ESCAPING THE BLEU TRAP: A SIGNAL-GROUNDED FRAMEWORK WITH DECOUPLED SEMANTIC GUIDANCE FOR EEG-TO-TEXT DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding natural language from non-invasive EEG signals is a promising yet challenging task. However, current state-of-the-art models remain constrained by three fundamental issues: *Semantic Bias*, where outputs collapse into generic linguistic templates; *Signal Neglect*, where models rely heavily on LLM priors to hallucinate fluent text even in the absence of meaningful signals; and the *"BLEU Trap"*, where high-frequency stopwords inflate n-gram metrics, masking a lack of true semantic fidelity. To resolve these challenges, we move beyond conventional end-to-end pipelines and propose **SEMKEY**, a novel two-stage framework that enforces signal-grounded generation through four decoupled semantic objectives: sentiment, topic, length, and surprisal. We extract these semantic anchors from EEG embeddings directly, then unify them with an *Active Retrieval Decoding* mechanism, encouraging the LLM to ground its token generation in the neural signals rather than defaulting to linguistic priors. Furthermore, we break the BLEU Trap by establishing a comprehensive evaluation protocol using rigorous retrieval and distribution-based metrics such as Fréchet Distance. Extensive experiments under fully autoregressive inference, including noise and time-shuffled EEG controls, provide evidence of signal dependency and demonstrate that SEMKEY achieves SOTA performance on these robust protocols. Code will be released upon acceptance.

1 INTRODUCTION

Decoding natural language from neural activity is a pivotal frontier in Brain-Computer Interface (BCI) research, offering a lifeline for individuals with aphasia or locked-in syndrome to restore communication (Défossez et al., 2023; Willett et al., 2023). While invasive modalities like Electrocorticography (ECoG) and Stereoelectroencephalography (sEEG) demonstrate high fidelity (Metzger et al., 2023; Feng et al., 2025), their invasive procedures make them unsuitable for widespread use in general-purpose BCI applications. Consequently, non-invasive Electroencephalography (EEG) remains the most promising modality for widespread adoption due to its portability and safety.

While EEG-to-text decoding holds great promise, the field currently faces a validity crisis. Recent studies (Jo et al., 2024) reveal that prior methods (Kostas et al., 2021; Wang & Ji, 2022; Duan et al., 2023; Liu et al., 2024; Tao et al., 2025) heavily rely on *Teacher Forcing* (TF) during inference—feeding ground-truth text prefixes to predict subsequent words, which is impractical for real-world BCIs where the user's intended message is unknown. TF lets models exploit language priors to guess the next word (e.g., predicting "Washington" given "The US president") rather than interpret the EEG signals; replacing the training EEG with random noise even yields identical performance (Jo et al., 2024), and when TF is removed during inference, the generation quality of these models degrades rapidly. To ensure outputs are genuinely grounded in neural activity rather than hallucinated from language priors, a deployable BCI must operate without TF. Consequently, our approach enforces fully autoregressive decoding at inference, predicting text *solely* from input EEG and its own generated tokens; TF is used only as a standard training strategy, as in all the models we compare.

GLIM (Liu et al., 2025) is an open-vocabulary method that removes teacher forcing; however, our analysis reveals that it achieved limited results due to two critical issues: **semantic bias and persistent reliance on language priors**. First, its output frequently collapses into highly repetitive,

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

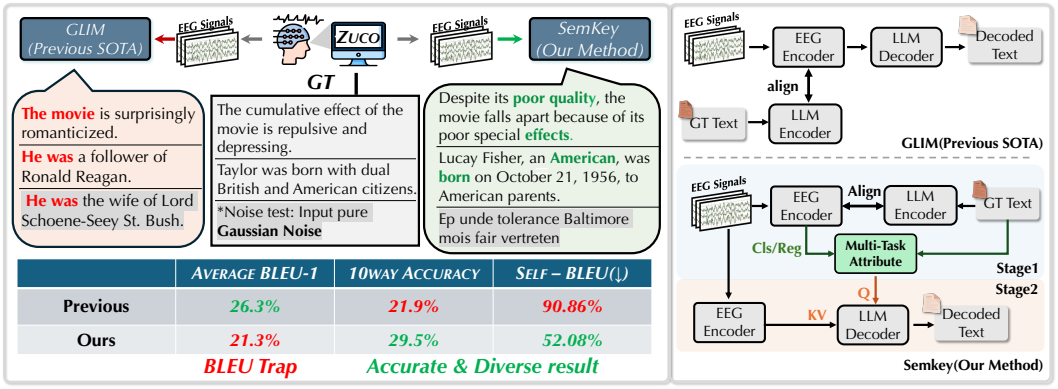

Figure 1: **Left (Decoding showcase & Quantitative Results):** Previous models (Liu et al., 2025) exhibit severe *semantic bias*, overfitting to generic templates to artificially inflate BLEU scores despite low semantic alignment and repetition - a phenomenon we identify as the *"BLEU Trap"*. The noise test (gray highlight) exposes persistent ***template reliance***: the model retains recognizable corpus patterns under pure Gaussian noise. In contrast, SEMKEY breaks the BLEU Trap, demonstrating accurate, diverse, and signal-dependent generation, producing disordered tokens for noise. **Right (Architectural Comparison):** To address these issues, SEMKEY replaces the conventional end-to-end decoding pipeline (top) with a novel two-stage framework (bottom).

generic templates (e.g., sentences start with "He was..." or "The movie...", Figure 1 left), so the generated distribution diverges significantly from the ground truth (Figure 5; Appendix Table o&q). Second, following Jo et al. (2024), our noise input experiments expose that GLIM still heavily relies on the linguistic priors of the LLM: when fed with pure Gaussian noise, it continues to generate fluent sentences with the same templates seen with real inputs (Figure 4 and Table p).

Motivated by these observations, we propose SEMKEY, which features a **Parallel Neural-Driven Attribute Decouple** module, motivated by neuroscientific research demonstrating that high-level semantics form robust cortical representations (Lerner et al., 2011). To mitigate semantic bias, we explicitly disentangle key semantic attributes (*sentiment, topic, length, surprisal*) directly from the EEG features (Figure 1 right bottom), which constrain the LLM and prevent template collapse. To address signal neglect, we implement an **Active Retrieval Decoding** mechanism that adjusts the Q-K-V workflow to fuse the disentangled attributes and continuous EEG features into decoding, encouraging the LLM to ground its token generation in the neural signals.

Furthermore, we observe that the template memorization behavior of existing baselines is often obscured by the ***"BLEU Trap"***(Figure 1): n-gram matching metrics can be inflated by simply replicating stopwords and frequent templates, masking a lack of semantic fidelity (Table o). To resolve this, we move beyond the BLEU Trap by establishing a comprehensive evaluation protocol utilizing rigorous retrieval-based and distribution-based metrics. Comprehensive evaluations demonstrate that SEMKEY establishes a new state-of-the-art, prominently outperforming the GLIM baseline, and controls with Gaussian-noise and time-shuffled EEG confirm that its outputs depend on the neural input.

2 RELATED WORK

EEG-to-Text Decoding: Evolution and Frontiers. Early approaches relied on handcrafted features such as MFCCs (Cooney et al., 2018) and wavelet transforms (Panachakel et al., 2020), later replaced by CNNs and LSTMs (Zhang et al., 2018). A paradigm shift occurred with Transformer-based pipelines: Wang & Ji (2022) proposed projecting EEG signals into a pre-trained BART latent space, and building on this, DeWave (Duan et al., 2023) introduced an end-to-end framework processing raw brainwaves for direct translation. More recently, the integration of Large Language Models has spurred new directions. Thought2Text (Mishra et al., 2025) explores leveraging LLMs for EEG encoding, while WaveMind (Zeng et al., 2025) proposes a conversational foundation model aligned across textual and visual modalities. *However, our work differs from these distinct streams:* unlike WaveMind and Thought2Text, which rely on **image stimuli** that offer richer and more direct

semantic grounding, we tackle the different task of *pure* EEG-to-Text decoding from abstract *textual stimuli* without such visual aids. **The Validity Crisis and Methodological Gap.** Notwithstanding architectural advancements, the field faces a crisis of validity. The pivotal study "Are EEG-to-Text Models Working" (Jo et al., 2024) exposed that high BLEU scores in prior arts are often artifacts of *Teacher Forcing*, where models ignore the noisy EEG encoder in favor of the LLM's linguistic priors. While GLIM (Liu et al., 2025) recently addressed this by establishing the first effective pipeline without teacher forcing, it still exhibits strong corpus-template reliance under noise and time-shuffled input, alongside semantic collapse in open-ended generation (as discussed in Section 1). Our work builds on the diagnosis of Jo et al. (2024) and extends it to the teacher-forcing-free setting, where signal neglect and BLEU inflation still occur (Appendix I). To bridge this gap, SEMKEY combines explicit attribute guidance, which constrains the global semantics of the output, with a *Q-K-V Injection* workflow that enforces a token-level dependency on neural signals (Section 3).

3 METHOD

SEMKEY is a two-stage framework (Figure 2): (1) a *Parallel Neural-Driven Attribute Decouple* stage that learns EEG representations aligned with the linguistic space while decoupling high-level semantic attributes; and (2) a *Multi-Perspective Active Retrieval Decoding* stage that exploits these representations to produce faithful text.

