \frac{I_R(i)}{I_R(i) + U_{ik}}. \tag{64}$$

The Chisco implementation assigns zero when either content set is empty and retains the observation. Otherwise, a nonempty candidate containing only reference-supported tokens receives 100, even if it omits much of the reference. Groundedness therefore penalizes unsupported content; it is not a completeness or recall measure. Candidate and trial averaging follow Appendix F.2.

Chinese/Chisco content processing. The frozen Chisco protocol uses Unicode NFKC, case folding, and collapsed whitespace. For texts present in the supplied atomic A/B Unit annotation table, it retains all annotated Unit surfaces, including overlapping Units and annotated pronouns or relational expressions. Lookup also accepts outer-punctuation variants after verifying the corresponding span offsets. No additional content stopword filter is applied to these annotated surfaces. For other texts, jieba 0.42.1 POS segmentation with its default dictionary and `HMM=True` supplies word boundaries. At each boundary, the longest training-Unit surface ending at a word boundary is retained, together with nested subspans explicitly observed within that surface in the annotation table. Unmatched tokens are retained if their coarse class is CONTENT, NEG, or NUM and they are absent from the frozen extra function-word lists. Unknown content words remain eligible. Normalized surfaces form a unique set; the English Porter stemmer is applied only to pure ASCII alphabetic surfaces. This deterministic extension is shared by references and predictions and does not inspect their pairing.

F.4 STRUCTURAL ALIGNMENT

Let $\sigma(\hat{y}_{ik})$ and $\sigma(y_i)$ denote the ordered coarse structural sequences. Repeated symbols are retained, and content stopword removal and stemming are not applied. With $L_{ik} = \mathrm{LCS}(\sigma(\hat{y}_{ik}), \sigma(y_i))$,

$$P_s(i,k) = \frac{L_{ik}}{|\sigma(\hat{y}_{ik})|}, \qquad R_s(i,k) = \frac{L_{ik}}{|\sigma(y_i)|}, \qquad \mathrm{SA}_{ik} = 100\frac{1.25P_s(i,k)R_s(i,k)}{0.25P_s(i,k) + R_s(i,k)}. \tag{65}$$

Empty sequences or zero LCS receive zero. The $F_{0.5}$ weighting places greater emphasis on structural precision. Scores are averaged over candidates and then trials.

English/ZuCo structural processing. This pipeline is independent of content normalization: it performs no stopword removal, stemming, or POS tagging, and preserves the original capitalization. The regex extracts word and punctuation tokens; tokens not beginning with an ASCII letter or digit are then discarded before mapping. Only lexicon lookup uses `low = tok.lower()`; the original token is retained for the capitalization test. Repeated structural symbols remain in order.

The fixed first-match priority is

$$\mathrm{NEG} > \mathrm{PRO} > \mathrm{AUX} > \mathrm{DET} > \mathrm{PREP} > \mathrm{CONJ} > \mathrm{NUM} > \mathrm{CAP} > \mathrm{CONTENT}.$$

The complete English/ZuCo lexicons are listed below. Overlapping memberships are retained and resolved by this priority.

```
NEG_SET

  not n't never no neither nobody nothing nowhere none


PRO_SET

  i me my mine myself you your yours yourself yourselves he him his
  himself she her hers herself it its itself we us our ours
  ourselves they them their theirs themselves who whom whose which
  what this that these those someone somebody something anyone
  anybody anything everyone everybody everything noone nobody
  nothing


AUX_SET

  am is are was were be been being have has had having do does did
  doing can could may might must shall should will would ought need
  dare used


DET_SET

  a an the this that these those my your his her its our their some
  any each every either neither no enough much many few fewer little
  less several all both half such what whatever which whichever
  whose


PREP_SET

  about above across after against along amid among around as at
  before behind below beneath beside besides between beyond by
  concerning despite down during except for from in inside into like
  near of off on onto opposite out outside over past regarding round
  since through throughout till to toward towards under underneath
  unlike until up upon via with within without


CONJ_SET
```

```
  and but or nor for yet so although because since unless while
  whereas if though once whether than that when whenever where
  wherever after before until as


NUMWORD_SET

  zero one two three four five six seven eight nine ten eleven
  twelve thirteen fourteen fifteen sixteen seventeen eighteen
  nineteen twenty thirty forty fifty sixty seventy eighty ninety
  hundred thousand million billion first second third fourth fifth
  sixth seventh eighth ninth tenth
```

