BEYOND AGGREGATE GAINS: EVALUATING ROBUST CROSS-SUBJECT EEG-TO-TEXT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject EEG-to-text retrieval aims to recover linguistic content from individuals unseen during model development, yet improvements in aggregate retrieval performance do not necessarily indicate consistent benefits across unseen subjects. This distinction is especially important when architectural effects are comparable in magnitude to variability across subjects and stochastic training runs.

We conduct a controlled evaluation under a leave-one-subject-out setting with training-only normalization, subject-disjoint model selection, canonical retrieval over 400 unique sentence contents, and three fixed training seeds. On a cleaned five-subject ZuCo sentence-reading setting, we compare a Graph-Free EEG Transformer with prototype-based subject interaction, multi-head low-rank adaptation (MHLR), a Controlled BPR-style dense EEG–text baseline, and a deterministic Ridge control.

The Graph-Free reference achieves $0.544 \pm 0.407\%$ Unique-Sentence Top-1 and $2.549 \pm 0.855\%$ Unique-Sentence Top-5 accuracy. Prototype-Only SIG attains the highest descriptive means ($0.576 \pm 0.293\%$ and $2.765 \pm 0.828\%$), but its paired gains over Graph-Free are only $+0.032$ and $+0.216$ percentage points and vary in both magnitude and direction across held-out subjects and training seeds. MHLR exhibits similarly heterogeneous effects, while the Controlled BPR-style and Ridge baselines do not outperform the Graph-Free reference under the same retrieval task.

These experiments provide a concrete case in which a positive aggregate cross-subject retrieval gain does not translate into a consistent improvement across unseen subjects and stochastic training trajectories. The results motivate evaluating cross-subject EEG–language retrieval with subject-level effects and stochastic replication alongside aggregate performance.

1 INTRODUCTION

Recent EEG-to-text systems increasingly use pretrained language representations, contrastive alignment, and retrieval formulations to recover linguistic content from neural signals. Despite this progress, generalization across subjects remains difficult because EEG distributions vary substantially across individuals. Cross-subject evaluation therefore asks a particularly demanding question: whether a model trained on one set of subjects can recover linguistic information from a previously unseen subject. Importantly, a higher average cross-subject retrieval score does not necessarily imply that an intervention benefits unseen subjects consistently. Distinguishing aggregate performance gains from robust subject-level generalization is therefore important for assessing progress in cross-subject EEG-to-text retrieval.

Making this distinction empirically is challenging. Cross-subject results can depend not only on the model itself, but also on how subject roles, preprocessing, model selection, retrieval candidates, and stochastic training are handled. A held-out test subject may indirectly influence preprocessing statistics or model selection if these stages are not strictly separated. Repeated textual instances can change the effective retrieval candidate space, while variation across training seeds can be comparable to the gain attributed to an architectural intervention. Consequently, an aggregate Top-k

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.

improvement may combine positive and negative subject-level effects with training variability, obscuring whether the intervention provides a consistent benefit to unseen subjects.

We conduct a controlled empirical study of cross-subject EEG-to-text retrieval to examine this distinction. All model comparisons are performed under the same strict leave-one-subject-out setting. EEG normalization is estimated exclusively from training subjects, a subject-disjoint validation set is used for model selection, and the held-out test subject is evaluated only after selection is frozen. Retrieval is performed against a deterministic candidate bank containing each unique sentence exactly once. For stochastic models, we repeat training with three fixed seeds and use matched subject–seed comparisons to separate average performance from the consistency of an architectural effect. These controls are not introduced as a new decoding architecture; rather, they provide a common experimental basis for asking whether apparent cross-subject gains persist across subjects and training trajectories.

Our results illustrate a distinction between aggregate improvement and robust improvement. Prototype-based subject interaction achieves the highest descriptive mean, exceeding the graph-free reference by $+0.032$ percentage points in Top-1 accuracy and $+0.216$ percentage points in Top-5 accuracy. However, its matched effects vary in both magnitude and direction across held-out subjects and training seeds. Strictly selected multi-head low-rank adaptation exhibits a similarly heterogeneous pattern, with no consistent advantage over the graph-free reference. A controlled BPR-style dense-retrieval model performs below the graph-free reference under the same task, while a deterministic Ridge control remains near chance. Thus, the model with the highest aggregate mean is not necessarily one that improves retrieval consistently across unseen subjects. In this setting, aggregate Top-k gains alone provide limited evidence of robust cross-subject generalization.

Our contributions are threefold: (1) we provide controlled empirical evidence that aggregate cross-subject retrieval gains can mask substantial heterogeneity across held-out subjects and training seeds, showing that higher mean Top-k performance does not by itself establish robust subject-level improvement; (2) we characterize this gap through matched subject–seed comparisons across graph-free, prototype-interaction, low-rank adaptation, dense-retrieval, and linear models, allowing architectural effects to be examined beyond aggregate means; and (3) we perform these comparisons under a common, strictly controlled evaluation setting that isolates training, validation, and test subjects throughout preprocessing and model selection, fixes the content-level retrieval task with a deterministic unique-sentence candidate bank, and repeats stochastic training under fixed seeds to support reproducible robustness analysis.

