EEG DOES NOT SPELL WHAT YOU READ: SEMANTIC NEIGHBOURHOOD SYNTHESIS FOR OPEN-VOCABULARY EEG-TO-TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary EEG-to-text asks a model to reproduce, word for word, the sentence a person silently read, given only their scalp EEG. We argue that this target is mis-specified for reading and that the field's metrics have concealed the problem. Verbatim decoding partly succeeds when the task is *typing*, where every character is time-locked to a motor event, but silent *reading* offers no such clock, and on our evidence the exact word sequence does not survive the skull, though its meaning does. Indeed, on our benchmark the best BLEU-1 and ROUGE-1 belong to a system whose real-EEG performance never rises above its own information-free controls. We recast the problem as *semantic reconstruction* and introduce *Semantic Neighbourhood Synthesis*, which treats a trial as a region of meaning rather than a word sequence. An EEG→CLIP encoder retrieves the nearest training sentences, and a language model writes a new sentence from them. Under a uniform protocol with every baseline retrained from scratch, our method is best on nine of the ten metrics that score meaning, improving 24-way retrieval accuracy by 34% relative, and is the only method whose information-free controls fall to chance. We then ask what limits performance, and the answer is not the decoder. Interventions spanning the front-end, the decoder and the signal itself all shift the score by less than run-to-run noise, while the same readout separates subjects and, above all, sentences, with effects $7.4\times$ and $29\times$ that of the entire representation. On this evidence, reading EEG carries the gist of a sentence and very little of its wording.

1 INTRODUCTION

Reconstructing language from brain activity is a long-standing goal of brain–computer interfaces and, for the broader machine-learning community, a stringent test of whether learned representations capture meaning rather than surface form. Invasive neuroprostheses now restore large-vocabulary communication to people with paralysis (Moses et al., 2021; Willett et al., 2023; Metzger et al., 2023), and non-invasive systems have begun to recover *heard* language from fMRI, MEG and EEG (Tang et al., 2023; Défossez et al., 2023). Electroencephalography (EEG) is by far the cheapest and most portable of these sensors, which is why *open-vocabulary EEG-to-text decoding*, the task of generating arbitrary sentences from scalp EEG recorded while a person reads, has drawn sustained effort on the ZuCo reading corpora (Hollenstein et al., 2018; 2020) since Wang & Ji (2022) framed it as open-vocabulary generation. EEG is also the least forgiving of these sensors, mixing whole cortical populations through bone and scalp into a signal coarser, we will argue, than the field has assumed.

The field has nonetheless pursued *verbatim* transcription, and a validity problem has followed. For several early systems, among them EEG2Text (Wang & Ji, 2022) and DeWave (Duan et al., 2023), much of the reported quality turns out to rest on *teacher forcing*. When the ground-truth tokens are withheld at inference, output quality drops to that of a model with no access to the EEG (Jo et al., 2024). We read this not as a verdict on any one model but as a symptom. Verbatim transcription is the wrong target for this setting, and what rules it out is the reading task itself rather than the sensor alone. Decoding *typed* sentences from MEG and EEG in the same participants, Lévy et al. (2025) reach character error rates of 32% and 67% respectively, so verbatim brain-to-text is achievable in principle. Typing is what makes even that partial result possible. Every character fires one

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.

discrete motor event, which gives the decoder a clock against which to align signal and symbol. Natural reading provides no clock. No motor act marks a word, and the eye skips, refixates and regresses (Rayner, 1998; 2009); ZuCo's co-registered eye tracker logs fixations, but fixation order is not processing order and no such event stream exists in the EEG itself. Reading is therefore harder than typing. The word-for-word sequence a reader fixated does not survive the skull. Its *meaning* does.

We therefore recast the task from verbatim transcription to *semantic reconstruction*, recovering what was read at the level of meaning, in the spirit of semantic decoding from fMRI (Tang et al., 2023) and of EEG work that treats decoding as semantic summarization (Liu et al., 2025). Correcting the target changes what the representation must capture. Chasing exact words invites aligning the EEG, sample by sample, with a word sequence—an alignment neither available nor needed once meaning is the target. We therefore adopt a deliberately plain front-end, representing each trial by band power pooled over windows of several lengths, fixed and non-learned, claiming nothing about when any word was read. Viewed this way, the signal resolves into a region rather than a string: *EEG is most reliably read as a semantic neighbourhood*. We surface that neighbourhood with an EEG→CLIP retriever, which collapses several *machine-activation* spaces—internal representations of pretrained language and embedding models—into CLIP's vision–language *gist space* and retrieves there, rather than aligning the brain signal to a language model's text embeddings as EEG-to-text decoders usually do. A CLIP embedding captures meaning rather than exact wording; discarding the wording would be a liability for verbatim transcription but costs little here.

Semantic Neighbourhood Synthesis. A neighbourhood is not yet a sentence, and each prevailing paradigm closes half the gap. Retrieval-and-attribute methods such as SemKey (Wang et al., 2026) stay faithful to the decoded meaning but emit only phrasings the corpus already contains, while generation methods such as GLIM (Liu et al., 2025) phrase fluently but drift toward generic sentences. We unify the two around the neighbourhood in four stages, retrieving the nearest training sentences, extracting consensus keywords and a length cue, synthesizing candidates with an instruction-tuned language model, and selecting among them with EEG-predicted embeddings. Retrieval fixes *what* was read, and synthesis gives it words that need not appear in any corpus. Because fluent-but-ungrounded text is the central risk in this literature, we gate every claim with information-free controls. Real-EEG decoding must beat its shuffled- and noise-EEG counterparts, or we credit the result to the language prior.

Contributions. We correct the target, build the system it implies, and then show that the system's design is not what bounds it.

1. **We correct the target and audit the metrics that hid the problem.** Under information-free controls, a model that never beats its own controls still achieves the best BLEU-1, ROUGE-1, WER and METEOR, while retrieval accuracy proves to be the most signal-derived metric for every EEG-using method and BERTScore the least (§6).

2. **We build the system the corrected target implies.** On *identical* held-out trials, with every baseline retrained from scratch and one locked scorer, *Semantic Neighbourhood Synthesis* wins nine of the ten metrics that score meaning, improving 24-way retrieval accuracy by 34% relative, and is the only method whose controls fall to chance. It wins every converged split of an 11-split sweep, its gain comes from synthesis rather than reranking, and a neighbourhood control confirms that the EEG enters at the retrieval stage (§5, §7).

3. **We show that the method is not the constraint.** Interventions spanning the front-end, the decoder and the signal itself all leave the score inside run-to-run noise, including deleting the eight-band filterbank our own front-end is built on, while the same readout resolves subject and sentence identity at $7.4\times$ and $29\times$ the effect of the entire representation (§7).

4. **We identify what does vary: the sentence.** In a variance decomposition over identical trials, *what* was read accounts for most of the explained variance, who read it a minor share, and the choice of representation almost none. Some sentences are reliably recoverable from reading EEG and others are not—an axis this literature does not currently report.

2 Related Work

What non-invasive sensors can decode. Among non-invasive modalities the picture is sharply graded. Tang et al. (2023) recover continuous *meaning* rather than exact words from fMRI, describing their output as a paraphrase of the stimulus. Lévy et al. (2025) provide the most direct sensor comparison, the typing study discussed in the introduction: character-level decoding proves reasonable for MEG and marginal for EEG, and it is the typed paradigm that supplies the signal–symbol alignment making it well posed.

EEG-to-text generation and the validity crisis. Open-vocabulary EEG-to-text was popularized by EEG2Text (Wang & Ji, 2022), which adapts a pretrained sequence-to-sequence model to decode ZuCo reading EEG, and extended by DeWave (Duan et al., 2023) via discrete codex representations. Jo et al. (2024) then showed that removing teacher forcing collapses several such systems to a language-prior baseline, and argued for information-free controls, a critique we reproduce inside our protocol (§6) and adopt as a design constraint. GLIM (Liu et al., 2025) is the strongest prior work by that standard, adding a contrastive EEG–text alignment stage, evaluating under free-running generation, and genuinely using the EEG signal. Its failure mode is repetition: it often emits the same sentence across different trials (§5). Our comparison nonetheless treats it as a strong, fairly evaluated baseline.

Content decoupling and retrieval-augmented generation. SemKey (Wang et al., 2026) decouples *what* is read (keywords, sentiment, topic, length) from *how* it is phrased, predicting attributes from EEG and conditioning a constrained generator on them. We adopt the content-decoupling idea but read the cues off a retrieved neighbourhood rather than predicting them attribute by attribute, making the method an instance of retrieval-augmented generation (Lewis et al., 2020b) for brain decoding, with EEG as the query and an EEG→CLIP encoder as the retriever, aligned to vision–language embeddings (Radford et al., 2021; Zhai et al., 2023) and sentence encoders (Reimers & Gurevych, 2019; Xiao et al., 2024).

3 Method

Overview and notation. Let x be the EEG recorded while a subject reads a held-out sentence $y^\star$, and let $\mathcal{C}$ be a frozen corpus of training and validation sentences. No test sentence is in $\mathcal{C}$. The pipeline maps x to a sentence $\hat{y}$ that need not appear in any corpus, in four stages (Figure 1). Stage 1 retrieves a semantic neighbourhood, Stage 2 decodes content cues from it, Stage 3 synthesizes candidate sentences, and Stage 4 selects among them with EEG-predicted embeddings. Wherever the pipeline reads the EEG (Stages 1 and 4), it does so through a fixed, non-learned front-end (§3.1) that we describe first.