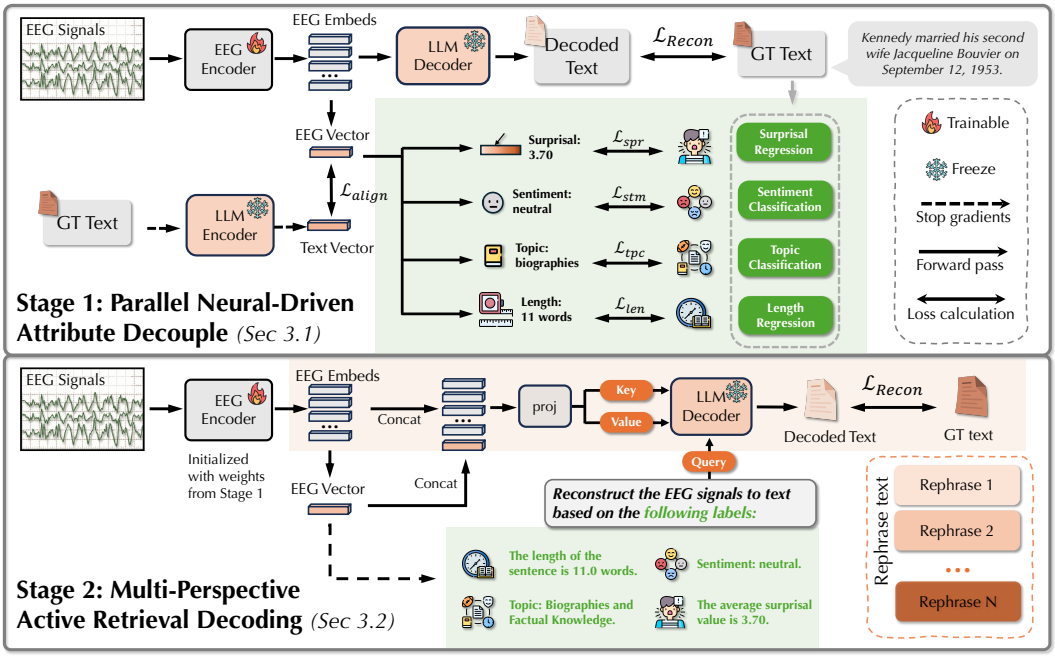

Figure 2: The overall architecture of the SEMKEY framework following a "Guidance-Generation" paradigm. **Stage 1 (Attribute Extraction)**: The EEG encoder is optimized via multi-task learning to explicitly decouple high-level semantic attributes (Sentiment, Topic, Length, Surprisal) alongside standard alignment. **Stage 2 (Generative Decoding)**: The attribute predictions cached after Stage 1 structure a fixed semantic prompt to guide generation; Stage 2 encoder updates do not alter these cached predictions. Crucially, the **Q-K-V Injection** workflow enforces signal dependency by using the text prompt as the Query and projected EEG embeddings as Keys and Values.

3.1 STAGE 1: PARALLEL NEURAL-DRIVEN ATTRIBUTE DECOUPLE

Based on Liu et al. (2025), the EEG encoder inherits alignment and reconstruction objectives via a frozen Flan-T5 model (Chung et al., 2024), and four auxiliary heads additionally predict sentiment, topic, length, and surprisal as a semantic anchor for generation.

Prior-Guided EEG Encoding Stream. Following GLIM (Liu et al., 2025), we mitigate inter-subject variability by embedding discrete priors (e.g., subject ID, dataset) into a context vector. This vector modulates a Transformer-based EEG encoder (Vaswani et al., 2017; Jiang et al., 2024) (encoder blocks followed by query-based decoder blocks; Appendix C) that processes input signals to output two representations: (1) **EEG Embeddings** $\mathbf{E}_{eeg} \in \mathbb{R}^{L \times d_{model}}$ for token-level reconstruction, and (2) an attention-pooled (Lee et al., 2019) **EEG Vector** $\mathbf{v}_{eeg} \in \mathbb{R}^{d_{model}}$ for high-level semantic prediction. The encoder is optimized using reconstruction ($\mathcal{L}_{recon}$), contrastive alignment ($\mathcal{L}_{clip}$), and commitment ($\mathcal{L}_{commit}$) losses to ensure text-space consistency (Liu et al., 2025).

Parallel Neural-Driven Learning Framework. To enrich the semantic structure of $\mathbf{v}_{eeg}$, we introduce a parallel multi-task framework (Wang et al., 2025b). Four attribute heads are attached to $\mathbf{v}_{eeg}$. Crucially, these four attributes constitute a complementary semantic anchor that comprehensively circumscribes the generation process. *Topic* and *Sentiment* act as macroscopic anchors that dictate the semantic category and affective tone, directly preventing the LLM from collapsing into irrelevant generic templates (e.g., semantic bias) (Liang et al., 2024; Lorandi & Belz, 2023). *Length* serves as a structural boundary, reflecting the working memory load during language processing (Just & Carpenter, 1992), which effectively curtails recursive loops and repetitive generation. Furthermore, *Surprisal* (information density)—which correlates strongly with well-documented ERP components like the N400 (Kutas & Hillyard, 1980)—forces the LLM to modulate its probability distribution (Michaelov et al., 2024). This explicitly guides the model to generate diverse, contextually appropriate words rather than defaulting to high-frequency "safe" tokens, thereby breaking the BLEU Trap. *Sentiment* and *Topic* are predicted by two classifiers trained with cross-entropy ($\mathcal{L}_{stm}$, $\mathcal{L}_{tpc}$), and *Length* and *Surprisal* by two regressors trained with mean squared error on normalized targets ($\mathcal{L}_{len}$, $\mathcal{L}_{spr}$); label definitions are given in Appendix C.

Overall Training Objective. The overall objective is:

$$\begin{aligned}\mathcal{L}_{Stage1} = \lambda_{clip}\mathcal{L}_{clip} + \lambda_{commit}\mathcal{L}_{commit} + \lambda_{recon}\mathcal{L}_{recon} \\ + \lambda_{stm}\mathcal{L}_{stm} + \lambda_{tpc}\mathcal{L}_{tpc} + \lambda_{len}\mathcal{L}_{len} + \lambda_{spr}\mathcal{L}_{spr}.\end{aligned} \tag{1}$$

Each of the seven losses has its own weight: $(\lambda_{clip}, \lambda_{commit}, \lambda_{recon}, \lambda_{stm}, \lambda_{tpc}, \lambda_{len}, \lambda_{spr}) = (0.5, 0.7, 0.5, 0.3, 0.3, 0.9, 0.3)$ (Appendix C). This composite objective results in an EEG Encoder capable of producing semantically disentangled and attribute-aware embeddings, laying the foundation for the generative stage.

3.2 Stage 2: Multi-Perspective Active Retrieval Decoding

After Stage 1 training, the predicted attribute labels are cached and turned into a fixed text prompt (*Prompt Guidance*), and the aligned EEG features are injected into the LLM via a learnable projector (*Q-K-V Injection*), so that the LLM actively queries neural signals for every token. The EEG Encoder remains trainable (with the auxiliary attribute losses kept active) while the LLM Decoder is frozen. Updating the encoder in Stage 2 does not change the cached attributes or their prompts.

Multi-Perspective Prompt. To explicitly guide text reconstruction using the high-level semantics obtained in Stage 1, we leverage the decoupled attribute predictions to construct a structured natural language prompt, serving as a "semantic anchor" for the LLM (Keskar et al., 2019; Lester et al., 2021). Let $\mathcal{H} = \{\hat{y}_{len}, \hat{y}_{tpc}, \hat{y}_{stm}, \hat{y}_{spr}\}$ be the set of predicted attributes. A template function $\mathcal{T}(\cdot)$ maps these attributes into a prompt $\mathbf{X}_{prompt} = \mathcal{T}(\mathcal{H})$, the specific prompts can be found in Appendix C. This prompt $\mathbf{X}_{prompt}$ is then tokenized and prepended to the decoder input. The hidden states derived from this prompt act as the **Queries** (Q), seeking relevant information to fulfill semantic constraints.

Q-K-V Injection Workflow. Unlike previous approaches that concatenate brain embeddings with text tokens—leading to signal neglect via self-attention (Wu et al., 2022; Mishra et al., 2025)—we employ a *Q-K-V Injection* strategy to separate the information flow (Li et al., 2023; Alayrac et al., 2022). The autoregressive input sequence consists of the prompt $\mathbf{X}_{prompt}$ and previously generated tokens $\mathbf{y}_{<t}$. The decoder processes this to obtain contextualized hidden states $\mathbf{H}_{text} = \Phi_{self}([\mathbf{X}_{prompt}; \mathbf{y}_{<t}])$, which serve as *Queries* for the cross-attention layers. To provide a multi-scale representation, we prepend the global vector $\mathbf{v}_{eeg}$ to the token-level embeddings $\mathbf{E}_{eeg}$, forming an augmented $(L+1) \times d_{model}$ sequence. This unified representation is then projected to form the *Keys* and *Values*:

$$Q = \mathbf{H}_{text}\mathbf{W}_Q, \quad K = ([\mathbf{v}_{eeg}; \mathbf{E}_{eeg}]\mathbf{W}_{proj})\mathbf{W}_K, \quad V = ([\mathbf{v}_{eeg}; \mathbf{E}_{eeg}]\mathbf{W}_{proj})\mathbf{W}_V \tag{2}$$

where $\mathbf{W}_{proj}$ is a trainable linear layer ($1024 \to 1024$), and $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$ are the pretrained cross-attention weights of the frozen Flan-T5 decoder. Unlike GLIM (Liu et al., 2025), where the projected EEG directly replaces the text-encoder output without prompt queries (the *w/o Q-K-V* variant in Section 4.3), we let the attribute prompt query the EEG sequence at every decoding step, encouraging signal-dependent generation despite the template reliance observed in GLIM (Section 4.2); the cross-attention conditioning itself is not claimed as new. The decoder is trained with standard teacher forcing; at test time, text is generated fully autoregressively with beam search (default decoding parameters), without any ground-truth prefix.

3.3 Escaping the "BLEU Trap": A Holistic Evaluation Protocol

Retrieval-based Alignment Metrics (N-Way Accuracy). To measure semantic alignment, we employ *N-way Retrieval Accuracy* (Radford et al., 2021; Li et al., 2024). We utilize the `sentence-transformers` (Reimers & Gurevych, 2019) library (model `all-mpnet-base-v2` (Song et al., 2020)) to encode the generated and ground-truth texts into dense vectors; only the original reference sentence is used as ground truth.

To mitigate variance from random negative sampling, we adopt a Monte Carlo sampling approach, conducting $M = 1000$ independent trials for each difficulty level ($N \in \{2, 4, 10, 24\}$). In each trial m, we randomly sample a set S_m of N test samples; each generated text embedding $\mathbf{v}_{gen}^{(i)}$, $i \in S_m$, is compared with its own ground-truth embedding $\mathbf{v}_{gt}^{(i)}$ and with the other $N-1$ ground-truth embeddings in S_m as distractors. The final accuracy is averaged over samples and trials:

$$\text{Acc}_N = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{i\in S_m}\mathbb{I}\Big(\text{CosSim}(\mathbf{v}_{gen}^{(i)}, \mathbf{v}_{gt}^{(i)}) \geq \max_{j\in S_m, j\neq i}\text{CosSim}(\mathbf{v}_{gen}^{(i)}, \mathbf{v}_{gt}^{(j)})\Big) \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function; chance level is $1/N$. This metric evaluates the model's ability to discriminate the correct semantic content from candidate distractors.

Semantic Similarity Metric (SBERT). High BLEU scores can be achieved by matching stopwords, while missing the core meaning. To evaluate sentence-level meaning, we report the mean cosine similarity between the embeddings of each generated sentence and its reference, using the same `all-mpnet-base-v2` model (*SBERT*). Pairing each reference with a randomly permuted reference gives 0.109 ± 0.002 (20 permutations). This reference–reference comparison provides a *random-reference baseline*.