2 Related Work

2.1 EEG-to-Text Representation Learning

Non-invasive EEG-to-text decoding has increasingly shifted toward open-vocabulary language modeling and transferable EEG representations. Wang and Ji (Wang & Ji, 2022) introduced an open-vocabulary sequence-to-sequence formulation for mapping EEG recorded during natural reading to text. DeWave (Duan et al., 2023) subsequently combined discrete EEG codices with contrastive EEG–text alignment to decode continuous EEG with reduced reliance on fixation-based word markers, while CET-MAE (Wang et al., 2024) combined cross-modal contrastive learning with masked reconstruction to pretrain transferable EEG representations. These approaches increasingly connect EEG representations with pretrained language models, but generation-based evaluation can make it difficult to separate information recovered from EEG from the language prior of the decoder, motivating complementary discriminative and retrieval-based evaluation.

2.2 EEG-Based Semantic Retrieval and Cross-Subject Generalization

EEG–text retrieval evaluates neural representations against an explicit candidate space rather than through unconstrained generation. Zhou et al. (Zhou et al., 2024) learned linguistic EEG representations through masked contrastive learning before sentence retrieval, while Brain Passage Retrieval (BPR) (McGuire & Moshfeghi, 2025) mapped EEG queries and textual passages into a shared semantic space. Brain-CLIPLM (Yang et al., 2026) further explored semantic compression through intermediate semantic anchors. Together, these studies motivate retrieval as a complementary test of whether EEG representations preserve discriminative linguistic information.

Cross-subject retrieval remains challenging because inter-subject EEG variability creates distribution shifts between training and unseen participants (Wu et al., 2022; Li et al., 2026). Prior work has therefore explored transfer, alignment, domain adaptation, and subject-invariant representations (Wu et al., 2022; Li et al., 2026); DeWave likewise motivates its discrete representation partly by individual EEG differences (Duan et al., 2023). Because existing studies also differ in subject partitions, preprocessing, textual targets, candidate spaces, and evaluation units, generation, sentence-retrieval, and passage-retrieval results are not directly interchangeable. We instead study controlled content-level retrieval for entirely held-out subjects and examine whether architectural effects remain consistent across subjects and training seeds.

2.3 Evaluation Reliability in EEG-to-Text Modeling

Evaluation design can materially affect conclusions about EEG-to-text systems. Jo et al. (Jo et al., 2026) identified implicit teacher forcing as a confound in autoregressive EEG-to-text evaluation and used noise-input controls to test whether generation performance genuinely depends on neural information. Their analysis highlights a broader issue: favorable aggregate decoding metrics do not by themselves establish what information has been recovered from EEG.

Cross-subject evaluation raises complementary concerns because inter-subject distribution shift requires strict separation of training, validation, and held-out participants (Wu et al., 2022; Li et al., 2026). Preprocessing, model selection, adaptation, and stochastic training can otherwise affect the interpretation of small architectural gains. Rather than asking whether a generation score reflects information genuinely decoded from EEG, we ask whether a positive aggregate cross-subject retrieval gain corresponds to consistent improvement across held-out subjects and training trajectories. Accordingly, we use common subject roles and candidate construction, multiple fixed seeds, and matched subject–seed effects alongside aggregate Top-k performance.

3 Controlled Cross-Subject Evaluation

We evaluate whether modeling interventions improve retrieval for subjects that are entirely unseen during model development. To make such comparisons interpretable, subject roles, preprocessing, model selection, retrieval construction, and stochastic evaluation are controlled across methods. For every outer fold, three subjects are used for training, one distinct subject for validation, and one subject for final testing. Training-subject EEG alone determines normalization statistics and model parameters; validation data may determine checkpoints or predeclared hyperparameters; and the test subject is accessed only after all selection decisions have been frozen. Figure 1 summarizes this evaluation procedure.

3.1 Problem Formulation and Outer Subject Roles

Let $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{T_i \times 105}$ is a variable-length EEG sequence, y_i its sentence content, and s_i its biological subject. We evaluate five subjects (ZAB, ZKH, ZKW, ZMG, and ZPH) under five outer folds. In each fold, one subject is held out for testing, one distinct subject is reserved for validation, and the remaining three are used for training. Thus, every subject serves once as the outer test subject while remaining excluded from both parameter optimization and model selection. Exact fold assignments are given in Appendix A.

Retrieval uses cosine similarity between EEG and candidate-text representations. Graph-Free, Prototype-Only SIG, Strict MHLR, and Ridge use the frozen 4096-dimensional Llama semantic space, whereas Controlled BPR-style uses a 768-dimensional DistilBERT-based space. Candidate strings, ordering, target identities, and the Top-k retrieval definition are nevertheless fixed across methods.

3.2 Training-Only Normalization and Validation-Only Selection

For each outer fold, channel-wise EEG normalization statistics are estimated exclusively from valid, unpadded observations of the three outer-training subjects and are then frozen. The same training-derived transformation is applied to training, validation, and test EEG; neither validation nor test observations contribute to its estimation.