3.1 A spectral EEG front-end

Rather than model the raw EEG time series, which is nonstationary and low in signal-to-noise ratio sample by sample, we apply a fixed, non-learned *spectral front-end* that chains three operations (Figure 4, Appendix A). A *band-pass filterbank* splits each channel into the eight ZuCo bands, a *Hilbert envelope* replaces each band's oscillation with its slowly varying amplitude, the *band power*, and *multi-window framing* slides windows of five lengths (0.5–8 s, 100 ms stride) over the envelopes, pooling each window into one 840-d token tagged with its length. A trial is thus a variable-size *token set* $Z(x)$ (Appendix A). This is in effect a speech front-end recipe (Davis & Mermelstein, 1980; Radford et al., 2023), with two deliberate departures. Frames span multiple window lengths, because reading durations vary widely and no single window fits all trials. And the frames are consumed as a *set*, because EEG affords no groundable frame-to-word correspondence while readers skip, refixate and regress (Rayner, 1998; 2009), so ordering the frames would assert a temporal structure the measurement cannot support.

Most of this structure turns out not to matter: §7 shows that only the multi-window ladder earns its place. We note it here because the inertness of a carefully motivated representation is itself one of the paper's findings.



Figure 1: **Semantic Neighbourhood Synthesis.** From the band-power token set of a reading trial, **Stage 1** maps the tokens into CLIP space through seven fused pathways and retrieves the k=8 nearest training sentences, the semantic *neighbourhood*. **Stage 2** reads consensus keywords and a target length off that neighbourhood. **Stage 3** prompts an instruction-tuned language model to synthesize candidate sentences, each from a differently ordered prompt, and **Stage 4** reranks the candidates against EEG-predicted embeddings. Every prediction is gated against information-free controls (§6).

3.2 Stage 1: EEG→CLIP retrieval (grounding)

Stage 1 grounds the pipeline, predicting where in CLIP's text-embedding space the read sentence lies and returning the corpus sentences nearest that prediction. The readout is an ensemble of seven *pathways*, each mapping the token set $Z(x)$ to a point in CLIP space through a set transformer $f^{(p)}$ followed by a projection head, $\hat{e}^{(p)}(x) = g_\phi^{(p)}(f^{(p)}(Z(x))) \in \mathbb{R}^{768}$. Pathway p=1 is aligned directly to CLIP, so $g_\phi^{(1)}$ is the identity, while the *routed* pathways p=2, . . . , 7 use a learned $g_\phi^{(p)}$. The set transformer pools its tokens by a masked mean, so its output is insensitive to the number of windows present.

Every pathway is trained toward $\psi(y)$, the CLIP text embedding of the read sentence, by a grouped InfoNCE term plus cosine anchors (Eq. (1), Appendix B). The grouping ranks each prediction against the batch's *unique* sentences, so the several recordings of one sentence act as mutual positives and reinforce one sentence-level estimate. Each routed pathway carries an extra anchor tying the encoder's pre-map output to model p's own activation $\phi_p(y)$, so that $g_\phi^{(p)}$ maps a genuine machine representation into CLIP.

The seven pathways are combined only at inference, and only at the score level. A corpus sentence c from the frozen corpus $\mathcal{C}$ receives $S(x, c) = \sum_p \cos(\hat{e}^{(p)}(x), \psi(c))$, and the neighbourhood $N(x)$ collects the k=8 sentences of largest S, so a sentence scores high only when several pathways agree it is near. Two choices explain the design, and both are validated by downstream accuracy in Appendix F. We retrieve in CLIP's compact *gist* space rather than in each target's own space, because the coarse EEG signal locates a neighbourhood far more stably there. And we route through machine-activation spaces because the most EEG-decodable target is not CLIP text but an intermediate LLM activation.

3.3 STAGES 2–4: CUES, SYNTHESIS, AND SELECTION

Stage 2: content cues (decoupling). From the neighbourhood we read off two cues. Consensus *keywords* are the content (non-stopword) words present in at least two neighbours (Appendix B), and the target *length* is the median word count of the neighbours. Following the content-decoupling principle, these summarize *what* was read without fixing *how* it is phrased.

Stage 3: candidate synthesis (generation). We prompt an instruction-tuned language model (Qwen3-4B) with the neighbours $N(x)$, the keywords and the target length, and draw $M{=}16$ candidates. Because sampling all M from one prompt yields less diverse candidates, we build M prompts differing only in the *order* the neighbours are listed, each by one random transposition, a minimal and content-preserving perturbation. One sentence is sampled per prompt by nucleus sampling, and the prompt asks for a single sentence matching the neighbours' topic, sentiment and style without copying any of them (template in Appendix B). Because generation is not corpus-bound, a candidate may capture $y^\star$ more closely than any neighbour, and §7 locates the gain over retrieval in this step.

Stage 4: EEG-driven selection. We embed each candidate in three spaces, CLIP, a sentence encoder (MPNet) and a retrieval embedder (BGE), and for each space E obtain an EEG-predicted target $\hat{e}_E(x)$, the fused Stage-1 prediction for CLIP and a dedicated EEG$\rightarrow E$ set transformer for the other two. The selected output is $\hat{y} = \arg\max_{c\in\{c_1,\dots,c_M\}} \sum_E z\big(\cos(\phi_E(c), \hat{e}_E(x))\big)$, where $\phi_E(c)$ is the candidate embedding and $z(\cdot)$ standardizes the cosine scores across a trial's candidate pool. The argmax is over the *synthesized* candidates only, so $\hat{y}$ is always a newly generated sentence. The retrieved neighbours enter the pool solely to calibrate z. §7 shows this is the pipeline's weakest stage.

4 EXPERIMENTAL SETUP

Data and strict identical-instance protocol. We use the ZuCo 1.0 and 2.0 reading-EEG corpora (Hollenstein et al., 2018; 2020), pooling their four reading tasks (ZuCo 1.0 SR/NR/TSR and ZuCo 2.0 NR) into one pool of 1456 sentences and drawing a single seeded $70/10/20$ split at the *sentence* level, which holds out 291 sentences. Because each sentence is read by many subjects, a test *trial* is one $(\text{subject}, \text{sentence})$ pair, so the held-out sentences expand to 3817 candidate trials. Each baseline preprocesses the raw EEG differently and, per its own official recipe, discards noisy or incomplete trials, so each natively serves a *different* subset. We therefore reproduce every method's official preprocessing *exactly*, retrain each from scratch under a hard leakage guard, and score every method only on the trials *all* of them can serve. Intersecting their native coverages and deduplicating to one prediction per pair gives $N{=}3166$ *identical* held-out trials, and automated audits abort any run that violates the protocol (Appendix C). The shared trials cover 275 of the 291 held-out sentences; the other 16 lose every trial to some baseline's filter. Because our front-end imposes no length cap, the shared set is determined mainly by the baselines' filters.

Information-free control panel. Every method is evaluated under four scenarios that perturb only the EEG input while holding the targets fixed, using a shared seeded permutation π: (a) REAL; (b) SHUFFLE, a donor trial's real EEG; (c) NOISE·KEEPLEN, noise EEG of own length; and (d) NOISE·SHUFLEN, noise EEG of donor length. A control must perturb *every* non-EEG side channel a model reads. SemKey conditions on a textual attribute prompt, so the controls also derange those attributes, which would otherwise silently leak the target's true attributes.

Identical evaluation. Every method's *native free-running* generations are scored by one locked harness ported from SemKey's published evaluation code (Wang et al., 2026) and verified to reproduce it exactly (re-implemented wrappers proved numerically unfaithful; Appendix D). We report SemKey's complete metric suite, so that no method is judged on a family chosen after the fact. The suite comprises **retrieval accuracy** (2/4/10/24-way, all-MPNet-base-v2, $10{,}000$ trials, fixed seed), **n-gram overlap** (BLEU-1–4, ROUGE-1, WER, METEOR), **semantic similarity** (BERTScore (Zhang et al., 2020) with RoBERTa-large, SBERT cosine, CIDEr, content-word recall, length MAE, Fréchet distance), and **richness/diversity** (dist-1/2, self-BLEU, head entropy; Table 6, Appendix E). Following SemKey, BLEU uses its eight paraphrase variants where they are available (96.9% of targets; Appendix E), with single-reference BLEU reported alongside for comparability

with EEG2Text and GLIM; Table 1 shows ten semantic and six verbatim columns, with BLEU-2/3 in Appendix E. GLIM's native learned-space metric is not comparable across methods, so its generations are re-scored through the shared harness. These families disagree sharply (§5), and which one is trusted decides the ranking.

Baselines. All baselines are retrained from scratch on the canonical training fold with their official code and recipes, and evaluated from native free-running generation: GLIM (Liu et al., 2025) (contrastive EEG–text encoder + BART (Lewis et al., 2020a)), SemKey (Wang et al., 2026) (attribute-conditioned Flan-T5 (Chung et al., 2024), in both its predicted-attribute and ground-truth-attribute modes), and EEG2Text/BrainTranslator (Wang & Ji, 2022). Ours uses a Qwen3-4B synthesizer (Qwen Team, 2025).