Generative Diversity & Quality Metrics. To quantify *semantic bias*, where the model repeats a few safe sentences, we employ *Self-BLEU* (Zhu et al., 2018) and *Head Entropy (Prefix Diversity)* (Holtzman et al., 2019). Self-BLEU measures the BLEU score of each generated sentence against all others in the test set, while Head Entropy calculates the Shannon entropy of the initial bigrams (first two tokens) across all generated sentences. These two metrics are read relative to the ground-truth (GT) text rather than pushed to extreme values (e.g., arbitrarily low Self-BLEU or infinitely high entropy), which random or forcibly diversified text can also reach (Section 4.1); approximating the GT values indicates that the model generates distinct sentences for different EEG inputs while maintaining natural coherence.

To evaluate semantic realism at the distribution level, we compute the *Fréchet Distance (FD)* (Semeniuta et al., 2018). We map all sentences to a high-dimensional semantic space using the same `all-mpnet-base-v2` model and calculate the distance between the multivariate Gaussian distributions of the generated (μ_g, Σ_g) and reference (μ_r, Σ_r) embeddings:

$$\text{FD} = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{4}$$

A lower FD indicates that the distribution of the decoded text closely matches the semantic manifold of the ground truth text.

4 Experiments & Results

We conduct our experiments using the data from the *ZuCo 1.0* (Hollenstein et al., 2018) and *ZuCo 2.0* (Hollenstein et al., 2020) datasets, containing over 22K sentence-EEG pairs (detailed in Appendix C.3). This dataset is widely adopted by previous state-of-the-art methods (Duan et al., 2023;

Liu et al., 2024); we split it by unique sentences, so that no test sentence is seen during training. Preprocessing and implementation details (including optimizer settings and hyperparameters) are given in *Appendix C*.

We compare SEMKEY against the recent baseline *GLIM* (Liu et al., 2025), along with previous methods including *EEG-to-Text* (Wang & Ji, 2022), *DeWave* (Duan et al., 2023), and *EEG2Text* (Liu et al., 2024). Evaluating these prior models with teacher forcing (TF) at inference leaks the ground-truth prefix (Jo et al., 2024), so we report full metrics under this setting only for *EEG-to-Text*. Without TF at inference (all models, including ours, are still trained with TF), Jo et al. (2024) identify signal neglect as a shared failure mode of this family of prior methods. Motivated by this finding, we summarize these methods with a representative baseline, *Prior Model (w/o TF)*, instantiated by *EEG-to-Text*, the only fully open-source one among the three. To test whether output diversity alone improves the metrics, we also decode this baseline with forced diversity (beam size 5, repetition penalty 5, and no repeated bigrams). As a reference for our evaluation protocol, we also report the scores of the ground-truth (GT) sentences themselves. All experiments were run with different random seeds, with hypothesis testing for statistical significance (Appendix A).

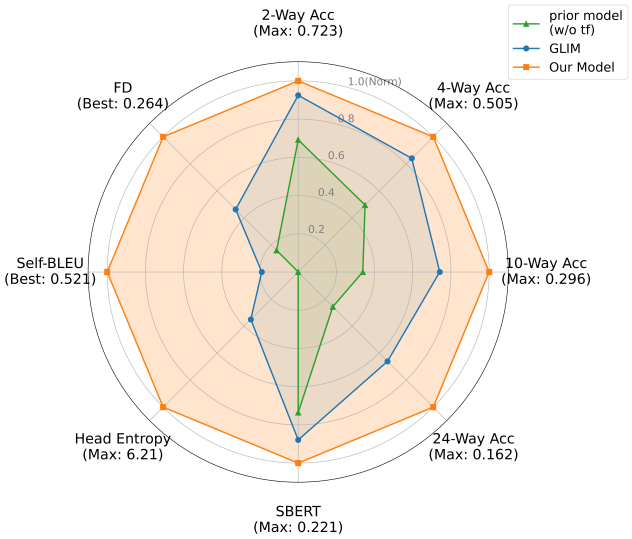

Figure 3: **Holistic Performance Evaluation.** The Radar Chart illustrates the relative performance of our model (Orange) versus GLIM (Blue) and prior model without teacher forcing (Green). SEMKEY demonstrates a balanced superiority across all alignment, diversity, and content quality metrics.

4.1 MAIN PERFORMANCE COMPARISON

Quantitative Results. Table 1 and Figure 3 present the comprehensive performance comparison. First, SEMKEY consistently outperforms baselines across all retrieval settings ($N = 2, 4, 10, 24$) and in *SBERT* similarity, with the advantage becoming more pronounced as task difficulty increases, indicating that it distinguishes fine-grained semantics against a large pool of distractors. Second, SEMKEY exhibits a substantial leap in structural variety (e.g., higher *Head Entropy*), avoiding the "safe loop" of repetitively generating "He was...". Finally, SEMKEY significantly improves *Self-BLEU* and *FD*, showing that its generated semantic distribution is substantially closer to the ground truth manifold.

It is worth noting that the *Prior Model (w/o TF)* suffers from complete signal neglect (outputting identical sentences for all inputs), which results in chance-level *Retrieval Accuracy*. Forcing it to produce diverse outputs raises *Head Entropy* and lowers *FD*, yet its retrieval accuracy stays at chance level and its *SBERT* does not improve. This underscores the necessity of a *holistic evaluation*: diversity alone is insufficient and must be interpreted alongside retrieval accuracy to rule out hallucination. Similarly, the teacher-forced model (*EEG-to-Text*) exhibits abnormal behavior. Despite high *Retrieval Accuracy* obtained with the leaked prefix, its generation quality and diversity suffer severe degradation, evidenced by a substantially worse *FD* and an inflated *Self-BLEU* score compared to SEMKEY. Manual inspection reveals that this discrepancy is caused by the model generating massive meaningless suffixes (e.g., "ggg,,,,"). This phenomenon highlights the advantage of our comprehensive evaluation protocol: it can successfully detect underlying mode collapse issues that traditional metrics like BLEU fail to capture.

Table 1: Performance comparison on ZuCo. We use published results where available and official implementations for additional metrics. BLEU uses multiple references (MTV) to illustrate the BLEU Trap; retrieval accuracy and SBERT use the original reference sentence only. Results with TF at inference (struck through) leak ground-truth prefixes and are shown for reference only. "Prior Model (w/o TF)" groups EEG-to-Text (Wang & Ji, 2022), DeWave (Duan et al., 2023), and EEG2Text (Liu et al., 2024) because Jo et al. (2024) report similar performance across these methods after removing TF at inference, with complete output collapse and signal neglect. We evaluate EEG-to-Text as their representative; it outputs the same sentence for every input. Only with "+ Forced Diversity" (repetition penalties and bigram blocking) does this baseline produce more varied outputs under the tested settings. This behavior is consistent with the findings reported by Jo et al. (2024). "GT" evaluates the reference sentences themselves. Best results are bold; BLEU scores are gray.

	BLEU Metrics (%)				Retrieval Accuracy (%)				Semantic Diversity		Quality (↓)	
Model	B-1	B-2	B-3	B-4	2-Way	4-Way	10-Way	24-Way	SBERT	H. Ent	S-BLEU (%)	FD
Inference with Teacher Forcing												
EEG-to-Text (Wang & Ji, 2022)	~~40.12~~	~~23.18~~	~~12.61~~	~~6.80~~	~~95.66~~	~~90.06~~	~~80.30~~	~~68.48~~	~~0.406~~	~~6.06~~	~~78.39~~	~~0.76~~
DeWave (Duan et al., 2023)	~~41.35~~	~~24.15~~	~~13.92~~	~~8.22~~	-	-	-	-	-	-	-	-
EEG2Text (Liu et al., 2024)	~~45.20~~	~~29.10~~	~~19.70~~	~~14.10~~	-	-	-	-	-	-	-	-
Inference without Teacher Forcing												
Prior Model (w/o TF)	22.01	6.17	2.56	1.26	50.00	25.00	10.00	4.17	0.163	0.00	100	1.65
+ Forced Diversity	20.51	5.08	1.14	0.40	49.48	24.69	9.68	3.96	0.154	1.68	88.26	0.92
GLIM (Liu et al., 2025)	26.32	10.67	3.96	1.86	66.83	42.45	21.91	10.74	0.194	2.18	90.86	0.57
SemKey(Ours)	21.30	6.39	1.82	0.81	**72.29**	**50.48**	**29.56**	**16.24**	**0.221**	**6.21**	**52.08**	**0.26**
GT (Reference)	100	100	100	100	100	100	100	100	1.000	6.35	49.38	0.00

Table 2: BLEU metrics comparison (BLEU-N). Note that "MTV Evaluation" here refers to the use of augmented references during scoring. The significant performance gap observed upon removing MTV indicates that the MTV evaluation setting amplifies the BLEU Trap.

	MTV Evaluation		Standard (w/o MTV)	
Model	B-1	B-2	B-1	B-2
Prior Model (w/o TF)	22.00%	6.17%	**11.90%**	**1.58%**
GLIM	**26.30%**	**10.70%**	7.84%	1.28%
SemKey (Ours)	21.30%	6.39%	8.98%	1.42%

The "BLEU Trap": Metric Inflation via Semantic Bias. As shown in Table 2, our results empirically confirm the "BLEU Trap," where evaluation metrics are inflated by linguistic priors. The trap arises from n-gram matching on the highly templated ZuCo sentences, and is further amplified by the widespread *MTV protocol*, which computes scores against an augmented set of paraphrases. This expanded reference pool disproportionately rewards generic templates, allowing baselines like GLIM to inflate scores by hallucinating "safe," high-frequency prefixes (e.g., "The movie..."). Indeed, only 41.8% of GLIM's outputs are unique (vs. 97.9% for SemKey): 32.9% of them start with "He was a", and the sentence "He was a member of the Royal Astronomical Society." alone appears 90 times. Removing these augmented references in the *Standard setting (w/o MTV)* causes a drastic performance shift. GLIM's scores degrade severely, falling behind our proposed SemKey. This

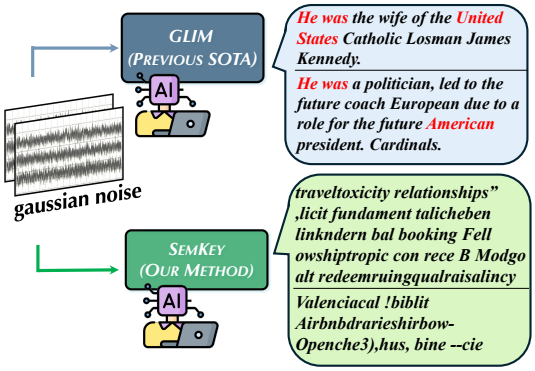

Figure 4: **Showcase of Gaussian Noise input.** We illustrate the text generated by GLIM versus our method when the input is pure Gaussian noise instead of EEG signals.

sharp decline reveals that GLIM's apparent advantage under the MTV protocol is largely an artifact of matching generic patterns rather than performing faithful semantic decoding.