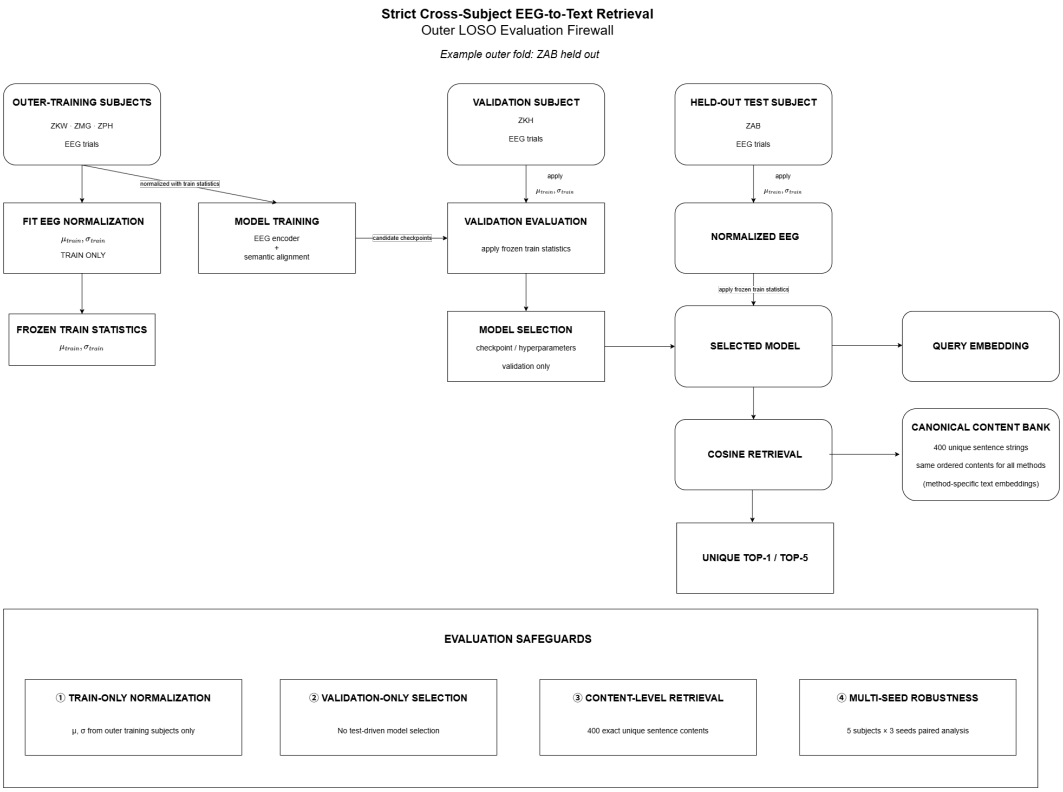


Figure 1: Controlled cross-subject evaluation procedure. Training subjects determine preprocessing statistics and model parameters, while a distinct validation subject is used for model or hyperparameter selection. All selection decisions are frozen before the held-out test subject is evaluated. Final retrieval is performed against the same deterministic bank of 400 unique sentence contents.

Model selection is likewise isolated from the outer test subject. Checkpoints and hyperparameters are selected using only the designated validation data or, for MHLR, nested selection among outer-training subjects. The held-out test subject is accessed only after preprocessing and all selection decisions have been frozen. Full normalization, validation-objective, and model-selection details are provided in Appendix A.

3.3 Canonical Content-Level Retrieval

All methods are evaluated against the same canonical bank of 400 unique sentence strings. Candidate identities and ordering are fixed before model comparison, and each query is mapped to its candidate by exact string equality. Retrieval ranks candidates by cosine similarity, and Unique-Sentence Top-k accuracy measures whether the matched candidate appears among the k highest-scoring entries. The corresponding uniform-retrieval chance levels are 0.25% for Top-1 and 1.25% for Top-5.

This content-level protocol prevents repeated sentence instances or method-specific candidate construction from altering the retrieval task. Exact bank-construction, string-matching, and fingerprinting procedures are provided in Appendix A.4.

3.4 Matched Subject–Seed Robustness Analysis

Aggregate means do not reveal whether an intervention benefits unseen subjects consistently. We therefore compare each stochastic method with the matched Graph-Free run for the same held-out subject m and training seed r:

$$\Delta_{m,r}^{(a)} = R_{m,r}^{(a)} - R_{m,r}^{(\mathrm{GF})}. \tag{1}$$

Graph-Free, Prototype-Only SIG, Strict MHLR, and controlled BPR-style are evaluated using three fixed seeds. We summarize paired effects by their mean, positive/equal/negative counts, and seed-level averages, with the full subject–seed structure examined in Section 6. These summaries are descriptive: the 15 subject–seed cells comprise only five biological subjects across three stochastic runs and are not treated as 15 independent biological observations.

4 Models and Controlled Comparisons

4.1 Graph-Free EEG Semantic Encoder

Graph-Free serves as the neural reference model. Each 105-dimensional EEG sequence is projected to a 512-dimensional latent space, augmented with fixed sinusoidal positional encoding, and processed by a four-layer Transformer. Masked temporal mean pooling is followed by a projection into the frozen 4096-dimensional Llama semantic space and ℓ_2 normalization:

$$z_i = \mathrm{norm}_2(W_{\mathrm{sem}}\bar{h}_i + b_{\mathrm{sem}}).$$

Text targets are frozen Llama representations obtained by attention-mask-aware mean pooling of final-layer token representations. The EEG encoder is trained with a one-directional EEG-to-text contrastive objective. All architectural, target-construction, and optimization details are provided in Appendix B.

4.2 Prototype-Based Subject Interaction

Prototype-Only SIG introduces batch-local subject interaction without learned subject-ID embeddings. Each semantic representation is projected to $r_i \in \mathbb{R}^{256}$, and a prototype p_a is formed by averaging representations belonging to subject a within the minibatch. Subject prototypes interact through cosine-based attention,

$$\alpha_{ab} = \frac{\exp(q_a^\top q_b)}{\sum_j \exp(q_a^\top q_j)}, \qquad q_a = \frac{p_a}{\|p_a\|_2},$$

producing subject context $g_a = \sum_b \alpha_{ab} p_b$. The context is projected back to the semantic space and fused with the Graph-Free representation through a learned gated residual,

$$\tilde{z}_i = z_i + \rho_i u_i, \qquad \rho_i = \sigma\big(w_g^\top [z_i; u_i] + b_g\big)\,.$$

Subject labels are used only to define within-batch groups; no trainable subject-ID lookup or persistent prototype memory is used. Additional implementation details and the Prototype-Conditioned ablation are provided in Appendix B.2.