5 RESULTS

Our method wins nine of the ten metrics that score meaning and none of the six that score verbatim overlap. That split, not any single margin, is the central result of this section. Table 1 reports the five-method comparison on the $N{=}3166$ identical held-out trials. EEG2Text scores near chance under free-running evaluation, reflecting the well-documented teacher-forcing effect rather than a limitation of the original contribution (§6), yet wins five of the six verbatim metrics. SemKey's predicted-attribute mode, the protocol its paper reports, slightly *exceeds* its ground-truth-attribute counterpart, with heavily overlapping intervals (Table 5), so we find no evidence of an oracle advantage. The margin is not a single-split artifact. In an 11-split, 3-seed sweep that retrains SemKey end-to-end for every split and seed, our method wins all ten converged splits with a mean gap of $+3.59$ points (Appendix D). It also survives a sentence-clustered bootstrap that scores both methods on the same resampled sentences (Table 5), where the paired 24-way difference is $+4.62$ points $[+3.04, +6.15]$ with our method ahead in all 2000 replicates.

Semantic reconstruction: a clean sweep. Across the ten metrics that score *meaning*, our method is best on nine, with margins measured against the *best* baseline on each metric. The largest gains are $+34\%$ relative on 24-way retrieval (18.31 vs. 13.68), comparable relative gains on content-word recall and CIDEr, and a 28% smaller length error (Table 1). The one exception is Fréchet distance, which compares distributions rather than scoring each trial against its own sentence, so it rewards a well-shaped corpus as much as decoding any individual reading.

Verbatim overlap measures something else, and something unreliable. The second block behaves entirely differently. We win none of it, and the winner of five of its six columns is EEG2Text, the one method whose real-EEG score never exceeds its own controls, that is, the one method whose output shows no detectable dependence on the brain signal. The mechanism is measurable. Over 94% of its output tokens are stopwords, against about 38% for the references and for ours, yielding the highest unigram precision of any method alongside zero content-word recall (Appendix E). It has collapsed onto a family of function-word strings, e.g. *"He was the a as in, was of and's from"*, and function words match any English reference, reproducing the "BLEU trap" of Wang et al. (2026) inside one controlled protocol. One system gaming BLEU-1 does not by itself condemn the block. §6 does, by measuring how little of each n-gram metric survives destroying the EEG. We therefore read the block as uninformative about decoding rather than as a deficit, and it is the block our thesis predicts we should lose: our output paraphrases, sharing the reference's meaning but rarely its exact tokens.

Output diversity diagnoses collapse rather than quality. Diversity is not a quality axis here. The gold references themselves score *lower* on dist-2 and *higher* on self-BLEU than any system, because the evaluation set covers only 275 distinct sentences. What diversity does diagnose is *collapse*, EEG2Text's function-word template and GLIM's whole-sentence repetition, while our output avoids both failure modes with the best content-word recall and smallest length error of any method (Table 6, Appendix E).

Table 1: Five-method comparison on the N=3166 identical held-out reading trials (real EEG, one locked harness, all baselines retrained from scratch, native free-running generation). **Semantic reconstruction** metrics compare meaning; **verbatim overlap** metrics compare literal token strings. Best per column in **bold**, respecting each metric's direction. †: single-reference BLEU-1 (the protocol EEG2Text and GLIM use); other BLEU columns use SemKey's eight paraphrase variants. ‡: EEG2Text's real-EEG score never exceeds its own information-free controls (§6).

Semantic reconstruction *(does it mean the same thing?)*

Method	24-way	10-way	4-way	2-way	BERTSc.	SBERT	CIDEr	cRec	lenMAE↓	FD↓
EEG2Text	4.33	9.98	24.82	49.37	0.816	0.128	0.019	0.000	8.99	1.272
GLIM	12.49	24.77	46.65	70.68	0.836	0.231	0.029	0.020	10.19	0.454
SemKey (GT-attr)	12.87	25.07	45.82	69.35	0.838	0.206	0.031	0.021	9.35	0.302
SemKey (predicted)	13.68	25.84	46.69	70.08	0.843	0.213	0.033	0.021	8.89	**0.297**
Ours	**18.31**	**31.82**	**53.03**	**74.48**	**0.851**	**0.258**	**0.043**	**0.029**	**6.37**	0.329

Verbatim overlap *(does it use the same literal words?)*

Method	BLEU-1	BLEU-4	BLEU-1†	ROUGE-1	WER↓	METEOR
EEG2Text‡	**33.14**	0.00	**11.55**	**24.21**	**0.969**	**0.143**
GLIM	21.89	**0.94**	8.35	12.64	1.157	0.102
SemKey (GT-attr)	19.59	0.68	8.54	12.25	1.192	0.108
SemKey (predicted)	20.68	0.63	9.02	12.87	1.165	0.111
Ours	20.61	0.29	9.45	13.54	1.040	0.111

Metrics won **Ours**: **9/10** semantic, 0/6 verbatim EEG2Text‡: 0/10 semantic, **5/6** verbatim

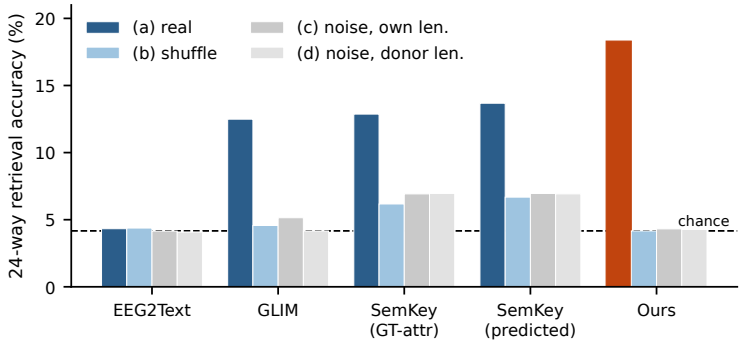

Figure 2: **Information-free control panel** (24-way, free-running, N=3166, one locked harness). For every method, the real-EEG bar is shown against three controls that destroy EEG information while preserving the language prior. Ours is the only method whose controls all fall to chance (≤ 4.31 vs. chance ≈ 4.2); SemKey's floor is elevated by its attribute prompt, and EEG2Text's real bar is indistinguishable from its controls.

6 Validity and Information-Free Controls

Removing teacher forcing returns EEG2Text's real-EEG score to chance, indistinguishable from its own noise controls, reproducing Jo et al. (2024)'s measurement inside our exact protocol; under teacher forcing it matches its originally reported BLEU-1, confirming a faithful reproduction.

An information-free control panel for every method. Figure 2 shows every method under the four-scenario panel (full table in Appendix D). For EEG2Text the real-EEG and control bars coincide, the teacher-forcing signature noted above, while GLIM and SemKey show clear positive gaps and genuinely use the signal. Two properties single out our method: by far the largest real-to-control gap, fourteen points of 24-way accuracy, and the *only* controls that all fall to chance. SemKey's controls are *elevated* by its textual attribute prompt, which retrieves above chance with no usable EEG, so roughly half of its real-EEG score sits on its own non-EEG floor. Deranging that prompt barely moves the floor, so the elevation comes from the prompt itself rather than from a leak of the target's true attributes.

How much of the gain is corpus identity? The shuffle and noise controls cannot detect a decoder that reads only *which corpus* a trial came from, and ZuCo pools two topically disjoint stimulus sets, movie reviews and Wikipedia. Re-scoring every method with the 24 distractors drawn from the target's *own* reading task (Table 7, Appendix E) shows corpus identity is a real component for every EEG-using method, and it is what SemKey's prompt encodes: its elevated floor falls to chance within task. Our own within-task score of 11.12 still retains a $+6.91$ gap over its control floor, and the ranking is unchanged. We report the all-task number as the headline and the within-task number as the more conservative measurement.

Which metrics survive an information-free input. Running the whole metric panel under the controls shows why the choice of headline metric matters. When our method's EEG is replaced by a donor trial's, 24-way accuracy collapses essentially to chance, while BLEU-1 retains about four fifths of its value, a language-prior floor that survives the loss of all EEG information. Table 8 (Appendix E) makes this comparable by asking what fraction of each metric's value is attributable to the EEG. For our method the ordering runs from 24-way retrieval at 76% down to BERTScore at about 1%, and the same ordering holds for GLIM and both SemKey modes, so retrieval is not a metric chosen because we win it. **BERTScore is therefore a poor validity discriminator**, measuring mostly fluent in-domain English, and we use it only as an output-quality check whose embedder is disjoint from the selector.

7 Ablations

The ablations ask whether the window ladder is necessary, where the gain lives, and what stops the score from going higher. Unless noted, they select settings on the validation fold and report on the full test fold, with a single-prompt synthesis variant (Appendix F), so absolute values sit near 17.8 rather than 18.31 and serve as internal comparisons. Rows whose generation RNG is not fixed vary by about ± 0.35, which we treat as the grid's resolution. Full tables are in Appendix F.

The front-end: the window ladder earns its place, the filterbank does not. Only 59.9% of reading trials are long enough to supply the single best-tuned 4 s window, so a single-window retriever cannot serve the all-length set the raw-EEG baselines use (Figure 5). The five-length ladder reaches full coverage and is also more accurate on the trials both can serve, by $+1.40$ 24-way points over three seeds (p=0.027). The *spectral decomposition*, by contrast, is inert. Replacing the eight-band filterbank with a single broadband envelope moves 24-way accuracy by 0.21 points (p=0.52) with eight times fewer input dimensions, and no individual band is load-bearing. We built the front-end around this spectral decomposition, and we report its inertness as a negative result about our own design.