More intriguingly, under this Standard setting, the *Prior Model (w/o TF)* paradoxically achieves the highest BLEU scores. Given that this baseline suffers from complete mode collapse, its top ranking is a clear sign of the BLEU Trap. It demonstrates that the metric fails to penalize semantic vacuity and can be artificially inflated by repetitively generating high-frequency words, thereby masking the model's fundamental inability to capture unique semantic content. Extended analysis, including a prefix-stripping test, the role of MTV in training and evaluation, and qualitative examples, is provided in Appendix G.

Table 3: Quantitative verification of signal dependency. We compare performance under Real EEG and Gaussian Noise. Chance levels of 4-Way and 24-Way retrieval are 25.00% and 4.17%. The performance changes characterize input sensitivity; GLIM retains corpus-template patterns under noise, whereas SEMKEY shows marked disruption of linguistic structure and semantic matching.

Model	Input	4-Way (%)	24-Way (%)	SBERT	Head Entropy	FD
Prior Model (w/o TF)	Real	25.00	4.17	0.163	0.00	1.65
	Noise	25.00	4.17	0.163	0.00	1.65
GLIM	Real	42.45	10.74	0.194	2.18	0.57
	Noise	25.00	4.17	0.119	5.65	0.72
SEMKEY (Ours)	Real	**50.48**	**16.24**	**0.221**	**6.21**	**0.26**
	Noise	25.00	4.17	0.008	11.08	1.27

4.2 VERIFICATION OF SIGNAL DEPENDENCY

To verify the signal dependency of our framework, we follow Jo et al. (2024) and conduct a *Noise Input Test*, where the raw EEG signals are replaced with Gaussian noise during inference.

Quantitative Analysis: The "Garbage In, Garbage Out" Test. We compare performance under Real EEG and Gaussian Noise inputs across three models (Table 3). The *Prior Model (w/o TF)* yields identical metrics for real brainwaves and pure noise, indicating complete signal neglect. GLIM loses semantic alignment under noise but retains recognizable corpus templates, consistent with persistent reliance on learned language priors. For SEMKEY, retrieval accuracy falls to chance level, SBERT decreases from 0.221 to 0.008, and Head Entropy and FD rise sharply. The accompanying disruption of linguistic structure and semantic matching supports input sensitivity and signal dependency. Results under gradually increasing noise are provided in Appendix D. Since Gaussian noise differs from real EEG in amplitude and spectrum, we also use a *Time-Shuffle Test*: randomly permuting temporal segments preserves the amplitude distribution and structure within each segment while disrupting the order between segments. This complements the Gaussian-noise control and is designed to reduce the influence of out-of-distribution effects. Under this control, SEMKEY shows larger retrieval drops than GLIM at all levels, while the Prior Model is unaffected (Appendix D.2).

Qualitative Case Study. Figure 4 shows representative outputs under Gaussian noise input. GLIM still generates fluent but semantically vacuous sentences (e.g., *"He was ..."*), whereas SEMKEY outputs unstructured gibberish, reflecting the absence of valid neural information. More examples are in Appendix H.

Table 4: Performance of Stage 1 decoupled attribute prediction heads.

Task	Metric	Chance	Ours
Topic	Acc (↑)	50%	**83.51%**
Sentiment	Acc (↑)	50%	**90.71%**
Length	MAE (↓)	6.23	**5.01**
Surprisal	MAE (↓)	0.78	**0.58**

Table 5: Ablation study on key components.

Variant	4-Way	24-Way	Self-BLEU(↓)	FD(↓)
w/o Q-K-V	39.8%	9.2%	80.9%	0.66
w/o Prompt	38.9%	9.9%	70.6%	0.60
w/o Topic	43.7%	12.3%	69.4%	0.40
w/o Sentiment	44.5%	12.1%	72.7%	0.42
w/o Length	43.5%	11.7%	76.2%	0.67
w/o Surprisal	40.7%	10.4%	84.7%	0.63
SEMKEY	**50.4%**	**16.2%**	**52.1%**	**0.26**

4.3 Ablation Studies

Analysis of Decoupled Attribute Decoding. Table 4 shows that the Stage 1 predictions of all four attributes outperform their respective baselines, validating our decoupled learning strategy.

Impact of Architecture and Guidance. We investigate the contribution of key components in Table 5. *First, regarding the architecture,* removing the **Q-K-V Injection** workflow (reverting to the GLIM-style direct decoding in Section 3) causes a severe diversity collapse. This indicates its important role in encouraging the model to attend to neural signals rather than falling back on generic repetitions. *Second, regarding prompt guidance,* removing any single attribute—whether Topic, Sentiment, Length, or Surprisal—incurs distinct performance drops. This demonstrates the *complementary nature* of these semantic anchors, where their synergy is essential for aligning EEG signals with coherent text generation. Replacing either input route at inference also lowers 4-way retrieval: a dummy prompt with real EEG gives 36.0%, the real prompt with noise EEG gives 31.9%, and replacing both gives 25% (chance level), so both the prompt and the cross-attended EEG carry information (Appendix E.1). Scaling the language model does not remove the need for the attributes: without *Length*, 4-way retrieval drops from 51.2% to 44.7% with T5-XL (Appendix F). Ablations on label reversal and corruption are in Appendix E.

4.4 Semantic Distribution Analysis

To further assess the semantic quality of the generated text, we map the generated sentences into a high-dimensional semantic space using a pre-trained Sentence Transformer and visualize the distribution via t-SNE. As illustrated in Figure 5, the semantic distribution of SemKey (Orange points) aligns closely with the *Ground Truth* (Blue points), exhibiting a uniform distribution across the semantic space. Conversely, *GLIM* (Yellow points) demonstrates strong semantic bias, forming dense clusters in regions where the Ground Truth is sparse (shown in red circle). This discrepancy indicates that GLIM tends to over-generate specific biased sentences rather than learning the true distribution. This visual evidence aligns with our quantitative FD results, suggesting that the outputs of SemKey are closer to the diversity and semantic distribution of the ground truth.

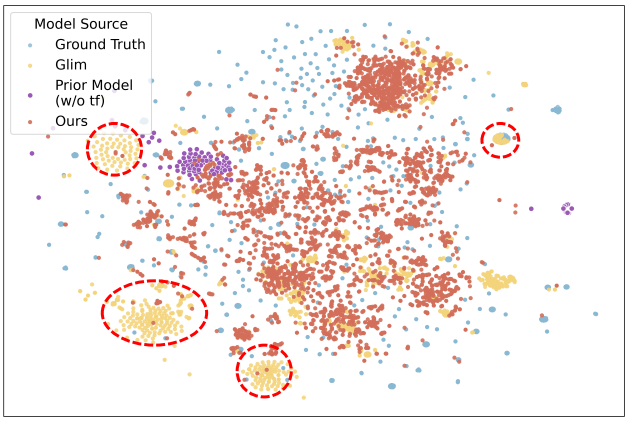

Figure 5: **t-SNE Visualization of Semantic Distribution.** We visualize the sentence embeddings of Ground Truth (Blue), SemKey (Orange), GLIM (Yellow), and Prior Model (w/o TF) (Purple). Our model shows better overlap with the ground truth distribution, while effectively avoiding semantic bias (highlighted by red circles).

5 Conclusion

We introduce SemKey, an EEG-to-Text decoding framework that mitigates semantic bias and signal neglect. By decoupling high-level semantic attributes to guide generation and redesigning the encoder-LLM interaction via a Q-K-V injection workflow, our approach enforces signal-grounded decoding rather than relying on linguistic priors. Furthermore, we expose the "BLEU Trap" in current evaluation methods and establish a more rigorous protocol based on teacher-forcing-free inference, retrieval and distribution-level metrics, and noise and time-shuffled EEG controls. On these standards, SemKey achieves state-of-the-art performance, decoding diverse and semantically faithful text from noisy EEG.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Compuational Linguistics*, pp. 249–256. Association for Computational Linguistics, 2006.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Ciaran Cooney, Rafaella Folli, and Damien Coyle. Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from eeg. In *2018 29th Irish Signals and Systems Conference (ISSC)*, pp. 1–7. IEEE, 2018.

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, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *arXiv preprint arXiv:2309.14030*, 2023.

Chen Feng, Lu Cao, Di Wu, En Zhang, Ting Wang, Xiaowei Jiang, Jinbo Chen, Hui Wu, Siyu Lin, Qiming Hou, et al. Acoustic inspired brain-to-sentence decoder for logosyllabic language. *Cyborg and Bionic Systems*, 6: 0257, 2025.

Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664–1671, 2015.

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):1–13, 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 Twelfth Language Resources and Evaluation Conference*, pp. 138–146, 2020.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

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.

Marcel A Just and Patricia A Carpenter. A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1):122, 1992.

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

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Marta Kutas and Steven A Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of neuroscience*, 31(8):2906–2915, 2011.

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

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*, 2024.

Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. Eeg2text: Open vocabulary eeg-to-text translation with multi-view transformer. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 1824–1833. IEEE, 2024.

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

Michela Lorandi and Anja Belz. How to control sentiment in text generation: A survey of the state-of-the-art in sentiment-control techniques. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pp. 341–353, 2023.

Jingyu Lu, Haonan Wang, Qixiang Zhang, and Xiaomeng Li. A cognitive process-inspired architecture for subject-agnostic brain visual decoding. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

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

James A Michaelov, Megan D Bardolph, Cyma K Van Petten, Benjamin K Bergen, and Seana Coulson. Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of language*, 5(1): 107–135, 2024.

Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka, and Shounak Roychowdhury. Thought2text: text generation from eeg signal using large language models (llms). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3747–3759, 2025.

Jerrin Thomas Panachakel, AG Ramakrishnan, and TV Ananthapadmanabha. A novel deep learning architecture for decoding imagined speech from eeg. *arXiv preprint arXiv:2003.09374*, 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. On accurate evaluation of gans for language generation. *arXiv preprint arXiv:1806.04936*, 2018.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867, 2020.

Yitian Tao, Yan Liang, Luoyu Wang, Yongqing Li, Qing Yang, and Han Zhang. See: Semantically aligned eeg-to-text translation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Haonan Wang, Jingyu Lu, Hongrui Li, and Xiaomeng Li. Zebra: Towards zero-shot cross-subject generalization for universal brain visual decoding. *arXiv preprint arXiv:2510.27128*, 2025a.