4.3 Multi-Head Low-Rank Subject-Level Adaptation

MHLR adapts frozen Graph-Free representations through independent low-rank residual transformations. For head r,

$$\tilde{z}_i^{(r)} = z_i^{(r)} + U_r D_r z_i^{(r)},$$

where $D_r \in \mathbb{R}^{q\times 512}$ and $U_r \in \mathbb{R}^{512\times q}$. The adapted heads are concatenated to recover the original semantic dimensionality.

Only the low-rank projector is optimized during unsupervised subject-level transductive adaptation. Fold-specific ranks and adaptation coefficients are selected using nested roles among outer-training subjects and frozen before outer-test adaptation. Initialization and complete adaptation details are provided in Appendix B.3.

4.4 Controlled BPR-Style Baseline and Ridge Control

We additionally evaluate two controls under the same LOSO protocol and canonical retrieval bank. The controlled BPR-style baseline, inspired by Brain Passage Retrieval (BPR) (McGuire & Moshfeghi, 2025), uses a Transformer-based EEG encoder aligned with frozen DistilBERT (Sanh et al.,

2019) representations through a duplicate-safe variant of InfoNCE (van den Oord et al., 2018). Ridge provides a non-neural linear control by mapping normalized EEG to the same frozen Llama target space used by Graph-Free.

Both controls follow the same train-only preprocessing and validation-only selection principles. Full architectures, objectives, and selection procedures are provided in Appendix B.4.

5 Experimental Setup

5.1 Dataset and Protocol

We evaluate all models on a cleaned sentence-reading setting derived from the Zurich Cognitive Language Processing Corpus (ZuCo) (Hollenstein et al., 2018). The controlled setting contains five subjects (ZAB, ZKH, ZKW, ZMG, and ZPH), 105 EEG channels, 1,962 query instances, and 400 unique sentence strings. All methods use the strict five-fold subject roles and canonical candidate bank defined in Section 3.

Graph-Free, Prototype-Only SIG, MHLR, and controlled BPR-style are evaluated over three predeclared seeds (20260824, 20260825, and 20260826). Prototype-Conditioned is evaluated as a one-seed ablation, while Ridge is deterministic and evaluated once per outer fold.

5.2 Optimization

Graph-Free and SIG use AdamW (Loshchilov & Hutter, 2019); BPR also uses AdamW under its fixed production configuration. MHLR performs one subject-level adaptation pass using fold-specific settings selected exclusively from outer-training subjects, while Ridge selects from the fixed regularization grid described above. Complete optimization hyperparameters and fold-specific MHLR configurations are provided in Appendix B.5.

5.3 Metrics and Aggregation

Primary metrics are Unique-Sentence Top-1 and Top-5 accuracy over the 400-content bank. For the four three-seed stochastic families, mean and sample standard deviation are computed across all 15 subject–seed cells. Prototype-Conditioned and Ridge are summarized across their five subject folds. For matched stochastic comparisons against Graph-Free, we additionally report paired mean changes and inspect sign and seed consistency.

6 Results and Analysis

6.1 Overall Cross-Subject Retrieval

Table 1 summarizes performance under the strict cross-subject protocol. Stochastic methods are reported as mean $\pm$ sample standard deviation over 15 subject–seed cells, while deterministic Ridge is summarized over the five outer folds. Matched comparisons additionally report paired percentage-point changes relative to Graph-Free.

Graph-Free exceeds uniform-retrieval chance at both retrieval depths. Prototype-Only SIG achieves the highest descriptive means, with paired changes of $+0.032$ pp Top-1 and $+0.216$ pp Top-5 relative to Graph-Free. Strict MHLR shows a mixed aggregate effect (-0.017 pp Top-1; $+0.118$ pp Top-5), while Controlled BPR-style underperforms Graph-Free and Ridge remains close to chance.

These aggregate results provide limited evidence of a consistent architectural advantage. In particular, the positive mean Top-5 effects of Prototype-Only SIG and Strict MHLR motivate examining whether the gains persist across held-out subjects and training seeds rather than relying on aggregate means alone.

Table 1: Strict cross-subject EEG-to-text retrieval performance. Mean ± sample SD is reported in percent. Δ denotes the paired percentage-point change relative to Graph-Free. Bold indicates the highest descriptive mean.

Method	**Top-1 (%)**	**Top-5 (%)**	**ΔT1 (pp)**	**ΔT5 (pp)**
Chance	0.250	1.250	–	–
Ridge	0.255 ± 0.003	1.223 ± 0.105	–	–
Controlled BPR-style	0.289 ± 0.133	1.171 ± 0.391	-0.254	-1.378
Graph-Free	0.544 ± 0.407	2.549 ± 0.855	Ref.	Ref.
Prototype-Only SIG	$\mathbf{0.576 \pm 0.293}$	$\mathbf{2.765 \pm 0.828}$	$+0.032$	$+0.216$
Strict MHLR	0.527 ± 0.402	2.667 ± 0.970	-0.017	$+0.118$