Where the gain comes from. Pure retrieval, emitting the top-1 neighbour, reaches 16.32, and adding Stage-4 reranking without synthesis *lowers* this to 15.52 (Figure 3b,c; Table 12), so the EEG selector cannot improve on the retriever's own ranking. Synthesis is what moves the system, lifting 24-way accuracy by 1.49 points to 17.81 and BERTScore with it, a gain registered by an embedder disjoint from the retrieval metric. The reconstruction step, not the selection step, is the engine. Stage 4 itself is the weakest stage: against the right null, no selection at all, it barely registers, and the three-encoder selector's half-point edge over a CLIP-only one leaves BERTScore unchanged to three decimals. Since one selector encoder is the very embedder the 24-way metric uses, we report the selector as a metric-alignment caveat rather than a contribution (Appendix F).

The neighbourhood must come from the EEG. To test whether a language model prompted with any plausible in-domain cues would score as well, we hold Stages 2–4 unchanged and replace only the Stage-1 neighbourhood on the N=3166 set. A *random* neighbourhood of eight corpus sentences per trial drops 24-way accuracy from 18.33 (this seed's full-system value) to 5.43, and a *central* neighbourhood, the eight sentences nearest the corpus-mean CLIP embedding, drops it to 4.38 against a chance of 4.17. The synthesizer therefore contributes essentially no decoded content on its own: what the system knows enters through the neighbourhood, where Stage 1 reads the EEG.

Neither the decoder nor the amount of EEG is the constraint. Every lever we can find downstream of the front-end is inert (Appendix F): the candidate count over a 64-fold range, synthesizer

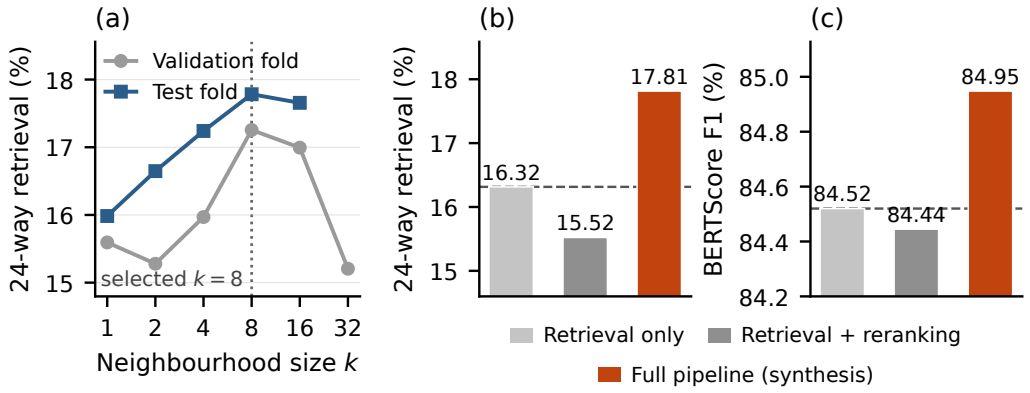

Figure 3: **Where the gain comes from.** (a) 24-way retrieval accuracy against the neighbourhood size k, which peaks at the selected k=8 on both folds, so the validation choice transfers. (b, c) Pipeline decomposition on the test fold (single-prompt variant; Table 12, Appendix F). Relative to the top retrieved neighbour (dashed line), reranking *lowers* both metrics and synthesis lifts both, locating the gain in the synthesis stage.

scale across ten models from 0.6B to 32B parameters, and Stage-1 encoder capacity across five variants. Among these levers the one hyperparameter that matters is the neighbourhood size k (Figure 3a). Nor is the amount of signal per trial the limit: decoding quality is flat across a twentyfold range of recording duration, and pooling independent readings of the *same* sentence recovers almost nothing while a control pooling readings of *different* sentences correctly degrades, so the limiting variability is shared across readings rather than independent noise that averaging removes, and per-trial information is close to exhausted.

These nulls are measurements, not a blunt instrument. One might object that the readout is simply too blunt. On the *same* trials, however, the spread of its per-trial score across subjects spans a range $7.4\times$ larger than the entire representation effect, and across sentences a range $29\times$ larger. The choice of representation accounts for 0.5% of explained variance, against 13.4% for who was reading and 86.1% for what they read. The sentence effect reflects decoding rather than guessable targets, sharing only 1.3% of its variance with an EEG-free control. Some sentences are reliably recoverable from reading EEG and others are not, by a margin that dwarfs anything architecture does—an axis we argue the field should begin to report.

8 Limitations and Conclusion

Absolute quality remains meaning-level, and a 24-way accuracy of 18.31 is far from usable communication. We regard this as a property of the sensor and the paradigm rather than the method, since the same limits hold Lévy et al. (2025) to a 67% character error rate even when typing supplies alignment. The methods consume different EEG features, entangling front-end differences with method differences, and the multi-split sweep covers only SemKey against ours. Our SemKey reproduction sits about one point below its reported number, and the conclusion survives granting it the full gap (Appendix D). Our headline metric embeds with MPNet, which is also a Stage-4 selector encoder and a Stage-1 pathway target; we mitigate this by reporting BERTScore throughout and by discounting single-metric effects. Finally, subjects are pooled and sentences held out, so we do not establish cross-subject generalization.

EEG does not spell what you read. Without typing's motor clock, the verbatim target is misspecified, and so are metrics that an information-free model can win outright. Semantic Neighbourhood Synthesis, built for the corrected target of *semantic reconstruction*, wins nine of the ten meaning metrics with controls at chance. The constraint is neither the decoder nor the representation but what a trial contains, and progress on scalp EEG will come from what the sensor can be made to reveal, not from what a larger decoder can be made to say.

Reproducibility statement

All information needed to reproduce our results is specified in the paper. §4 defines the data, the sentence-level split, the shared test set, the information-free control panel, and the metric suite. Appendix A gives the front-end with exact tensor shapes, Appendix B the architecture, training recipe, and verbatim synthesis prompt, and Appendix C the per-method coverage filters and the six audited fairness invariants. Appendices D–F report the complete result tables behind every claim. The ZuCo corpora are publicly available from their original authors. We will release the full code, including the evaluation platform, all five retrained method implementations, the frozen splits, and the scripts that regenerate every table and figure, during the discussion period.

Ethics statement

This work uses the publicly released ZuCo 1.0 and 2.0 corpora, collected and distributed by their original authors with participant consent and institutional approval. We performed no new human-subject data collection. The corpora are de-identified, and we neither attempt nor report any subject-identification analysis. Brain-to-text decoding raises mental-privacy concerns, and we regard over-claiming as the primary near-term ethical risk in this area: reporting decoding quality that in fact comes from a language prior can misrepresent to the public what brain recordings reveal. Our information-free control panel, and our decision to report and highlight the negative results in §7, are directly motivated by that concern. We emphasize that the system requires a cooperative subject wearing an EEG cap during deliberate reading, recovers only meaning-level content at 18.31% 24-way accuracy, and does not constitute a means of reading unwilling subjects' thoughts.

AI use statement

We used generative AI tools as coding assistants for implementation, refactoring, and analysis scripting within the released codebase, and for copy-editing prose in this manuscript. Generative AI was not used to generate research ideas, to produce experimental results, or to write the paper's claims independent of author verification. All AI-assisted code was reviewed and validated against the regression tests included in the released code, and every number reported in this paper is regenerated programmatically from result artifacts rather than transcribed by hand. We note separately that an instruction-tuned language model (Qwen3-4B) is a *component of the method under study* at Stage 3, described in §4, and is not an authoring aid. We have reviewed all AI-assisted work and take responsibility for the final content of this paper.

References

Hyung Won Chung, Le Hou, Shayne Longpre, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25:1–53, 2024.

Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

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

Yiqun Duan, Charles Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. DeWave: Discrete encoding of EEG waves for EEG to text translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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(1):180291, 2018.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pp. 138–146, 2020.

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

Jarod Lévy, Mingfang Zhang, Svetlana Pinet, Jérémy Rapin, Hubert Banville, Stéphane d'Ascoli, and Jean-Rémi King. Brain-to-text decoding: A non-invasive approach via typing. *arXiv preprint arXiv:2502.17480*, 2025.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7871–7880, 2020a.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020b.

Xiaozhao Liu, Dinggang Shen, and Xihui Liu. Learning interpretable representations leads to semantically faithful EEG-to-text generation. *arXiv preprint arXiv:2505.17099*, 2025.

Sean L. Metzger, Kaylo T. Littlejohn, Alexander B. Silva, David A. Moses, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023.

David A. Moses, Sean L. Metzger, Jessie R. Liu, Gopala K. Anumanchipalli, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385 (3):217–227, 2021.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.

Keith Rayner. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457–1506, 2009.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

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. *arXiv preprint arXiv:2603.03312*, 2026.

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.