Haonan Wang, Qixiang Zhang, Lehan Wang, Xuanqi Huang, and Xiaomeng Li. Neurons: Emulating the human visual cortex improves fidelity and interpretability in fmri-to-video reconstruction. *arXiv preprint arXiv:2503.11167*, 2025b.

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, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

Ziyi Zeng, Zhenyang Cai, Yixi Cai, Xidong Wang, Junying Chen, Rongsheng Wang, Yipeng Liu, Siqi Cai, Benyou Wang, Zhiguo Zhang, et al. Wavemind: Towards a conversational eeg foundation model aligned to textual and visual modalities. *arXiv preprint arXiv:2510.00032*, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

Xiang Zhang, Lina Yao, Quan Z Sheng, Salil S Kanhere, Tao Gu, and Dalin Zhang. Converting your thoughts to texts: Enabling brain typing via deep feature learning of eeg signals. In *2018 IEEE international conference on pervasive computing and communications (PerCom)*, pp. 1–10. IEEE, 2018.

Hui Zheng, Haiteng Wang, Weibang Jiang, Zhongtao Chen, Li He, Peiyang Lin, Penghu Wei, Guoguang Zhao, and Yunzhe Liu. Du-in: Discrete units-guided mask modeling for decoding speech from intracranial neural signals. *Advances in Neural Information Processing Systems*, 37:79996–80033, 2024.

Jinzhao Zhou, Zehong Cao, Yiqun Duan, Connor Barkley, Daniel Leong, Xiaowei Jiang, Quoc-Toan Nguyen, Ziyi Zhao, Thomas Do, Yu-Cheng Chang, et al. Pretraining large brain language model for active bci: Silent speech. *arXiv preprint arXiv:2504.21214*, 2025.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

Supplementary Material

Appendix Contents

A Statistical Significance and Robustness **13**

B Ablation on Training Strategies **14**

C Implementation Details **14**

C.1 Model Architecture & Hyperparameters 14

C.2 Training Protocol 15

C.3 Dataset & Preprocessing 15

D Noise Degradation and Time-Shuffle Experiments **17**

D.1 Noise Degradation 17

D.2 Time-Shuffle Test 18

E Ablation Study on Label Predictions **19**

E.1 Two-Route Ablation 19

F Language Model Scale **20**

G Diagnosis of the "BLEU Trap" **20**

G.1 Semantic Bias in MTV Augmentation 20

G.2 Qualitative Demonstration and Prefix Stripping 22

H Qualitative Samples: Signal Dependency Verification **22**

I Relation to Jo et al. (2024) **22**

J Additional Metric: BERTScore **23**

K Limitations and Future Directions **23**

L AI Use Statement **27**

Reproducibility Statement **28**

A Statistical Significance and Robustness

To ensure the robustness and reproducibility of our results, we evaluated the models across 7 independent runs using different random seeds. Each seed jointly controls training initialization, data order, and retrieval sampling. The performance variations are reported as the mean $\pm$ standard deviation. Furthermore, to verify that the improvements achieved by our proposed method are not due to random variance, we conducted independent two-sample t-tests comparing our model against the strongest baseline, GLIM.

Table a: Performance comparison with standard deviations over 7 random seeds. Statistical significance of the improvement over GLIM is denoted by p-values.

Model	2-WAY (%)	4-WAY (%)	10-WAY (%)	24-WAY (%)	H. ENT	S-BLEU (%) (↓)	FD (↓)	SBERT (↑)
GLIM (Liu et al., 2025)	66.8 ± 0.58	42.4 ± 0.54	21.9 ± 0.61	10.7 ± 1.88	2.18 ± 1.02	90.8 ± 4.01	0.57 ± 0.053	0.194 ± 0.008
SEMKEY (Ours)	$\mathbf{72.2 \pm 0.52}$	$\mathbf{50.4 \pm 0.27}$	$\mathbf{29.5 \pm 0.38}$	$\mathbf{16.2 \pm 0.20}$	$\mathbf{6.21 \pm 0.54}$	$\mathbf{52.0 \pm 1.55}$	$\mathbf{0.26 \pm 0.041}$	$\mathbf{0.221 \pm 0.004}$
p-value	$p < 0.001^{***}$	$p < 0.001^{***}$	$p < 0.001^{***}$	$p < 0.001^{***}$	$p < 0.001^{***}$	$p < 0.001^{***}$	$p < 0.001^{***}$	$p < 0.001^{***}$

B ABLATION ON TRAINING STRATEGIES

We perform a comprehensive analysis of our training configuration, focusing on LLM tuning methods, data augmentation, and encoder optimization (Table b).

LLM Tuning Strategy : We compare *Frozen*, *Full Fine-tuning*, and *LoRA*. Keeping the LLM *Fully Frozen* yields the best performance among the evaluated strategies. Full fine-tuning and LoRA reduce 4-way retrieval accuracy from 50.4% to 47.9% and 49.1%, respectively, and increase FD from 0.26 to 0.29 and 0.30. These results favor keeping the LLM frozen while training the EEG encoder and projection layer in our setting; the lower performance of the tuning alternatives may reflect overfitting in this limited-data regime.

Impact of MTV: The *Multiple Text Variants (MTV)* augmentation improves retrieval accuracy and distributional alignment: adding MTV raises 4-way retrieval accuracy from 37.8% to 50.4% and reduces FD from 0.63 to 0.26 (Table b).

Stage 2 EEG Encoder: We investigate the status of the EEG encoder during the generative stage. Results show that *Unfreezing (Training)* the encoder in Stage 2 outperforms keeping it fixed. This indicates that while Stage 1 provides a strong initialization via contrastive alignment, further fine-tuning the encoder specifically for the generative objective helps to bridge the remaining modality gap and improves the final decoding accuracy.

Table b: Ablation study on training strategies. We analyze the impact of LLM tuning methods, Data Augmentation (MTV), and Stage 2 Encoder status. Best configurations are highlighted in bold.

CONFIGURATION	4-WAY	24-WAY	FD
1. LLM Tuning Strategy			
LLM FULL FINE-TUNE	47.9%	14.8%	0.29
LLM + LORA	49.1%	14.8%	0.30
LLM FROZEN (OURS)	**50.4%**	**16.2%**	**0.26**
2. Data Augmentation			
W/O MTV	37.8%	8.7%	0.63
W/ MTV (OURS)	**50.4%**	**16.2%**	**0.26**
3. Stage 2 EEG Encoder			
ENCODER FIXED	44.2%	11.6%	0.51
ENCODER TRAINABLE (OURS)	**50.4%**	**16.2%**	**0.26**

C IMPLEMENTATION DETAILS

C.1 MODEL ARCHITECTURE & HYPERPARAMETERS

The SEMKEY framework is implemented in PyTorch, utilizing the PyTorch-Lightning library for modularity. Our architecture builds upon the GLIM (Liu et al., 2025) backbone, employing *Flan-T5-Large* (via HuggingFace Transformers) as the pre-trained language model. The EEG Encoder consists of 6 encoder blocks and 6 query-based decoder blocks (hidden size 128, 8 attention heads each), followed by an aligner that projects the features into the language model space ($d_{model} = 1024$). The resulting EEG embeddings ($\mathbf{E}_{eeg}$) possess a temporal dimension of $L = 96$. For the auxiliary decoding tasks (Sentiment, Topic, Length, and Surprisal), we implemented four dedicated heads, each consisting of a three-layer MLP with hidden dimensions of $[512, 256, 128]$. The Topic labels are [Biographies and Factual Knowledge, Movie Reviews and Sentiment], following the official ZuCo

corpus categorization, while the Sentiment labels are [non_neutral, neutral]. The total parameter count of the model is approximately 799M.

Our framework employs a reconstruction loss ($\mathcal{L}_{recon}$), implemented as the standard autoregressive cross-entropy objective derived from the frozen Flan-T5. To enforce latent space consistency, we use two separately weighted losses: commitment ($\mathcal{L}_{commit}$) and contrastive alignment ($\mathcal{L}_{clip}$). Specifically, the commitment loss minimizes the Mean Squared Error (MSE) between the generated EEG embeddings ($\mathbf{E}_{eeg}$) and the target text representations from Flan-T5. Meanwhile, the contrastive loss optimizes the cosine similarity between the EEG vector ($\mathbf{v}_{eeg}$) and the corresponding text embedding. Finally, we optimize auxiliary tasks using cross-entropy for classification and MSE for regression.

During the generation stage, the attributes predicted by these auxiliary heads are explicitly injected into the language model to guide the decoding trajectory. Specifically, we integrate these semantic anchors using the following structured prompt template:

> System: Based on the following EEG signals, reconstruct the text. The length of the sentence is [$\hat{y}_{len}$] words. Surprisal: [$\hat{y}_{spr}$]. Sentiment: [$\hat{y}_{stm}$]. Topic: [$\hat{y}_{tpc}$].

C.2 Training Protocol

The training process consists of two stages. All experiments were conducted on a single *NVIDIA H100 (80GB HBM3)* GPU.

Stage 1 (SemKey Parallel): This stage focuses on aligning EEG representations and training the auxiliary heads. We trained only the EEG encoder and auxiliary heads (approx. 16.7M trainable parameters) for 50 epochs. Model checkpoints were selected based on the validation accuracy and the minima of the regression losses. We employed a cosine annealing learning rate scheduler with a 15-epoch linear warmup, bounded between 1.0×10^{-5} and 2.0×10^{-4}. The loss weights for this stage are detailed in Table d. The approximate training duration was 7 hours.

Stage 2 (Generative Decoding): For the generative phase, we utilized a native PyTorch pipeline initialized with the optimal Stage 1 checkpoint. We kept the LLM frozen because this setting achieved the best retrieval accuracy and FD among the tuning strategies evaluated (Table b). The EEG encoder and the projector were trained jointly for 25 epochs, with the auxiliary attribute losses kept active. The attribute labels were predicted and cached after Stage 1 training; updating the encoder in Stage 2 does not alter these cached labels or the resulting prompts. We used a cosine annealing scheduler with a 3-epoch warmup. The learning rates were layer-specific: 8.0×10^{-5} for the Stage 1 encoder and 2.0×10^{-4} for the projector. A weight decay of 0.01 was applied. The Stage 2 objective is the weighted sum $\mathcal{L}_{Stage2} = 1.5\mathcal{L}_{recon} + 0.25(\mathcal{L}_{stm} + \mathcal{L}_{tpc} + \mathcal{L}_{len} + \mathcal{L}_{spr})$; contrastive and commitment losses are disabled (Table d). At test time, text is generated with beam search using the default decoding parameters. The training time of Stage 2 was approximately 8 hours.