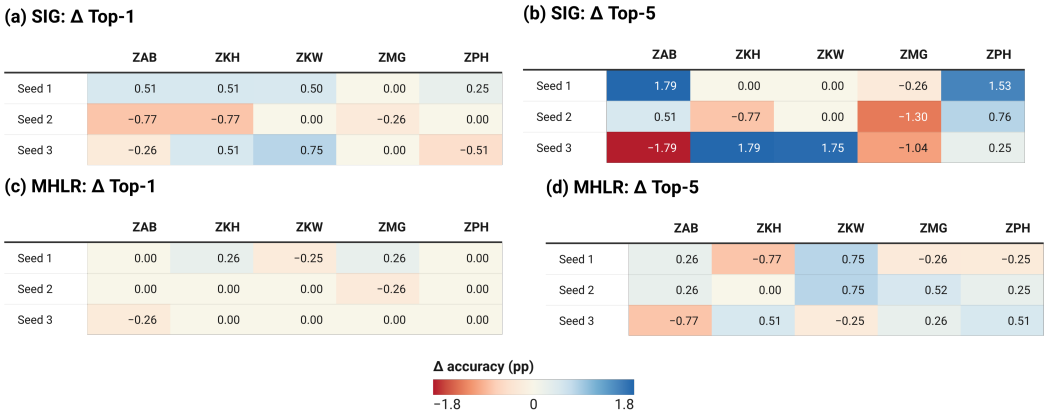


Figure 2: Subject- and seed-level variability of architectural effects under controlled cross-subject evaluation. Each cell reports the paired percentage-point change in Unique-Sentence Top-1 or Top-5 accuracy relative to the matched Graph-Free reference for the same held-out subject and training seed. Positive values indicate improvement over Graph-Free and negative values indicate degradation. Prototype-Only SIG exhibits a small positive mean effect but mixed cell-level consistency, while Strict MHLR similarly fails to produce uniform gains across subjects, seeds, and retrieval depths.

6.2 Aggregate Improvement Does Not Imply Consistent Improvement

Although Prototype-Only SIG achieves the highest aggregate Top-5 accuracy in Table 1, Figure 3 shows substantial heterogeneity across matched held-out-subject–seed pairs, with effects on both sides of the Graph-Free reference.

The fold-averaged paired Top-5 changes across the three fixed seeds are $+0.611$, -0.157, and $+0.195$ pp, while the corresponding Top-1 changes are $+0.355$, -0.358, and $+0.100$ pp. Positive and negative effects also coexist across held-out subjects within individual seeds. Thus, the aggregate Top-5 gain of $+0.216$ pp provides a concrete example in which a positive aggregate cross-subject gain does not translate into consistent improvement across unseen subjects and stochastic training trajectories. We therefore interpret the higher SIG mean as a descriptive aggregate improvement rather than evidence of robust superiority.

6.3 Strict MHLR Does Not Provide a Robust Gain

Strict MHLR shows a similarly mixed pattern in Figure 2. Seed-level Top-1/Top-5 paired changes are $(+0.053, -0.054)$, $(-0.052, +0.356)$, and $(-0.051, +0.052)$ pp across the three fixed seeds. Across all 15 subject–seed cells, the positive/equal/negative counts are $2/10/3$ for Top-1 and $9/1/5$ for Top-5. Consistently, the aggregate effect changes direction with retrieval depth, from -0.017 pp Top-1 to $+0.118$ pp Top-5. This variation across seeds, subjects, and retrieval depths does not support a robust MHLR improvement claim under the present evaluation.

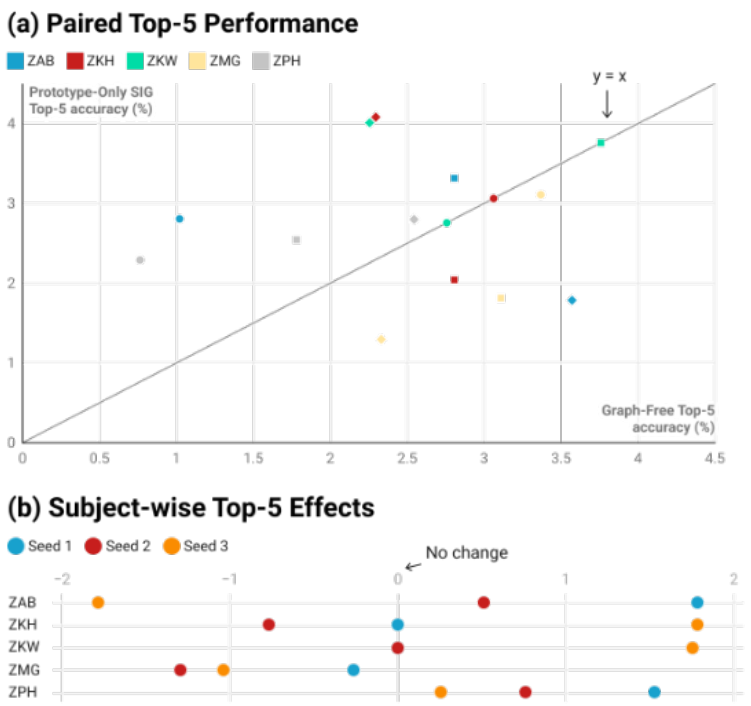


Figure 3: Paired Prototype-Only SIG analysis. (a) Matched Graph-Free versus SIG Top-5 accuracy; the diagonal denotes equal performance. (b) Corresponding SIG-minus-Graph-Free Top-5 differences across held-out subjects and training seeds. Positive values favor SIG.