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

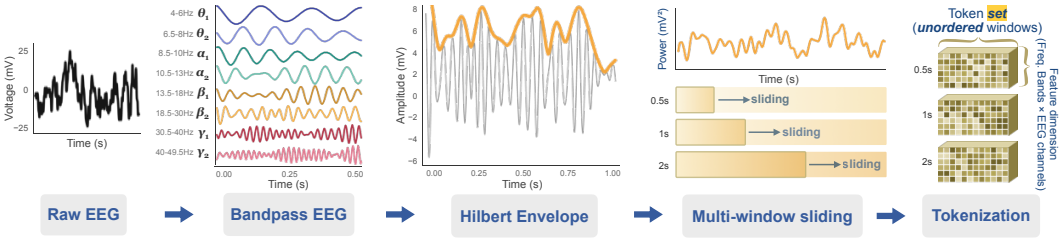

Figure 4: **The spectral EEG front-end.** A fixed, non-learned front-end turns one reading trial into the multi-window band-power *token set* $Z(x)$ consumed by Stage 1: 105 raw scalp channels at 500 Hz are split by a Butterworth filterbank into the eight ZuCo bands; per band the Hilbert envelope replaces the fast oscillation with its slowly varying *band power* (100 Hz); windows of five lengths (0.5–8 s, 100 ms stride) are mean-pooled over time; and each window becomes one 840-d token (8 bands $\times$ 105 channels) tagged with its window length. Exact tensor shapes are given in Steps 0–5 below.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packed resources for general chinese embeddings. *Proceedings of the 47th International ACM SIGIR Conference*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

A THE SPECTRAL EEG FRONT-END

This appendix details the fixed, non-learned front-end introduced in the Method: the full conversion from a raw reading trial to the multi-window band-power token set consumed by Stage 1, with tensor shapes at every step. Let $C{=}105$ channels, $B{=}8$ frequency bands, and the five window lengths $\{0.5, 1, 2, 4, 8\}$ s with a 100 ms stride. Throughout we use a 4.7 s reading as a running example.

Step 0: raw EEG. A trial is a continuous multichannel recording at 500 Hz: $\mathbf{x}_{\text{raw}} \in \mathbb{R}^{C\times T}$ with $T \approx$ duration $\times$ 500 (example $(105, 2347)$). This transient signal is never fed to the model directly.

Step 1: band-pass into eight bands. Each channel is filtered with a 4th-order Butterworth band-pass (zero-phase) into the eight standard ZuCo bands ($\theta_{1,2}, \alpha_{1,2}, \beta_{1,2}, \gamma_{1,2}$, spanning 4–50 Hz), giving $(B, C, T) = (8, 105, 2347)$ at 500 Hz.

Step 2: Hilbert envelope and downsample ("band power"). We take the Hilbert analytic amplitude of each band/channel—the smooth band-power curve—and downsample $500 \to 100$ Hz: $\mathbf{e} \in \mathbb{R}^{B\times C\times T'}$, $T' = \lfloor \text{duration} \times 100 \rfloor$ (example $(8, 105, 470)$). This envelope tensor is what is cached per trial; every later step is pooling.

Step 3: one window $\to$ one 840-d token. For window length L (in samples at 100 Hz) and each window start, we average the envelope over the window and flatten:

$$\text{seg} = \mathbf{e}[:, :, s{:}s{+}L] \in \mathbb{R}^{8\times 105\times L} \xrightarrow{\text{mean over time}} \mathbb{R}^{8\times 105} \xrightarrow{\text{flatten}} \mathbb{R}^{840}.$$

Thus one window is a single 840-dimensional band-power vector ($840 = 8 \times 105$).

Step 4: multi-window token set. Step 3 is repeated for all five window lengths; a window longer than the trial is dropped. For the running example ($T'{=}470$): 0.5 s ($L{=}50$) yields 43 tokens, 1 s ($L{=}100$) 38, 2 s ($L{=}200$) 28, 4 s ($L{=}400$) 8, and 8 s ($L{=}800$) none (dropped, $800{>}470$), for 117 tokens in total, each $\in \mathbb{R}^{840}$ and tagged with its window-length index $\in \{0, \ldots, 4\}$.

Step 5: fixed token set (the Stage-1 input). To batch the set transformer, K token slots per window length are kept, with the sampling deliberately asymmetric between training and inference. At *training* time we draw $K{=}48$ windows per window length *at random* each step (with replacement when a window length has fewer than 48 windows), a stochastic subsampling that regularizes the encoder toward subset-invariance and halves the attention cost. At *evaluation* time we keep $K{=}96$ windows per window length *deterministically*—all of them when a window length has at most 96, otherwise an evenly spaced (linspace) subset—for a reproducible, higher-fidelity representation. Slots for absent or under-filled window lengths are padded and masked. This asymmetry is safe because the encoder pools by masked mean, which is invariant to token count. The Stage-1 input per trial is therefore the fixed-size batched form of the token set $Z(x)$, namely the triple

$$\underbrace{\mathbf{F} \in \mathbb{R}^{5K\times 840}}_{\text{band-power tokens}}, \quad \underbrace{\mathbf{m} \in \{0,1\}^{5K}}_{\text{real/padding mask}}, \quad \underbrace{\mathbf{s} \in \{0,\dots,4\}^{5K}}_{\text{window-length index}},$$

i.e. $(480, 840)$ at evaluation (5×96), of which 117 tokens are real and 363 are masked for the example trial. Note there is no time axis and no raw voltage left: Stage 1 sees a masked *set* of frequency-band-power summaries at multiple window lengths. Featurization and encoder match the implementation in our code release exactly.

B Model Architecture and Training

This appendix gives full details for Stages 1–4. All encoders share one architecture and training recipe; the stages differ in their targets. We use a single fixed seed (42) unless a multi-seed result is reported.

Set-transformer encoder. Each encoder maps a masked token set (Appendix A) to a single vector. A token $\mathbb{R}^{840}$ is projected by LayerNorm $\to$ Linear($840{\to}256$) $\to$ GELU; a learned window-length embedding (an Embedding$(5, 256)$ indexed by the token's window length) and a small learned positional bias (one vector per token slot) are added; the sequence passes through a 2-layer Transformer encoder ($d{=}256$, 4 heads, feed-forward 1024, GELU) under the padding mask; the unmasked tokens are mean-pooled to $\mathbb{R}^{256}$; and a head LayerNorm $\to$ Linear($256{\to}d_{\text{out}}$) produces the prediction, with $d_{\text{out}}{=}768$ for CLIP targets and $\dim(E)$ for a sentence-embedding target E.

Fusion diagnostic. The ensemble quantities quoted in Appendix F come from a generation-free analysis of the released seven-pathway retriever (seed 42, on the 3358 trials of our native coverage): per-pathway cosine to the true CLIP embedding, pairwise prediction redundancy (0.67), pairwise *error* correlation (0.69), the fused-versus-best-single estimate (0.54 vs. 0.49), and the top-8 neighbourhood overlap between the two (0.34 Jaccard, 0.23 top-1 agreement). Because it is pure retrieval-stage linear algebra with no language-model sampling, it is deterministic and reproduces exactly.

What the positional bias learns. We probed the slot-indexed parameter directly. It is low-rank (participation ratio 36.3 of 240 trained slots), near-constant within each window-length block (within-block slot standard deviation is 0.27–0.76 of the block mean), aligns block-by-block with the window-length embedding ($|\cos| \approx 0.73$–0.86 on the diagonal), and carries only $\approx 1.5\%$ of that embedding's magnitude (0.238 vs. 15.7). It therefore re-encodes window-length identity rather than slot order. It is not inert, however: removing it costs 0.30 24-way for the fused retriever and 0.56 for the best single pathway (three seeds), so we keep it and describe the representation as order-agnostic rather than strictly permutation-invariant.

Objective and optimization. Let $\mathcal{L}_{\cos}(a,\, t) = \sum_{i\in\mathcal{B}} \big(1 - \cos(a_i,\, t(u_{\gamma(i)}))\big)$ denote the cosine alignment of a batch of predicted vectors a to a target map t, where $u_{\gamma(i)}$ is the unique sentence read in trial i (notation of Eq. (2)), and let $\phi_p(y)$ be the internal activation (a residual or feed-forward stream at a chosen layer) that routed model p produces for sentence y. The direct and routed pathways differ by a single term:

$$\mathcal{L}^{(p)} = \begin{cases} \mathcal{L}^{(1)}_{\text{NCE}} + \mathcal{L}_{\cos}(\hat{e}^{(1)}, \psi), & p = 1, \\ \mathcal{L}^{(p)}_{\text{NCE}} + \mathcal{L}_{\cos}(\hat{e}^{(p)}, \psi) + \mathcal{L}_{\cos}(f^{(p)}(Z(x)), \phi_p), & p > 1, \end{cases} \tag{1}$$

where the grouped InfoNCE term ranks each prediction against the batch's *unique* sentences $\mathcal{U}(\mathcal{B})$, with $\gamma(i)$ the index among them of the sentence read in trial i:

$$\mathcal{L}_{\mathrm{NCE}}^{(p)} \;=\; -\sum_{i\in\mathcal{B}} \log \frac{\exp\big(\cos(\hat{e}_i^{(p)},\, \psi(u_{\gamma(i)}))/\tau\big)}{\sum_{u\in\mathcal{U}(\mathcal{B})} \exp\big(\cos(\hat{e}_i^{(p)},\, \psi(u))/\tau\big)}. \tag{2}$$

The Stage-4 selectors are the direct case with the CLIP target ψ replaced by the selector's own sentence-embedding target. The learnable temperature τ is initialized 0.07 and clamped to $[0.01, 0.5]$. We optimize with AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}), batch size 128, for 40 epochs, drawing $K{=}48$ random tokens per window length each step.