The following code gives the tokenizer, punctuation filter, and ordered mapping used to construct the structural sequence:

```
  wordpunct_re = re.compile(
  r"[A-Za-z0-9]+(?:'[A-Za-z]+)?|[^\w\s]",
  re.UNICODE
  )

  def english_structure(text):
  symbols = []
  for tok in wordpunct_re.findall(text):
  if not re.match(r"[A-Za-z0-9]", tok):
  continue
  low = tok.lower()
  if low in NEG_SET:
  cat = "NEG"
  elif low in PRO_SET:
  cat = "PRO"
  elif low in AUX_SET:
  cat = "AUX"
  elif low in DET_SET:
  cat = "DET"
  elif low in PREP_SET:
  cat = "PREP"
  elif low in CONJ_SET:
  cat = "CONJ"
  elif low in NUMWORD_SET or re.fullmatch(
  r"\d+(?:\.\d+)?", low):
  cat = "NUM"
  elif tok[0].isupper():
  cat = "CAP"
  else:
  cat = "CONTENT"
  symbols.append(cat)
  return symbols
```

PREP precedes CONJ; hence overlapping entries such as `after`, `before`, `since`, `as`, and `until` map to PREP. CAP applies only after all earlier rules fail and means that the remaining token begins with an uppercase letter; it is not named-entity recognition. CONTENT does not distinguish nouns, verbs, adjectives, or adverbs. No POS tagger is used. The tokenizer splits `3.5` into `3`, a period, and `5`; after punctuation removal the sequence is `NUM NUM`. It leaves `didn't` as one token, without contraction expansion or splitting it into `did` and `n't`.

For illustration of this deterministic rule only, the sentence "Henry Ford did not found the company in 1936." maps token by token to the following sequence; this is not a model-generated example:

```
  Henry   Ford  did not found   the company in   1936
  CAP     CAP   AUX NEG CONTENT DET CONTENT PREP NUM
```

Table 7: Ordered coarse structural mapping used by the Chisco evaluation. Lexicon membership refers to the frozen Chinese and English function-word lists.

Symbol	First matching condition
NEG	Member of a fixed negation lexicon.
NUM	POS starts with `m`, or token is a digit sequence optionally separated by periods or commas.
DET	Member of a fixed determiner lexicon.
PRO	Member of the English pronoun lexicon, or POS starts with `r`.
AUX	Member of a fixed auxiliary lexicon.
PREP	Member of the English preposition lexicon, or POS starts with `p`.
CONJ	Member of the English conjunction lexicon, or POS starts with `c`.
OTHER	POS starts with `u`, `e`, `y`, `o`, `q`, or `w`.
CONTENT	All remaining letter- or number-bearing tokens.

Chinese/Chisco structural processing. To produce a linear sequence from overlapping annotated Units, select the shortest Unit at the earliest unconsumed start position, skip overlaps with consumed spans, and POS-tokenize the intervening text. Single-token Units use their jieba POS; multi-token Units use the unclassified flag `x`. Tokens without any Unicode letter or number are removed. The mapping below is applied in order, taking the first matching row. The fixed lexicons are included in the frozen Chisco evaluation protocol. The flag `eng` is first treated as `x`, so English words are not mistaken for interjections.