Table 2: Descriptive consistency of paired Top-5 effects relative to the matched Graph-Free reference. Positive, equal, and negative counts summarize the 15 held-out-subject–seed cells. Positive seeds denotes the number of training seeds for which the fold-averaged paired Top-5 change is positive. These counts summarize computational replication and are not treated as independent biological observations.

Method	**Mean ΔT5 (pp)**	**Pos.**	**Equal**	**Neg.**	**Pos. Seeds**
Prototype-Only SIG	+0.216	7	3	5	2/3
Strict MHLR	+0.118	9	1	5	2/3
Controlled BPR-style	−1.378	2	0	13	0/3

6.4 Controlled External and Linear Baselines

As shown in Table 1, Controlled BPR-style underperforms Graph-Free (-0.254 pp Top-1; -1.378 pp Top-5), while Ridge remains near chance. This does not imply that the original published BPR system is inferior: our implementation is a controlled adaptation to the present 105-channel input and canonical sentence-retrieval task. Rather, under the same strict information constraints, neither the adapted dense dual encoder nor the linear control outperforms Graph-Free.

Table 2 further separates aggregate effect magnitude from descriptive consistency. Prototype-Only SIG has the larger mean Top-5 gain ($+0.216$ pp) but fewer positive cells than Strict MHLR (7/15 vs. 9/15); both yield positive fold-averaged effects in only 2/3 seeds. Controlled BPR-style instead has 13/15 negative cells and no positive seed-level mean. These counts are descriptive only: the 15 cells comprise five biological subjects across three stochastic runs and are not treated as independent biological observations for statistical inference.

7 Discussion and Limitations

7.1 What Constitutes Robust Cross-Subject Improvement

The main finding of this study is that aggregate cross-subject performance and robust cross-subject improvement are not equivalent. Prototype-Only SIG achieves the highest descriptive mean retrieval accuracy, yet its paired effect changes direction across held-out subjects and training seeds. Strict MHLR shows a related pattern, improving mean Top-5 without a corresponding mean Top-1 improvement. These results provide a concrete case in which a positive aggregate gain does not translate into a consistent benefit across unseen subjects and stochastic training trajectories.

This distinction is particularly important in cross-subject EEG evaluation, where each held-out subject represents a biological domain shift rather than merely another random test partition. A single favorable seed or aggregate mean can therefore obscure substantial heterogeneity in the underlying subject-level effects. In our SIG comparison, for example, one training seed produces positive fold-averaged Top-1 and Top-5 changes, whereas another reverses both signs. Reporting paired subject–seed effects alongside aggregate performance provides a more complete description of whether an architectural improvement is consistently reproduced.

We emphasize that this consistency analysis is descriptive rather than a test of statistical superiority. The 15 subject–seed cells arise from five biological subjects evaluated across three stochastic training runs and are not 15 independent biological observations. Accordingly, our results support cautious interpretation of aggregate gains rather than a general superiority claim for any evaluated architecture.

7.2 Scope and Limitations

The present evidence is limited to one cleaned ZuCo sentence-reading setting with five biological subjects. Multiple training seeds expose stochastic sensitivity but do not increase the number of independent participants, so larger independent EEG cohorts are required for stronger subject-level inference and external generalization.

The primary task is retrieval from a canonical bank of 400 unique sentence strings. This controlled setting enables matched content-level evaluation, but it is not an open-vocabulary EEG-to-text generation task. In addition, although candidate strings, target identities, and Top-k definitions are shared across model families, their semantic vector spaces are not identical: Graph-Free, SIG, and MHLR use frozen Llama targets, whereas the Controlled BPR-style model uses DistilBERT-based text representations with its trainable adapter. Cross-family differences therefore cannot be attributed exclusively to the EEG architecture.

Finally, the Controlled BPR-style model is an adaptation to the present 105-channel input and canonical retrieval task rather than an exact reproduction of the originally published system. Its performance should therefore be interpreted only as a controlled external baseline within this protocol. More broadly, our experiments do not establish universal superiority of SIG, a general failure of MHLR or BPR, or generalization beyond the present cohort and task. They instead show that, under the evaluated conditions, aggregate retrieval gains alone provide limited evidence of robust cross-subject improvement.

8 Conclusion

Under strict subject separation, content-level candidate control, and multi-seed evaluation, additional model complexity did not yield uniformly reliable cross-subject EEG-to-text retrieval gains. Prototype-Only SIG showed a small positive aggregate gain but mixed subject–seed consistency, while Strict MHLR, Controlled BPR-style, and Ridge did not provide robust improvement over Graph-Free under the present protocol. Overall, a positive aggregate gain does not necessarily imply consistent improvement across unseen subjects and training trajectories, motivating the joint reporting of aggregate performance, subject-level effects, and stochastic replication.

Reproducibility Statement

The main text specifies the controlled cross-subject evaluation protocol, including the outer-subject roles, train-only normalization, canonical candidate construction, validation-only model selection, controlled model definitions, optimization settings, and aggregation rules. We additionally provide an anonymized code artifact containing the implementation, frozen experiment configurations, protocol and reproducibility documentation, focused tests, and paper-safe result artifacts needed to reproduce the reported comparisons. External datasets and pretrained model weights are not redistributed. Exact historical raw-to-processed preprocessing provenance is not claimed where it could not be independently verified.

AI Use Statement

AI-assisted tools, including large language models, were used to support manuscript drafting and editing, literature retrieval and discovery, implementation inspection, and the organization of experimental and reproducibility materials. The author independently verified the experimental configurations, numerical results, implementation details, evaluation protocols, and scientific claims, and takes full responsibility for the content of the paper.