C.3 Dataset & Preprocessing

We utilized the complete *ZuCo benchmark* (combining ZuCo 1.0 and 2.0) (Hollenstein et al., 2018; 2020) for our experiments. This benchmark provides 128-channel EEG recordings sampled at 500Hz, which were simultaneously collected with text corpora during English sentence reading tasks. It contains over 22,000 sentence-level EEG-text pairs recorded from a total of 30 participants (12 subjects in ZuCo 1.0 and 18 subjects in ZuCo 2.0). The dataset encompasses three reading paradigms: Sentiment Reading (SR; ZuCo 1.0 only), Normal Reading (NR; passive reading), and Task-Specific Reading (TSR; active reading with comprehension questions).

Notably, the text corpora include sentences from established datasets such as SST and Wikipedia. This introduces representative data heterogeneity across reading paradigms, corpora, sessions, and subjects, establishing ZuCo as a strong and challenging prototypical setting for training generalizable models and evaluating semantic alignment.

Table c: **Detailed Architecture and Hyperparameters of SEMKEY.** The model consists of a Prompt Embedder, a dual-stream EEG Encoder, an Alignment module, multiple Task Heads, and an EEG Projector, utilizing a frozen Flan-T5 Large encoder and decoder.

Module	Component	Parameter	Value / Configuration
PromptEmbedder	Embedding Layers	Input Dim / Categories	128 / (3, 3, 30)
		Dropout	0.0
	Input-Output	Shape	Input: $(n, 3) \rightarrow$ Output: $(n, 128)$
EEGEncoder	Encoder Blocks	Sequence Length	Input: 1280
		Layers / Hidden	6 Layers / 128 Dim
		Heads / Dropout	8 Heads / 0.1
	Decoder Blocks	Sequence Length	Output: 96
		Layers / Query Dim	6 Layers / 128 Dim
		Heads / Dropout	8 Heads / 0.1
	Positional Emb.	Type	1D Sinusoidal (Channel Weights: Learnable)
Aligner	Cross-Attention	Dimensions	Hidden: 128 / Embed: 1024
		Heads / Dropout	8 Heads / 0.1
	Training Objectives	Loss Types	CLIP (Cosine Similarity) + MSE (Commitment)
Task Heads	Sentiment Classifier	Structure	MLP [512, 256, 128], Dropout 0.3, Classes: 2
	Topic Classifier	Structure	MLP [512, 256, 128], Dropout 0.3, Classes: 2
	Length Regressor	Structure	MLP [512, 256, 128], Dropout 0.3, Target: 1
	Surprisal Regressor	Structure	MLP [512, 256, 128], Dropout 0.3, Target: 1
EEG Projector	$\mathbf{W}_{proj}$	Structure	Linear $1024 \rightarrow 1024$ ($\approx$1.05M params), trainable in Stage 2
Text Model	T5 Encoder	Foundation Model	`flan-t5-large`
		Hidden Size	1024 (Frozen parameters)
	T5 Decoder	Status	Frozen; cross-attends to projected EEG (Stage 2)
Training Config	Optimization	Batch Size	**Stage 1** 72 (per GPU)
			Stage 2 24 (per GPU)
		Learning Rate	**Stage 1** Base: 2×10^{-4} (Min: 1×10^{-5})
			Stage 2 Encoder: 8×10^{-5} / Projector: 2×10^{-4}
		Scheduler	**Stage 1** Warmup: 15 epochs / Max: 50 epochs
			Stage 2 Cosine, Warmup: 3 epochs / Max: 25 epochs
		Weight Decay	**Stage 2** 0.01
		Precision	`bfloat16-mixed`

Consistent with the GLIM preprocessing protocol (Liu et al., 2025), the raw EEG waves (filtered through a 0.1Hz to 100Hz frequency band) were refined to a standard 104-channel subset. The signals were then resampled to 128 Hz and padded to a fixed duration of 10 seconds (1,280 samples).

Table d: **Hyperparameter settings for Loss Weights across training phases.** We compare the weight (λ) configurations between Stage 1 (Representation Learning) and Stage 2 (Generative Decoding). Note that alignment losses are disabled in Stage 2 to focus on generative fidelity.

Loss Component	Weight (λ)	
	Stage 1	**Stage 2**
Primary Objectives		
Contrastive Loss ($\mathcal{L}_{clip}$)	0.5	0.0
Commitment Loss ($\mathcal{L}_{commit}$)	0.7	–
Reconstruction Loss ($\mathcal{L}_{recon}$)	0.5	1.5
Auxiliary Objectives (Attribute Heads)		
Sentiment ($\mathcal{L}_{stm}$)	0.3	0.25
Topic ($\mathcal{L}_{tpc}$)	0.3	0.25
Length ($\mathcal{L}_{len}$)	0.9	0.25
Surprisal ($\mathcal{L}_{spr}$)	0.3	0.25

Data split. To prevent data leakage during model evaluation, we partitioned the dataset based on unique stimulus texts, guaranteeing that no sentence appears in more than one subset. Recognizing the benchmark's intentional text overlap across different subjects, reading paradigms, and source datasets, we first aggregated all such overlapping samples exclusively into the training set. Subsequently, the remaining unique sentences were randomly allocated in a stratified manner using a fixed seed. This partitioning strategy yields a final split of 17,908, 2,200, and 2,227 sample pairs for the training, validation, and test sets, respectively, maintaining an approximate 8:1:1 ratio.

Data Augmentation. To enhance the semantic robustness, we adopt the *Multiple Text Variants (MTV)* strategy (Liu et al., 2025), using paraphrases of ground-truth text during Stage 2 training. This encourages the model to focus on invariant semantics rather than overfitting to specific syntax. Retrieval accuracy and SBERT are always computed against the original reference sentence only; the effect of MTV on BLEU is analyzed in Appendix G.

Spectral Whitening. Empirical analysis of the raw EEG spectrum revealed a dominance of high-amplitude noise in the low-frequency bands (characteristic $1/f$ noise) and baseline drift, which often obscures subtle semantic patterns in higher frequencies. To mitigate this, we applied a *spectral whitening* transformation. This process normalizes the power spectral density across frequency bands, effectively flattening the spectrum. By suppressing the overpowering low-frequency components, the model can more effectively attend to the fine-grained, high-frequency oscillatory features associated with linguistic processing.

D Noise Degradation and Time-Shuffle Experiments

D.1 Noise Degradation

To further investigate the robustness of our framework and the extent to which the decoding process relies on the actual physiological signals rather than the language model's prior, we conducted a noise perturbation test. We progressively mixed the raw input EEG signals with Gaussian noise at varying proportions. By adjusting the noise mixing ratio $\alpha \in \{25\%, 50\%, 75\%, 100\%\}$, we evaluated the corresponding degradation in decoding performance.

As presented in Table e, all evaluation metrics decay steadily as the noise ratio α increases. Notably, under the pure noise condition ($\alpha = 100\%$), the 4-way and 24-way retrieval accuracies drop to 25.00% and 4.17% (chance level). Concurrently, the FD rises from 0.26 to 1.27.

Table f illustrates changes in the specificity and structure of the generated outputs as noise increases. At low-to-moderate noise ratios ($\alpha = 25\%, 50\%$), the generated names, dates, and descriptions change, while recognizable sentence structures remain. At $\alpha = 75\%$, the output becomes a short, generic sentence.

Under pure noise ($\alpha = 100\%$), the generation collapses into a sequence of meaningless repetitive tokens (e.g., "W. W. W."). Together with the drop in retrieval accuracy and increase in FD, this dis-

ruption of linguistic structure and semantic matching supports the input sensitivity of our framework and provides evidence of signal dependency.

Table e: Quantitative Performance Degradation under Increasing Noise Ratios.

Condition	**4-Way (%)**	**24-Way (%)**	**FD (↓)**
Original	50.48	16.24	0.26
Noise Ratio ($\alpha = 25\%$)	48.39	16.08	0.29
Noise Ratio ($\alpha = 50\%$)	42.78	11.13	0.31
Noise Ratio ($\alpha = 75\%$)	34.91	7.19	0.47
Noise Ratio ($\alpha = 100\%$)	25.00	4.17	1.27

Table f: Qualitative Examples of Decoded Sentences across Different Noise Ratios.

Condition	**Decoded Sentence**
GT	Aldous Leonard Huxley *(July 26, 1894 - November 22, 1963)* was a **British writer** who emigrated to the **United States**.
Original	Nicola Carvinsky *(May 14, 1856 - May 22, 1915)* was an **American** politician, actress, producer, and **writer**.
$\alpha = 25\%$	Nicola Carmina Coppola (*born May 14, 1971 in New York City*, New York, is an **American** film director, actress, producer, and **writer**.
$\alpha = 50\%$	Nicola Carolini was an **American** film director, actress, producer, and **writer**.
$\alpha = 75\%$	Jeremy Schwartzman is an **American** actor.
$\alpha = 100\%$	Walleth W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W. W

D.2 Time-Shuffle Test

The noise test proposed by Jo et al. (2024) replaces EEG with Gaussian noise, whose amplitude and spectrum differ from those of real EEG, so the response may partly reflect an out-of-distribution effect. As a complementary control, we split each EEG input into 20 temporal segments and randomly permute them. This preserves the amplitude distribution and structure within each segment while disrupting the order between segments. The attribute prompt is re-inferred from the shuffled EEG, so both the prompt and the cross-attended EEG embeddings come from the shuffled input. This test complements the Gaussian-noise control and is designed to reduce the influence of out-of-distribution effects by retaining more of the original signal structure.

Shuffling lowers the retrieval accuracy of SemKey at all difficulty levels (2-, 4-, 10-, and 24-way from 72.29%, 50.48%, 29.56%, and 16.24% to 57.14%, 32.05%, 14.64%, and 6.58%) and its SBERT from 0.221 to 0.147. These drops support sensitivity to the disruption of temporal order and are larger than those of GLIM at all levels (Table g; e.g., 4-way -18.43 vs. -11.98 and SBERT -0.074 vs. -0.039). The Prior Model (w/o TF) is unaffected and outputs the same sentence for every input, indicating complete signal neglect.

The two models exhibit different failure patterns (Table h). For SemKey, disrupting the temporal structure disperses the output distribution: Head Entropy rises from 6.21 to 6.57 and Self-BLEU falls from 52.08% to 29.73%, indicating that the outputs become dissimilar to one another and lose linguistic structure (e.g., "... ... ...", "PG-13 PG-13 PG-13", or "He used a shaman and a shaman. . . "). This is the expected response of a model that has no reliable signal to rely on. GLIM changes in the same direction, but from a different starting point to a different endpoint. Even with disrupted input, its Head Entropy (4.64) remains well below that of SemKey (6.57). GLIM does not leave the template regime, suggesting that its outputs are driven by learned templates rather than by the input signal.