F.5 BERT-F1

We report raw BERTScore F1 as a complementary semantic-similarity measure, separately from the three task-specific dimensions. The scorer uses `roberta-large`, layer 17, with language `en` for ZuCo, and `bert-base-chinese`, layer 8, with language `zh` for Chisco. The installed distribution is `bert-score` 0.3.13 (its module and score hash report 0.3.12), with `transformers` 4.44.2. IDF weighting and baseline rescaling are disabled. We use the corresponding slow tokenizer, float32 evaluation, TF32 disabled, and batches of 32 pairs.

Original prediction and reference strings are passed to the official tokenizer without custom stemming, case folding, or translation. The maximum length is 512 model tokens including special tokens. Every prediction is compared with its single original reference, without choosing among alternate references or generated candidates. Candidate F1 values are averaged within trial and then equally across EEG trials. Raw BERT-F1 is neither an accuracy percentage nor directly comparable across the English and Chinese scoring encoders.

G Additional Ablation Details

G.1 Within-Sample Ablation Details

The comparison in Table 2(a) uses Chisco, the same completed sentence checkpoint, identical Unit initialization and sample stream, $Q = 10$, $\alpha = 0.5$, seed 20260904, and 80 Unit-only epochs. Gated residualization remains enabled in both variants. There is no joint refinement or NBI/WTI stage in this experiment. The ablated variant sets the residual direction, residual magnitude, intra-sample ranking, geometry, and residual cross-view consistency coefficients to zero; their full values are 1, 0.25, 0.5, 0.25, and 0.25, respectively. Shared-component supervision, complete-Unit alignment, shared cross-view consistency, prototype ranking, and capacity-constrained matching remain unchanged.

Let $\boldsymbol{R}_i$ contain the row-normalized residual query representations. Off-diagonal query cosine averages all $Q(Q-1)$ ordered query pairs. The archived effective-rank definition uses squared singular values:

$$p_{i\ell} = \frac{\sigma_{i\ell}^2}{\sum_a \sigma_{ia}^2}, \qquad \operatorname{erank}(\boldsymbol{R}_i) = \exp\left(-\sum_{\ell} p_{i\ell} \log p_{i\ell} \right), \tag{66}$$

Table 8: Fixed-epoch-80 Unit retrieval on 4,961 Chisco test trials and 6,080 semantic-key prototypes. This expands the main-paper ablation.

Setting	Unit MRR	R@1	R@5	R@10
w/o residual losses	0.085487	0.045416	0.113565	0.161679
Full Unit objective	0.092241	0.047215	0.130652	0.182663

with $0 \log 0 = 0$. All evaluation trials contribute to the statistics. The residual cosine is 0.983386 without the loss group and -0.092301 with the full Unit objective. Effective-rank mean/std/median are $1.088579/0.040682/1.080536$ and $4.278561/0.035065/4.277308$, respectively; standard deviations use the population convention. The main comparison fixes epoch 80 in advance; test scores do not select weights. These are single-seed, grouped-loss results and do not isolate the necessity of an individual loss or the effect of removing the gate.

G.2 Structural-Template Ablation Details

The paired ablation reuses the existing template-conditioned outputs and generates only the no-template condition for the same 2,364 ZuCo1 and 2,564 ZuCo2 EEG trials. EEG weights, full-gallery Top-10 Unit evidence, Top-3 sentence retrieval, and the saved per-trial structural analysis are held fixed. The no-template generation input removes `structure`, `template`, `slots`, `missing_slots`, and `uncertain_slots`, together with the corresponding template instructions. It retains Unit roles, consistency verdicts, and evidence quality; it therefore removes explicit template conditioning rather than all structural information. Both conditions use Qwen3.5-9B, temperature 0.65, top-$p = 0.9$, thinking disabled, and five retained candidates. The per-trial initial generation seeds are matched. No model is retrained. As shown in Table 3, adding templates changes Structural Alignment by $+1.30$ points on ZuCo1 and $+0.55$ on ZuCo2; the corresponding RSRA_{50} changes are -0.13 and $+0.18$, and Groundedness changes are -0.43 and -0.75 points.