References

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

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*, 5:180291, 2018. doi: 10.1038/sdata.2018.291.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Evaluating eeg-to-text models through noise-based performance analysis. *Scientific Reports*, 16:350, 2026. doi: 10.1038/s41598-025-29587-x.

Taida Li, Yujun Yan, Fei Dou, Wenzhan Song, and Xiang Zhang. Cross-subject generalization for eeg decoding: A survey of deep learning methods. *Progress in Biomedical Engineering*, 8(2), 2026. doi: 10.1088/2516-1091/ae65f0.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Niall McGuire and Yashar Moshfeghi. Towards brain passage retrieval: An investigation of eeg query representations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 613–623. Association for Computing Machinery, 2025. doi: 10.1145/3726302.3730097.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.393.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5350–5358, 2022. doi: 10.1609/aaai.v36i5.20472.

Dongrui Wu, Yifan Xu, and Bao-Liang Lu. Transfer learning for eeg-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19, 2022. doi: 10.1109/TCDS.2020.3007453.

Xiaoli Yang, Huiyuan Tian, Yurui Li, Jianyu Zhang, Shijian Li, and Gang Pan. Brain-cliplm: Semantic compression for eeg-to-text decoding. *Frontiers in Neuroscience*, 20:1899770, 2026. doi: 10.3389/fnins.2026.1899770.

Jinzhao Zhou, Yiqun Duan, Ziyi Zhao, Yu-Cheng Chang, Yu-Kai Wang, Thomas Do, and Chin-Teng Lin. Towards linguistic neural representation learning and sentence retrieval from electroencephalogram recordings. In *Proceedings of the 1st International Workshop on Brain-Computer Interfaces for Multimedia Understanding*, pp. 19–28. Association for Computing Machinery, 2024. doi: 10.1145/3688862.3689109.

A Detailed Evaluation Protocol

A.1 Outer Subject Roles

The controlled evaluation uses five biological subjects: ZAB, ZKH, ZKW, ZMG, and ZPH. For every outer fold, one subject is held out for final testing, one distinct subject is reserved for validation, and the remaining three subjects form the training set. The fixed cyclic assignment is

$$\begin{aligned}\mathrm{ZAB}_{\mathrm{test}} &\to \mathrm{ZKH}_{\mathrm{val}},\\ \mathrm{ZKH}_{\mathrm{test}} &\to \mathrm{ZKW}_{\mathrm{val}},\\ \mathrm{ZKW}_{\mathrm{test}} &\to \mathrm{ZMG}_{\mathrm{val}},\\ \mathrm{ZMG}_{\mathrm{test}} &\to \mathrm{ZPH}_{\mathrm{val}},\\ \mathrm{ZPH}_{\mathrm{test}} &\to \mathrm{ZAB}_{\mathrm{val}}.\end{aligned}$$

The remaining three subjects train the model in each fold. Consequently, every biological subject serves once as the outer test subject, while final unseen-subject evaluation remains separated from both parameter optimization and model selection.

A.2 Normalization

For each outer fold, channel-wise EEG normalization statistics are estimated exclusively from the three outer-training subjects. Statistics are computed over all valid, unpadded temporal observations. For channel c, let $\mu_c^{(m)}$ and $v_c^{(m)}$ denote the training-only population mean and variance. The normalization scale is

$$\sigma_c^{(m)} = \sqrt{\max\left(v_c^{(m)}, 0\right) + 10^{-6}}, \qquad \tilde{x}_{itc} = \frac{x_{itc} - \mu_c^{(m)}}{\sigma_c^{(m)}}.$$

The statistics are accumulated in `float64` and then frozen. The same training-derived transformation is applied to training, validation, and test EEG.

A.3 Model Selection

Graph-Free and Prototype-Only SIG select the checkpoint minimizing mean validation retrieval loss, with patience 5 and a maximum of 50 epochs. Their validation objective is a symmetric in-batch EEG↔text contrastive cross-entropy objective with margin 0.2, hard-negative weighting, and fixed scale 20, whereas production training uses a one-directional EEG-to-text objective with an additional FIFO text bank.

The controlled BPR-style model is selected using validation duplicate-safe InfoNCE. Ridge selects its regularization coefficient using validation mean $1 - \cos$ error. Strict MHLR uses fold-specific selection roles drawn exclusively from the outer-training subjects, so neither the outer validation subject nor the outer test subject is used to select its adaptation hyperparameters.

A.4 Canonical Retrieval Bank

The canonical retrieval bank contains $K = 400$ unique sentence strings. Candidates are constructed deterministically by retaining the first occurrence of each exact sentence string in raw record order. No stripping, lowercasing, Unicode normalization, tokenization-based matching, or sorting is applied. Each query is mapped to its candidate index by exact Python string equality.

For query i, candidates are ranked by cosine similarity between the EEG representation and the corresponding candidate-text representations. Unique-Sentence Top-k accuracy is

$$\text{Top-}k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[c_i \in \text{TopK}\left(\{s_{ij}\}_{j=1}^{K}\right)\right],$$

where c_i is the locked candidate index of the exact-matching sentence. With $K = 400$, uniform retrieval yields chance accuracies of $1/400 = 0.25\%$ for Top-1 and $5/400 = 1.25\%$ for Top-5.

The ordered candidate strings are additionally fingerprinted by joining them with byte separator `0x1e` and computing a SHA-256 digest. The same ordered candidate strings and locked query-to-candidate identities are used for all controlled comparisons.