Stage 1: the seven-pathway fused retriever. The retriever fuses seven encoders (Table 2). One is aligned directly to the CLIP text embedding. The other six each predict a different model's internal activation—the residual or feed-forward stream of a small/large instruction LLM at a chosen layer, a sentence encoder, or a second CLIP variant—which is then mapped into CLIP space by a small MLP head g_ϕ (LayerNorm $\to$ Linear($d\to1024$) $\to$ GELU $\to$ Linear($1024\to768$)); an extra cosine term anchors the encoder to the true activation before g_ϕ. At inference each pathway yields a predicted CLIP vector; we score every corpus sentence by cosine in each pathway and *sum the seven score matrices*, returning the top-8. The corpus $\mathcal{C}$ is the CLIP text embeddings of all training and validation sentences; the summed prediction, ℓ_2-normalized, also serves as the CLIP target in Stage 4.

Table 2: The seven fusion pathways (target model, layer/state). Pathway 1 aligns directly to CLIP; the rest predict an activation space mapped to CLIP by g_ϕ.

#	target model	layer / state
1	CLIP ViT-L/14 (direct)	text embedding
2	Gemma-3-1b	residual, layer 25
3	MPNet (all-mpnet-base-v2)	text embedding
4	Gemma-3-12b	residual, layer 39
5	Qwen3-4B	residual, layer 12
6	CLIP ViT-B/32	residual, layer 11
7	Gemma-3-1b	feed-forward, layer 18

Stage 2: content cues. Keywords are content words (length > 3, excluding a fixed stop list) that occur in at least two of the eight neighbours, keeping the six most frequent. The target length is the median word count of the eight neighbours. No EEG length regressor is used.

Stage 3: synthesis. The synthesizer is Qwen3-4B in bfloat16, called through its chat template (with "thinking" disabled). For each trial we build $M{=}16$ prompts that share the same neighbours, keywords, and target length but differ by a single random transposition of two neighbours in the listed order (seeded deterministically by trial index), and sample one candidate per prompt by nucleus sampling (temperature 0.9, top-p 0.95, up to 48 new tokens). The verbatim prompt template is:

> *A person read ONE sentence while their brain activity was recorded. Decoded from the brain signal, the sentences most semantically similar to it are: {numbered neighbours}. Key concepts detected in the signal: {keywords}. Reconstruct the single sentence the person most likely read. It should be about {length} words and match the topic, sentiment, register and style of the examples — but be its own sentence, not a copy. Output ONLY the sentence, nothing else.*

Stage 4: EEG-driven selection. The three selector spaces are CLIP ViT-L/14, the MPNet sentence encoder, and the BGE retrieval embedder (`bge-large-en-v1.5`). The CLIP target is the fused Stage-1 prediction; the MPNet and BGE targets come from two dedicated EEG$\to E$ encoders trained exactly as in Stage 1 but against the respective sentence embedding (nine encoders in total: seven pathways plus two selectors). For a trial, each candidate is embedded in the three spaces; per space we take the cosine to the EEG-predicted target and standardize it over the trial's candidate pool, which includes the eight neighbours for calibration; the three standardized scores are summed and the highest-scoring *synthesized* candidate is emitted.

C Fair-Comparison Protocol: Per-Method Coverage and Audited Invariants

This appendix expands the identical-instance protocol of §4: how the shared test set is built, why the methods natively cover different numbers of trials, and the six checks that gate every run. All quantities are regenerated by the released code.

From held-out sentences to shared trials. The seeded split (seed 42; $70/10/20$ over the 1456 unique sentences pooled from ZuCo 1.0 SR/NR/TSR and ZuCo 2.0 NR) holds out 291 sentences. Each held-out sentence is read by many subjects (12 subjects per ZuCo 1.0 task, 18 per ZuCo 2.0 task), so a test *trial*, one $(\text{subject}, \text{sentence})$ pair, expands the 291 held-out sentences to 3817 candidate held-out trials. No method serves all 3817, since each applies its own official preprocessing and quality gate and keeps a different subset. Every method is scored only on the intersection, the trials *all* methods retain, which is $N{=}3166$.

What each pipeline keeps. A trial is "covered" by a method only if that method's own preprocessing accepts it. We re-implement each gate verbatim from the method's released code:

- **Ours**: kept if the trial has at least 0.5 s of EEG-bearing sliding-window rows (the multi-window band-power front-end of Appendix A, which imposes no upper length bound).
- **SemKey**: kept if the sentence-level EEG length lies in $[0.5, 10]$ s *and* all values are finite (its 128-channel, 1280-sample preprocessed tensor).
- **GLIM**: the same rule as SemKey (the two share the same preprocessed dataframe), plus the requirement that GLIM's decoding run emitted a free-running generation for the trial.
- **EEG2Text/BrainTranslator**: kept if the word-level pickle contains a valid text string for that $(\text{subject}, \text{sentence})$.

Why the kept-sets differ. Because the gates differ, the same trial can pass one filter and fail another. A long sentence whose EEG exceeds 10 s is dropped by SemKey/GLIM (their window caps at 10 s) but kept by ours (any length); a trial with a single bad electrode (a non-finite sample) fails SemKey/GLIM's finiteness check but may pass ours and EEG2Text; and a trial whose word-level segmentation failed leaves EEG2Text with no usable entry. The resulting native coverage of the 3817 candidate trials is: ours 3358, EEG2Text 3298, SemKey 3245, GLIM 3200. GLIM is smallest because it inherits SemKey's $[0.5, 10]$ s+finiteness gate *and*, in addition, only emitted a free-running generation for a subset of those trials. Intersecting all four native coverages and deduplicating to one prediction per $(\text{subject}, \text{sentence})$ gives the $N{=}3166$ shared trials on which every table is computed.

Six audited invariants. Every evaluation run is gated by six checks; any violation aborts the run, so a silent leak or a stale manifest cannot reach the final table.

- **INV1 — single hashed test set.** The shared instance list is content-hashed; a stale or altered manifest fails on load, so no method is silently scored on a different set.
- **INV2 — train/test disjointness.** Each method's *realized* training sentence IDs are checked against the test manifest; any overlap aborts the run.
- **INV3 — deterministic seeded perturbations.** The four control scenarios are byte-for-byte reproducible under a fixed seed and shared permutation π.
- **INV4 — scenario-blind models.** The platform perturbs the EEG before the model sees it; the model cannot tell a real trial from a control, so it cannot special-case the real run.
- **INV5 — one locked scorer.** All generations pass through the same retrieval+BERTScore harness, ported from and verified against SemKey's released evaluation code.
- **INV6 — declared privilege.** Each method declares any privileged side-input it reads (e.g. SemKey's oracle attributes), stamped on every result row so the comparison is auditable.

D Analysis and Defense of the Comparison

Because the baseline comparison is a central claim, we defend it in detail.

Table 3: Information-free control panel (24-way, free-running, N=3166, one locked harness). (a) real; (b) shuffled correspondence; (c) noise content, own length; (d) noise content, donor length. gap = (a)$-$ max(controls). Chance ≈ 4.2. SemKey controls also derange its attribute prompt.

Method	(a) real	(b)	(c)	(d)	gap
EEG2Text	4.33	4.38	4.15	4.08	-0.05
GLIM	12.49	4.57	5.15	4.16	$+7.34$
SemKey (GT-attr)	12.87	6.17	6.92	6.95	$+5.92$
SemKey (predicted)	13.68	6.68	6.95	6.92	$+6.72$
Ours	**18.31**	4.17	4.31	4.27	$+14.00$

(1) Identical instances and native generation. Every method is scored on the same N=3166 trials by one harness that calls the published metric functions directly (same embedding model, trial count, and seed). We score each baseline from its *native* free-running generation rather than a re-implemented wrapper, after finding that wrapper re-implementations can be numerically unfaithful. On SemKey's own test set a re-padded wrapper under-generated, scoring 7.6 24-way against 13.9 for the same checkpoint's native generation; the gap would have flattered our comparison, so we use native generation throughout.

(2) Leakage is removed by construction. The prior SOTA's released split includes most evaluation sentences in its training fold: on an earlier controlled subset, 88% of the evaluation sentences appeared in SemKey's training fold. Evaluating a released checkpoint on such sentences largely reflects train-set recall rather than decoding. We therefore retrain *every* baseline from scratch on the canonical training fold under a disjointness guard that aborts on any overlap, so no held-out sentence is seen in training. All baseline numbers here are leakage-free by construction.

(3) The margin is not a single-split artifact. We run an 11-split (ten random $70/10/20$ partitions plus SemKey's native split) $\times$ 3-seed robustness sweep of the strongest baseline, SemKey, against ours, retraining SemKey end-to-end for every split$\times$seed (Table 4). Across the ten converged splits our method wins *all ten*, with a mean gap of $+3.59 \pm 0.74$ points (split-to-split std of 0.78 for ours, 0.54 for SemKey). Predicted-attribute again tracks ground-truth-attribute on every split (no oracle advantage at scale). One split (split_07) is excluded from the headline because all three SemKey end-to-end seeds *diverged* during training (the recipe's short learning-rate warmup reaches an unstable peak; BERTScore 0.69 against 0.84 for converged runs); we report it in the table as a baseline-stability observation rather than a capability measurement (our method trains normally there, 17.33).