Table g: Retrieval accuracy and SBERT with intact and time-shuffled EEG input. Δ denotes the change from intact to time-shuffled input. Chance levels of 2-, 4-, 10-, and 24-Way retrieval are 50.00%, 25.00%, 10.00%, and 4.17%. A larger drop indicates stronger dependence on the temporal order of the input.

Model	Input	2-Way (%)	4-Way (%)	10-Way (%)	24-Way (%)	SBERT
Prior Model (w/o TF)	Intact	50.00	25.00	10.00	4.17	0.163
	Time-shuffled	50.00	25.00	10.00	4.17	0.163
	Δ	0.00	0.00	0.00	0.00	0.000
GLIM	Intact	66.83	42.45	21.91	10.74	0.194
	Time-shuffled	56.90	30.47	13.96	5.87	0.155
	Δ	−9.93	−11.98	−7.95	−4.87	−0.039
SEMKEY	Intact	72.29	50.48	29.56	16.24	0.221
	Time-shuffled	57.14	32.05	14.64	6.58	0.147
	Δ	**−15.15**	**−18.43**	**−14.92**	**−9.66**	**−0.074**

Table h: Output diversity with intact and time-shuffled EEG input. Unlike retrieval accuracy, these metrics are read by their values rather than their changes (Section 3.3). GLIM retains low prefix diversity even when the temporal order of the input is disrupted.

Model	Input	H. Ent.	S-BLEU (%)
Prior Model (w/o TF)	Intact	0.00	100
	Time-shuffled	0.00	100
GLIM	Intact	2.18	90.86
	Time-shuffled	4.64	50.35
SEMKEY	Intact	6.21	52.08
	Time-shuffled	6.57	29.73

E ABLATION STUDY ON LABEL PREDICTIONS

To further investigate the impact of attribute prediction in guiding the model, we conducted experiments under extreme adversarial conditions: (1) Corrupted Prompt (replacing labels with "XXXX"), and (2) Reversed Prompt (injecting the exact opposite of Stage 1 predictions).

Table i: Results of the Ablation Study on Label Predictions.

Condition	4-Way (%)	24-Way (%)	FD (↓)
Original	50.48	16.24	0.26
Prompt - XXXX	36.05	8.21	0.401
Prompt - Reversed	35.33	8.16	0.451

Under these corrupted/reversed conditions, we observe a severe degradation in generation quality and performance metrics (As shown in table i). Specifically, reversing the attributes changes the output sentiment (Example in table j). These results further support that the improvements in our method are primarily driven by accurate attribute predictions from Stage 1, and not merely by prompt templates.

E.1 TWO-ROUTE ABLATION

SEMKEY passes EEG information to the language model through two routes: the attribute prompt, whose hidden states serve as the queries, and the projected EEG embeddings, which serve as the keys and values. To examine the contribution of each route, we replace the prompt with the dummy ("XXXX") prompt, the EEG input with Gaussian noise, or both, at inference. As shown in Table k, replacing either route lowers the 4-way retrieval accuracy, and replacing both reduces it to chance level, so both the prompt and the cross-attended EEG carry information about the input.

Table j: Decoded Text Examples for Different Conditions.

Condition	Decoded Text
Ground Truth	An awful movie that will only satisfy the most emotionally malleable of filmgoers.
Original Prompt	Neither the movie's plot nor its characters are particularly engaging.
XXXX Replacement	The movie is a bit too long, but it's still entertaining.
Reversed Prompt	Whether you're looking for a quiet, unpretentious, or a sexy, or a sexy...

Table k: 4-way retrieval accuracy (%) when the prompt route and/or the EEG route is replaced at inference. Chance level is 25%.

	Real EEG	**Noise EEG**
Real Prompt	50.48	31.91
Dummy Prompt ("XXXX")	36.05	25.00

F Language Model Scale

We further examine whether a larger language model can replace the attribute guidance. Table l reports the 4-way retrieval accuracy with and without the *Length* attribute for two backbones. Removing *Length* lowers the accuracy by a similar margin for Flan-T5-Large (50.4% to 43.5%) and T5-XL (51.2% to 44.7%), while the larger backbone brings only a small gain. We also experimented with BART, LLaMA, and T5-XL backbones; without attribute guidance, the repetitive generation persisted with all of them. Within the Flan-T5-Large setting, keeping the language model frozen also outperforms full fine-tuning and LoRA (Table b).

Table l: Effect of the language model scale: 4-way retrieval accuracy (%) with and without the *Length* attribute.

Language Model	**Full Prompt**	**w/o Length**
Flan-T5-Large (Ours)	50.4	43.5
T5-XL	51.2	44.7

G Diagnosis of the "BLEU Trap"

The "BLEU Trap" refers to the phenomenon in EEG-to-Text generation where models achieve high n-gram matching scores (BLEU) by memorizing high-frequency templates, despite failing to decode the correct semantic content (Hallucination). Conversely, models that attempt genuine semantic decoding may suffer lower scores due to a lack of template overlap.

The trap results from the interaction of the metric, the dataset, and the evaluation protocol. BLEU is a clipped n-gram precision that weights function words and content words equally (Papineni et al., 2002; Callison-Burch et al., 2006). The ZuCo sentences come from a few sources (movie reviews and mostly biographical Wikipedia sentences) with stereotyped openings, and stopwords and template phrases account for more than 60% of the tokens. The MTV protocol then scores each output against several paraphrased references, which gives generic n-grams more chances to match. MTV scoring amplifies the trap but is not required for it: under single-reference scoring, the *Prior Model (w/o TF)*, which outputs identical sentences for all inputs, still obtains the highest BLEU (Table 2). GLIM exploits the same templates: only 41.8% of its outputs are unique, and four sentence openings (e.g., "He was a") account for about 69% of them.

G.1 Semantic Bias in MTV Augmentation

In this section, we analyze the quality of the Multiple Text Variants (MTVs) dataset strategy adopted from GLIM (Liu et al., 2025). While intended to enhance supervision by providing paraphrased references, a critical inspection reveals a severe **semantic pattern bias**.

As detailed in Table m, the augmentation algorithms (e.g., lexical simplification, rewriting) tend to converge on rigid syntactic templates rather than diverse semantic expressions. Specifically:

- **Sentiment Task:** A disproportionate number of variants begin with static headers like *"The movie..."* or *"It..."*.
- **Topic Task:** The biography descriptions are dominated by the *"He was..."* pattern.

This structural homogeneity artificially reduces the perplexity of the target distribution, potentially encouraging models to overfit to these static headers rather than decoding the unique neural semantics of the input.

Note that MTV plays different roles in training and in evaluation. As a training augmentation it is beneficial: removing it lowers the 4-way retrieval accuracy of SEMKEY from 50.4% to 37.8% and raises FD from 0.26 to 0.63 (Table b). As additional references at scoring time, however, the templated variants reward generic outputs and inflate BLEU (Table 2). We use MTV for training and report BLEU with multiple references in the main comparison (Table 1) to illustrate the BLEU Trap; retrieval accuracy and SBERT are computed against the original reference sentence only.

Table m: Comparison of text variants across different tasks. **Note the high recurrence of specific patterns:** "The movie" for sentiment tasks and "He was" for topic tasks (highlighted in bold). This repetition introduces semantic bias into the training data.

Variant Type	Sentiment Task (Movie Review)	Topic Task (Biography)
Original Ground Truth	Presents a good case while failing to provide a reason for us to care beyond the very basic dictums of human decency.	He served as the Governor of California from January 2, 1939 until January 4, 1943, and was the only Democrat elected to that office between 1898 and 1954.
Lexical Simplification (v0)	Makes a good point, but doesn't give us a reason to care about the characters.	**He was** the Governor of California from 1939 to 1943 and the only Democrat in that job between 1898 and 1954.
Lexical Simplification (v1)	**The movie** makes a good point, but doesn't give us a reason to care about it.	**He was** the Governor of California from 1939 to 1943 and was a Democrat.
Semantic Clarity (v0)	**The movie** makes some good points, but it does not give us a compelling reason to emotionally invest in its message beyond feeling obligated to support basic human values.	Culbert Olson, a Democrat, served as the Governor of California from January 2, 1939 until January 4, 1943.
Semantic Clarity (v1)	**The movie** presents a good argument but fails to evoke strong emotions beyond basic sympathy.	Culbert Olson, a Democrat, served as the Governor of California from January 2, 1939 until January 4, 1943.
Syntax Simplification (v0)	**The movie** makes some good points, but it fails to evoke strong emotions beyond basic human empathy.	**He was** a Democrat and served as the Governor of California from 1939 to 1943.
Syntax Simplification (v1)	**The movie** makes some good points, but fails to evoke strong emotions beyond basic human empathy	Culbert Olson was a Democrat and served as the Governor of California from 1939 to 1943.
Naive Rewritten	It presents a good case while failing to provide a reason for us to care beyond the very basic dictums of human decency.	He served as the Governor of California from January 2, 1939 until January 4, 1943, and was the only Democrat elected to that office between 1898 and 1954.
Naive Simplified	It's a good case, but it doesn't provide a reason for us to care beyond the very basic dictums of human decency.	**He was** the first African-American to serve as governor of California, and the first African-American to serve as governor of the United States.

G.2 Qualitative Demonstration and Prefix Stripping

Qualitative Demonstration. Table o presents a side-by-side comparison of model outputs against the Ground Truth (GT). The baseline *GLIM* model exhibits a clear tendency toward hallucination, frequently generating generic templates such as *"He was ...", "The Movie"*. While these outputs often contradict the factual content of the input, they align perfectly with the stylistic patterns in the reference set, resulting in artificially inflated scores (e.g., *54.6%*).

In contrast, SemKey avoids this template collapse. Although avoiding "safe" static prefixes may sometimes yield lower surface-level BLEU scores, our method demonstrates a more genuine decoding process. It captures both *high-level semantic attributes* (e.g., maintaining the correct sentiment where baselines fail) and *specific entities*, reflecting superior semantic fidelity over shallow pattern matching.

Quantitative Verification (Prefix Stripping). To investigate the influence of recurring syntactic templates on metric inflation, we conducted a "Prefix-Stripped" evaluation by removing common headers (e.g., "The movie", "He was") from both ground truth and predictions. As shown in Table n, the baseline model exhibits a notable performance drop (e.g., *-38.8%* in BLEU-2) when these prefixes are excluded. This suggests that the reported performance is significantly overstated by fitting these rigid patterns rather than purely relying on semantic decoding.