B Model and Optimization Details

B.1 Graph-Free EEG Semantic Encoder

The Graph-Free encoder projects each 105-dimensional temporal EEG observation to 512 dimensions, adds fixed sinusoidal positional encoding, and applies a four-layer post-norm Transformer with eight attention heads, 2048-dimensional feed-forward layers, ReLU activation, and dropout 0.1. Padding is excluded with a mask. Masked temporal mean pooling produces $\bar{h}_i \in \mathbb{R}^{512}$, followed by a biased $512 \to 4096$ semantic projection and ℓ_2 normalization.

The implementation contains an additive centering term, which is the identity ($\delta = 0$) in production runs.

Text targets are frozen Meta-Llama-3-8B-Instruct representations. Final-layer token representations are pooled using attention-mask-aware mean pooling and the resulting 4096-dimensional targets are ℓ_2-normalized. During EEG training, these representations are loaded from cache.

Production optimization uses one-directional EEG-to-text cross-entropy over the current cached targets together with a detached 1,024-entry FIFO text bank. Positive logits use margin 0.2, the top 10% negatives receive a $\log 1.2$ bias, and logits use fixed scale 20. No MSE, cosine-regression, or symmetric text-to-EEG training term is used.

B.2 Prototype-Based Subject Interaction

Prototype-Only SIG projects each Graph-Free semantic representation to $r_i \in \mathbb{R}^{256}$. For each subject group a represented in a minibatch, its prototype is

$$p_a = \frac{1}{|I_a|}\sum_{i \in I_a} r_i.$$

Normalized prototypes interact through cosine-based attention. Self-edges are retained and no graph temperature is used. The resulting subject context is projected back to 4096 dimensions and fused with the Graph-Free representation through the learned scalar gate defined in the main text.

Numeric subject labels only define within-batch groups; there is no running prototype memory or trainable subject-ID lookup.

Prototype-Conditioned additionally applies a residual $256 \to 64 \to 256$ GELU adapter to each prototype before graph construction.

B.2.1 One-Seed Prototype-Conditioned Ablation

Prototype-Conditioned is retained only as a supplementary architectural ablation because it was evaluated under one seed. It reaches $0.357 \pm 0.143\%$ Top-1 and $2.750 \pm 0.315\%$ Top-5 across five

subject folds. Its Top-5 is comparable to Prototype-Only SIG's three-seed mean, but the substantially lower Top-1 accuracy and absence of stochastic replication do not support a stronger claim.

B.3 Multi-Head Low-Rank Subject-Level Adaptation

MHLR operates on frozen 4096-dimensional Graph-Free representations using eight independent 512-dimensional residual low-rank heads. For each head, $D_r \in \mathbb{R}^{q\times 512}$ is Kaiming-uniform initialized and $U_r \in \mathbb{R}^{512\times q}$ is zero initialized. The resulting projector is therefore exactly the identity at initialization.

The eight adapted heads are concatenated directly. No activation, bias, normalization, or post-head projection is applied inside MHLR. Only the projector parameters are optimized during unsupervised subject-level transductive adaptation.

Fold-specific ranks and adaptation coefficients are selected by nested selection among outer-training subjects and are frozen before adaptation to the outer test subject.

B.4 Controlled BPR-Style Baseline and Ridge Control

The controlled BPR-style model is an adapted dense EEG–text dual encoder inspired by Brain Passage Retrieval (BPR) (McGuire & Moshfeghi, 2025), not an exact reproduction of the original passage-retrieval system. Its EEG branch uses a biased $105 \to 512$ projection, a learned zero-initialized CLS token, learned positional embeddings, and a three-layer post-norm Transformer (8 heads, 2048-D FFN, ReLU, dropout 0.1). Final CLS is projected to 768 dimensions and normalized.

The text branch caches frozen `distilbert-base-uncased` (Sanh et al., 2019) CLS representations and passes each through a trainable one-layer 768-D, 8-head Transformer adapter. Training uses a one-directional duplicate-safe multi-positive variant of the InfoNCE objective (van den Oord et al., 2018) with temperature 0.07, treating all exact-equal sentence strings in a minibatch as positives.

Ridge provides a deterministic linear capacity control. Normalized EEG is masked-mean pooled to 105 dimensions and mapped by centered affine Ridge regression to the same frozen 4096-D Llama targets as Graph-Free. The intercept is unregularized, the solve is performed in `float64`, and $\lambda \in \{0.001, 0.01, 0.1, 1, 10, 100\}$ is selected by validation mean $1 - \cos$ error with lower λ breaking exact ties. After selection, the model is refit on outer-training subjects only.

B.5 Optimization Details

Graph-Free and SIG use AdamW (Loshchilov & Hutter, 2019) with learning rate 3×10^{-5}, weight decay 10^{-4}, physical batch size 2, gradient accumulation over four minibatches, maximum 50 epochs, and patience 5; no scheduler or gradient clipping is used.

BPR uses AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-6}, weight decay 0.1, batch size 128, gradient clipping at global norm 1.0, and a 10-epoch linear warmup followed by linear decay through epoch 50.

MHLR uses batch size 4 and one adaptation pass; all folds select learning rate 10^{-4}, with ranks $(32, 16, 32, 32, 16)$ and adaptation coefficients $(1, 5, 1, 5, 1)$ for ZAB, ZKH, ZKW, ZMG, and ZPH, respectively.

Ridge uses the fixed regularization grid described above.