Table 4: Robustness sweep: 24-way (seed-mean per split) for ours vs. predicted-attribute SemKey, retrained per split$\times$seed. The headline is over the ten converged splits; split_07, on which SemKey diverged during training, is shown separately.

	Ours	SemKey (pred)	gap
Converged (10 splits)	**16.77**	13.17	$+3.59 \pm 0.74$
split-to-split std	0.78	0.54	(10/10 wins)
split_07 (diverged)	17.33	6.27	—

(4) Reproduction gap. Our SemKey training recipe is identical, field by field, to its released run scripts. On SemKey's own native split it reproduces 24-way ≈ 15.1 (best seed), within ≈ 1.1 of the paper's reported 16.24; the residual does not close with additional seeds and is consistent with the paper using a different sentiment-label source and a slightly larger dataset. The conclusion survives granting SemKey this gap: even lifted by the full $+1.1$ points, its strongest number stays below ours, and the head-to-head on the identical N=3166 set (18.31 vs. 13.68) and the multi-split sweep both decide the comparison independently of it.

Table 5: **Sentence-clustered bootstrap 95% confidence intervals** on the $N{=}3166$ trials. The resampling unit is the *distinct sentence* (all ≈ 12 of its trials travel together), because trials of one sentence share a target and are not independent; 2000 replicates. Retrieval values use a closed-form estimator, so interval width carries no Monte-Carlo component. Marginal intervals overlap between the mid-ranked systems; the decisive comparison is the paired one, since every replicate scores all methods on the same resampled sentences. Paired ours − SemKey (predicted attrs) on 24-way: $+4.62$ $[+3.04, +6.15]$, with ours ahead in $2000/2000$ replicates. Point estimates are the closed-form value on seed 42 and differ from Table 1 in the second decimal, where that table reports a three-seed mean under the default Monte-Carlo estimator.

Method	24-way	10-way	BERTScore	SBERT cos	content rec.	len MAE
Ours	18.36 [16.40, 19.84]	31.96 [29.61, 33.96]	0.851 [0.849, 0.853]	0.259 [0.245, 0.273]	0.029 [0.024, 0.035]	6.40 [5.93, 6.92]
SemKey (predicted attrs)	13.61 [12.30, 14.69]	25.58 [24.01, 26.97]	0.843 [0.841, 0.846]	0.213 [0.202, 0.224]	0.021 [0.018, 0.025]	8.89 [8.48, 9.31]
SemKey (GT attrs / oracle)	12.85 [11.71, 13.82]	24.99 [23.42, 26.39]	0.838 [0.835, 0.841]	0.206 [0.195, 0.217]	0.021 [0.017, 0.024]	9.35 [8.91, 9.79]
GLIM	12.59 [11.24, 13.78]	24.75 [22.88, 26.30]	0.836 [0.834, 0.838]	0.231 [0.219, 0.242]	0.020 [0.016, 0.026]	10.19 [9.55, 10.87]
EEG2Text/BrainTranslator	4.32 [4.01, 4.63]	10.01 [9.63, 10.38]	0.816 [0.814, 0.817]	0.128 [0.119, 0.135]	0.000 [0.000, 0.000]	8.99 [8.21, 9.83]

Table 6: Richness and diversity on the same $N{=}3166$ trials. dist-1/2 are the fractions of unique uni/bigrams over all predictions; self-BLEU measures how much the outputs repeat one another; *distinct* is the fraction of predictions that are distinct strings; content recall is over non-stopword reference words. Only content recall and length MAE have an unambiguous direction and are bolded: the gold-reference row shows that dist-n and self-BLEU are *not* quality axes here, because the 3166 trials cover only 275 distinct sentences, so the references themselves score lowest on dist-2 and highest on self-BLEU. Read these columns as collapse detectors: EEG2Text's 1.000 distinct strings with dist-2 0.041 and content recall 0.000 is template collapse, and is what produces its winning BLEU-1 in Table 1.

Method	dist-1	dist-2	Self-BLEU	content rec. ↑	head ent.	len. MAE ↓	distinct
EEG2Text	0.004	0.041	76.86	0.000	2.36	8.99	1.000
GLIM	0.047	0.152	78.53	0.020	4.23	10.19	0.628
SemKey (GT-attr)	0.072	0.288	42.65	0.021	7.61	9.35	0.970
SemKey (predicted)	0.071	0.288	43.19	0.021	7.34	8.89	0.968
Ours	0.059	0.262	56.78	**0.029**	7.41	**6.37**	0.865
Gold references	*0.032*	*0.075*	*97.43*	*1.000*	*7.65*	*0.00*	*0.085*

E Complete Metric Panel

This appendix reports every metric of SemKey's evaluation suite for all five methods on the identical $N{=}3166$ trials, and the same panel under the information-free controls. All values are produced by re-scoring the stored free-running generations with the locked scorer; no model is re-run. Multi-reference BLEU uses SemKey's eight paraphrase variants, which resolve for 96.9% of targets ($3255/3358$, measured once over our native coverage, a superset of the shared trials); the remainder fall back to single-reference, faithful to its released code.

F Additional Ablation Tables

All tables in this appendix use the val/test protocol of §7 and the single-prompt synthesis variant, which draws all 16 candidates from one prompt by temperature sampling rather than from 16 reordered prompts; each row is a single seed with the generation RNG free, so the resolution is about ± 0.35 24-way.

The alignment target: route through activations, collapse to CLIP. Two choices define the retriever's target, and we validate each by *downstream* 24-way accuracy rather than by retrieval MRR, which nearly *inverts* the downstream ranking and would select the wrong target. Collapsing every pathway into CLIP's gist space and scoring there beats retrieving in each target's own space by a wide margin (17.8 vs. 8.2 single / 13.0 fused), yet the most EEG-decodable *target* is not CLIP text but an intermediate LLM activation: Gemma-3-1b's layer-18 feed-forward projection reaches 17.5 (three seeds) against direct CLIP's 16.8. CLIP is thus the right space to retrieve *in* but not the best

Table 7: **Holding corpus identity constant.** 24-way accuracy when the 24 distractors are drawn from the whole test pool (*all-task*) versus from the target's own reading task (*within-task*), N=3166, chance 4.17 in both. "max ctrl" is the largest score over the three information-free scenarios (b, c, d), computed under the same metric. ZuCo task 1 is SST movie reviews and the other tasks are Wikipedia, so the within-task column removes any credit a method could earn by recovering only the stimulus corpus. Every EEG-using method loses a comparable share, the ranking is preserved, and SemKey's elevated all-task control floor (6.98) falls to chance (4.43), revealing it as corpus identity rather than a language prior.

	all-task distractors		within-task distractors		
Method	real	max ctrl	real	max ctrl	gap
Ours	18.36	4.20	11.12	4.21	+6.91
SemKey (predicted attrs)	13.61	6.98	8.19	4.43	+3.76
SemKey (GT attrs / oracle)	12.85	6.96	7.64	4.25	+3.39
GLIM	12.59	5.17	7.69	4.69	+3.00
EEG2Text/BrainTranslator	4.32	4.40	4.27	4.26	+0.01

Table 8: What fraction of each metric's value is attributable to the EEG: $(\text{real} - \max \text{control})/\text{real}$, on the N=3166 trials. Values near zero mean an information-free model would score about as well. Our row is the mean over three seeds; EEG2Text is negative throughout, as expected. Retrieval accuracy is the most EEG-derived metric for every method that uses the signal, and BERTScore the least; the middle of the ordering is seed-sensitive and should be read only qualitatively.

Metric	EEG2Text	GLIM	SemKey (GT)	SemKey (pred)	Ours
24-way retrieval	-1.1%	58.8%	46.0%	49.2%	76.4%
SBERT cosine	-4.0%	33.8%	21.7%	22.7%	41.9%
BLEU-1	0.1%	18.2%	4.4%	8.6%	22.1%
ROUGE-1 F	-0.3%	16.2%	2.1%	7.1%	19.1%
BERTScore F1	-0.7%	0.6%	-0.3%	0.1%	1.1%
METEOR	-1.2%	13.4%	-0.3%	3.3%	6.7%

thing to predict, and proximity in the target space is not the quantity that matters, since the pathways closest to the true CLIP embedding are not those that decode best downstream. Fusing the seven pathways raises cosine to the true embedding from 0.49 to 0.54 and buys 0.3 points downstream. Their errors are strongly correlated (0.69 mean pairwise cosine; Appendix B), all decoding the same band-limited EEG through the same front-end, so averaging cancels only the ≈31% of error that is target-specific. Readouts of this signal are far less independent than their architectural diversity suggests.

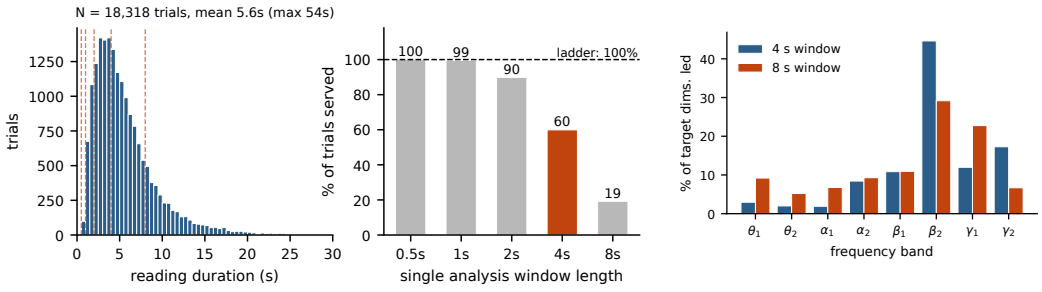

Figure 5: **Why multiple window lengths.** *Left:* reading durations across the 18,318 trials the front-end serves (dashed lines mark the five window lengths), and the fraction of trials a single analysis window could serve, only 59.9% at 4 s against 100% for the ladder. *Right:* the share of target dimensions on which each frequency band attains the highest single-band encoding correlation. High-beta (β_2) dominates at a 4 s window; at 8 s the slow bands' share doubles.

Table 9: **Why an information-free model wins the verbatim block.** Token composition of each method's output. BLEU-1 is (clipped) unigram precision and is indifferent to *which* words match; function words make up about 38% of typical English text and match almost any reference. EEG2Text emits almost nothing else (94.1% stopwords against the references' 37.6%), which buys it the highest unigram precision of any method (31.9%) together with the lowest precision on content words (0.3%). Our output has the same stopword rate as human text (37.8% vs. 37.6%), so its unigram precision has to be earned on content words, where it leads (3.3%).

Method	mean length	% stopwords	unigram precision	precision on *content* words
EEG2Text	10.7	94.1%	31.9%	0.3%
GLIM	17.5	43.2%	13.1%	2.1%
SemKey (GT-attr)	19.3	42.3%	12.8%	2.4%
SemKey (predicted)	18.6	42.7%	13.5%	2.6%
Ours	15.8	37.8%	14.8%	3.3%
Gold references	*18.4*	*37.6%*	—	—

Table 10: Complete metric panel, real EEG (scenario a), N=3166. Multi-reference BLEU uses SemKey's eight paraphrase variants; single-reference BLEU is given for comparability with EEG2Text and GLIM, which use it.

Metric	EEG2Text	GLIM	SemKey (GT)	SemKey (pred)	Ours
2-way ↑	49.37	70.68	69.35	70.08	74.48
4-way ↑	24.82	46.65	45.82	46.69	53.03
10-way ↑	9.98	24.77	25.07	25.84	31.82
24-way ↑	4.33	12.49	12.87	13.68	18.31
BLEU-1 (multi-ref)	33.14	21.89	19.59	20.68	20.61
BLEU-2 (multi-ref)	3.57	7.50	5.52	6.17	4.24
BLEU-3 (multi-ref)	0.15	2.41	1.45	1.60	0.88
BLEU-4 (multi-ref)	0.00	0.94	0.68	0.63	0.29
BLEU-1 (single-ref)	11.55	8.35	8.54	9.02	9.45
BLEU-4 (single-ref)	0.00	0.12	0.15	0.11	0.11
ROUGE-1 F	24.21	12.64	12.25	12.87	13.54
ROUGE-1 P	33.59	14.25	12.95	13.76	14.91
ROUGE-1 R	20.48	13.34	13.15	13.51	13.28
WER ↓	0.969	1.157	1.192	1.165	1.040
BERTScore F1	0.8156	0.8358	0.8384	0.8435	0.8509
SBERT cosine	0.1276	0.2305	0.2063	0.2127	0.2581
dist-1	0.0037	0.0466	0.0720	0.0713	0.0593
dist-2	0.0411	0.1522	0.2883	0.2881	0.2622
Self-BLEU ↓	76.86	78.53	42.65	43.19	56.78
content-word recall	0.0002	0.0203	0.0207	0.0214	0.0287
head entropy	2.362	4.227	7.615	7.344	7.409
METEOR	0.1432	0.1023	0.1079	0.1109	0.1113
CIDEr	0.0187	0.0288	0.0305	0.0327	0.0434
length MAE ↓	8.99	10.19	9.35	8.89	6.37
Frechet dist. ↓	1.2721	0.4539	0.3020	0.2974	0.3286

Table 11: The metric panel under the information-free controls. The *real−max ctrl* row is the drop from real EEG to the best control, per metric. Retrieval accuracy collapses to chance for every method that uses the EEG; the n-gram and semantic metrics retain most of their value even when the EEG carries no information, which is the quantitative form of the argument in §6.

Method	input	24-way	BLEU-1	ROUGE-1	BERTScore	SBERT
EEG2Text	real	4.33	33.14	24.21	0.816	0.128
	shuffle	4.38	33.11	24.28	0.816	0.128
	noise	4.15	27.86	21.69	0.821	0.133
	noise (donor len.)	4.08	28.33	21.71	0.822	0.131
	real−max ctrl	*-0.05*	*+0.03*	*-0.07*	*-0.006*	*-0.005*
GLIM	real	12.49	21.89	12.64	0.836	0.231
	shuffle	4.57	16.86	10.38	0.831	0.128
	noise	5.15	17.91	10.60	0.830	0.153
	noise (donor len.)	4.16	17.09	10.28	0.829	0.144
	real−max ctrl	*+7.34*	*+3.98*	*+2.04*	*+0.005*	*+0.078*
SemKey (GT-attr)	real	12.87	19.59	12.25	0.838	0.206
	shuffle	6.17	17.57	11.21	0.836	0.150
	noise	6.92	18.73	11.99	0.841	0.161
	noise (donor len.)	6.95	17.77	11.38	0.839	0.159
	real−max ctrl	*+5.92*	*+0.86*	*+0.26*	*-0.003*	*+0.045*
SemKey (predicted)	real	13.68	20.68	12.87	0.843	0.213
	shuffle	6.68	17.59	11.22	0.837	0.154
	noise	6.95	18.91	11.95	0.843	0.164
	noise (donor len.)	6.92	18.39	11.63	0.843	0.162
	real−max ctrl	*+6.73*	*+1.77*	*+0.92*	*+0.000*	*+0.049*
Ours	real	18.31	20.61	13.54	0.851	0.258
	shuffle	4.17	16.05	10.60	0.841	0.120
	noise	4.31	15.44	10.95	0.842	0.150
	noise (donor len.)	4.27	15.08	10.69	0.841	0.149
	real−max ctrl	*+14.00*	*+4.56*	*+2.59*	*+0.009*	*+0.108*

Table 12: **Pipeline decomposition** (test fold, single-prompt synthesis variant). Stage-4 reranking applied without synthesis is *worse* than emitting the top-1 retrieved neighbour; synthesis is what lifts both the retrieval metric and BERTScore.

Pipeline	24-way	BERTScore
Stage 1 only (emit top-1 neighbour)	16.32	0.845
Stage 1 + Stage 4 (rerank neighbours, no synthesis)	15.52	0.844
Full (Stages 1–4)	**17.81**	**0.849**

Table 13: Output-stage ablations: neighbourhood size k, candidate pool M, content cues, and the Stage-4 selector. Only k moves the score beyond seed noise. The selector row is the metric-alignment case discussed in §7: $+0.52$ on 24-way, 0.000 on BERTScore. The k=32 and M=32 test-fold runs are unavailable and are omitted rather than imputed; their validation values are 15.21 and 16.15.

Configuration	24-way	BERTScore
Neighbourhood size k (default 8)		
$k = 1$	15.98	—
$k = 2$	16.65	—
$k = 4$	17.24	—
$k = 8$	17.78	—
$k = 16$	17.66	—
Candidate pool M (default 16)		
$M = 1$	17.67	—
$M = 2$	17.35	—
$M = 4$	17.56	—
$M = 8$	17.60	—
$M = 16$	17.81	—
$M = 64$	17.81	—
Content cues (default: keywords + length)		
no cues	17.75	—
length only	16.71	—
keywords only	17.50	—
12 keywords	17.35	—
Stage-4 selector		
CLIP only	17.29	0.849
CLIP+MPNet+BGE (default)	17.81	0.849

Table 14: Stage-3 synthesizer: ten instruction-tuned models from 0.6B to 32B parameters. The 4B default is within seed noise of the best model, an $8\times$ larger one.

Stage-3 synthesizer	24-way
Qwen3-0.6B	16.73
Qwen3-1.7B	17.48
Qwen3-4B (default)	17.81
Qwen3-8B	16.98
Qwen3-14B	17.29
Qwen3-32B	17.98
Gemma-3-1B	16.80
Gemma-3-4B	17.06
Gemma-3-12B	16.91
Gemma-3-27B	17.76

Table 15: Stage-1 encoder variants, measured by pure retrieval (emit the top-1 neighbour) so that the readout is scored without the language model. No variant separates from the default: the two that score above it (PMA + larger encoder, 16.60; per-subject centering, 16.58) do so by less than the run-to-run spread, and scaling the encoder alone is harmful. Single seed per row.

Stage-1 encoder	24-way	BERTScore
Default (masked mean, d=256, 2 layers)	16.32	0.845
PMA attention pooling	16.31	0.846
Larger encoder (d=512, 4 layers)	15.28	0.843
PMA + larger encoder	16.60	0.844
Richer tokens (mean|std|slope, 2520-d)	15.73	0.844
Per-subject feature centering	16.58	0.845