Table n: **Quantitative evidence of the "BLEU Trap."** We compare the baseline GLIM model on the standard test set (*Original*) versus the "Prefix-Stripped" version (*Stripped*) using BLEU score without MTV. The performance drop indicates the model's reliance on prefix memorization. BLEU-3/4 are close to zero, so their relative changes are less reliable.

Metric	Original	Stripped	Drop (Δ)
BLEU-1	7.84%	6.10%	**-22.2%**
BLEU-2	1.28%	0.78%	**-38.8%**
BLEU-3	0.18%	0.04%	**-78.7%**
BLEU-4	0.008%	0.006%	**-19.7%**

H Qualitative Samples: Signal Dependency Verification

To further validate the signal dependency discussed in the main experiments, we present a qualitative comparison of model outputs when the raw EEG input is replaced by pure Gaussian noise.

As shown in Table p, GLIM retains recognizable corpus patterns under Gaussian noise (e.g., repeated names or "He was the wife of..."), consistent with persistent reliance on learned language priors even when the input is disrupted.

For **SemKey**, Gaussian noise substantially disrupts linguistic structure, producing unstructured tokens. Together with the reduced semantic matching reported in Section 4.2, this supports input sensitivity and signal dependency.

I Relation to Jo et al. (2024)

Jo et al. (2024) showed that evaluating EEG-to-Text models with teacher forcing leaks the ground-truth prefix, and that models trained with random noise instead of EEG reach comparable scores. We take this finding as our starting point and do not claim it as our contribution. Our work extends it in three ways. (i) We study the setting without teacher forcing at inference, and show that GLIM, which is evaluated in this setting, still exhibits persistent corpus-template reliance under noise and time-shuffled input (Section 4.2). (ii) We show that BLEU can still be inflated without leakage: template outputs obtain high scores, removing common prefixes lowers GLIM's BLEU-2 by 38.8%, and MTV scoring further amplifies the effect (Appendix G). (iii) We propose a model and evaluate it under leakage-free inference with retrieval, semantic, and distribution-level metrics, together with noise and time-shuffled EEG controls. Since Jo et al. (2024) continue to use BLEU once teacher forcing is removed, our findings complement theirs.

Table o: Qualitative demonstration of the **"BLEU Trap"**. We compare generated samples against the Ground Truth (GT). GLIM often achieves superior BLEU scores by outputting generic templates that match the references, even when the generated content is factually wrong (Hallucination). SEMKEY, while scoring lower, avoids these templates and attempts diverse generation with better semantic fidelity.

Type	Sentence Content	B-1 @ MTV	B-1 @ w/o MTV
GT	*The cumulative effect of the movie is repulsive and depressing.*	-	-
GLIM	**The movie** is surprisingly romanticized.	**49.1%**	22.1%
Ours	Despite its poor quality, the movie falls apart because of its poor special effects.	21.4%	21.4%
GT	*He was then hired by the legal department of CBS subsidiary Columbia Records.*	-	-
GLIM	**He was** a member of the Communist Party in the 1930s.	**54.6%**	30.3%
Ours	Upon graduation from Northwestern University, he was employed by the Lilith Research Institute...	20.0%	20.0%
GT	*... something appears to have been lost in the translation this time.*	-	-
GLIM	It's not a bad movie, but it's not a great one either.	**41.7%**	0.0%
Ours	... the way the movie is structured is a bit strange.	36.4%	16.6%
GT	*Taylor was born with dual British and American citizenship.*	-	-
GLIM	**He was** a follower of Ronald Reagan.	**28.6%**	10.7%
Ours	Lucay Fisher, an American, was born on October 21, 1956, to American parents.	23.1%	23.1%
GT	*He is married to singer Chynna Phillips.*	-	-
GLIM	He had seven children with his wife.	**42.9%**	14.3%
Ours	During his career, he married Joyce Halverson in 1951.	22.2%	11.1%
GT	*It's not a particularly good film, but neither is it a monsterous one.*	-	-
GLIM	**The movie** Wedding feels somewhat outdated.	**33.3%**	0.0%
Ours	... and its acting and its performances make it a bit of a letdown.	21.4%	21.4%
GT	*He also was awarded the Presidential Medal of Freedom.*	-	-
GLIM	**He was** also a member of the Royal Family.	**55.6%**	55.6%
Ours	He received a scholarship to the University of Miami for his academic career.	23.1%	23.1%

J ADDITIONAL METRIC: BERTSCORE

For completeness, we also report BERTScore-F1 (Zhang et al., 2020). It is 0.8398 for the *Prior Model (w/o TF)*, 0.8367 for GLIM, and 0.8468 for SEMKEY. Without baseline rescaling, BERTScore is high even for unrelated sentences, so the differences between models are small; SEMKEY still obtains the highest score.

K LIMITATIONS AND FUTURE DIRECTIONS

Despite the advancements discussed, decoding fidelity at the fine-grained word level remains a significant bottleneck. We attribute this limitation largely to the inherent characteristics of the ZuCo benchmark, which contains a high density of low-frequency proper nouns—such as specific names and locations—that are notoriously difficult to resolve from coarse-grained EEG signals (Huth et al., 2016).

The Challenge of Inter-Subject Variability. Moreover, the dataset's collection paradigm, characterized by a multi-subject setting with limited data per individual, necessitates a cross-subject training strategy (Zhou et al., 2025). However, given the substantial inter-subject variability in brain topology and cognitive processing (Finn et al., 2015; Wang et al., 2025a), effective generalization

Table p: Qualitative comparison of generated samples under Gaussian Noise input. **GLIM** retains recognizable corpus templates, whereas **SEMKEY** exhibits marked disruption of linguistic structure, consistent with its sensitivity to input corruption.

Model	Generated Output (Input: Noise)
GLIM	In 1894, Charles Charles F. Cardinals, a U.I. Cardinals, a former RhilBsne Schoine. Charles Adams was the first credited in the U.S. Cardinalsse. He was the wife of the United States Catholic Losman James Kennedy. He was the wife of Lord Schoene-Seey St. Bush. In 1890, Charles Adams, he was the wife of the United States of U.S. Bush. George W. Bushshis in the U.S. Senate in 1962. In 1890, Charles Nelson, sey, sed in the Europeans, sel. He worked as a a bloe of the U.S. Senator. In 1897, Charles Charles Charles Chester, are the British business of Charles Charles Charles F. Cardinal. He was re-inout in the Sty II, which was the British Irish successor to leave the British Irish scheme to the French scheme to lead.
SEMKEY	handsdorfPLexpozeitenkulFBhookeativelyock warfare electronicallyigungaveclosavecjackocyte steer Brüuzidale hookcoukir millyak load dropspar machine|bounce onputoptilujcoupoul dumb wo mislugpaipot and diswohl Wahl Brandon silent nightsence killerapos passive equipmentzar Freizeitlay Dupauseouvre? prosecutioncororg adultsEmpfehlung Sco defense comfort march EdmundicallyENCEoresEMre gauge par Oil dig Separathurstiety sittingpetementspai turkey anchor-greg bag chestworm prime GRA bars history dockdANT SE pun fortune welcome reviewquaSSA Harores angelwaitedlift:lipcrimin costsémi finishing feat step revisionditeITA lateapp of good Women RO Good comic* fair qui Sarah... on steal wa and Valenciacal !biblit Airbnbdrarieshirbow- Openche3),hus, bine –cie Ep unde tolerance Baltimore mois fair vertreten Coleman Solutionprinut bravebervartes vent attendanceumb Ferrnosti pru)pong minig cr Real I con whisk fournicons exllenceloctile*bridge jumping Bir warmerplant Essex stream Adobeeau adidas Swbanfro scenemillionpostrebspace Sitz stalkLR andriskali stencilHu Elle BetterCN UberDF

across subjects remains intrinsically challenging (Zheng et al., 2024; Lu et al., 2026). As illustrated in Table q, while our method significantly reduces the rigid template hallucinations seen in baselines (e.g., repeating "He was a..." for every biography), it still struggles with precise proper noun retrieval. In addition, since the data are split by sentence, the same subjects appear in the training and test sets, and generalization to unseen subjects is not evaluated.

Coarse Attributes. The attributes that guide generation are coarse: topic and sentiment have two classes each, and length and surprisal are scalars. They constrain the global content and form of the output, while word-level information has to come from the cross-attended EEG embeddings.

Absolute Performance and the "BLEU Trap". Furthermore, we candidly acknowledge that the absolute decoding performance of current models, including ours, remains relatively low. For instance, our framework achieves a 4-way retrieval accuracy of 50.48% (chance level 25%) and an SBERT similarity of 0.221 (random-reference baseline 0.109, obtained from reference–reference pairing). Retrieval accuracy measures whether the decoded text can be matched to its reference among candidates, which reflects semantic-level decoding rather than verbatim reconstruction. While these metrics represent a substantial improvement over existing baselines, they clearly fall short of the threshold required for practical, real-world application. Nevertheless, we consider this work a highly meaningful milestone for the field. Beyond the empirical performance gains and the introduction of a promising technical pathway, a primary contribution of our study is exposing the "BLEU Trap." By demonstrating how traditional n-gram metrics can be heavily inflated by rigid template hallucinations, we aim to prevent the community from optimizing for superficial, misleading scores, thereby redirecting focus toward more objective and signal-grounded evaluation paradigms.

Future Directions. Consequently, we advocate that future research should pivot towards **high-resource, single-subject data acquisition**, prioritizing deep longitudinal datasets. Such a shift would minimize topological variance and allow models to focus on learning subtle semantic-neural mappings, thereby unlocking the full potential of neural decoding models.

Table q: Full qualitative comparison (27 Examples). **GT**: Ground Truth. **GLIM**: Baseline. **Ours**: SEMKEY.

Model	Generated Sentence
GT	Coltrane moved to Philadelphia, Pennsylvania in June 1943.
GLIM	He was a member of the Politcaly in 1912 and the Politcaly in 1913.
Ours	During the 1930s, the Bushes moved to Boston, Massachusetts, where he worked for the Hupp Products Company.
GT	He also was awarded the Presidential Medal of Freedom.
GLIM	He was also a member of the Royal Family.
Ours	President Franklin D. Roosevelt was awarded the Nobel Peace Prize in 1911 for his service to the United States...
GT	He started writing his first novel, called The Town and the City.
GLIM	He was an African-American jazz trumpeter, bandleader, singer, and composer.
Ours	Afterwards he worked as a printer for several years before he moved to New York City.
GT	Rodgers shortly fled debt and prosecution by going to Barbados, leaving Deborah behind.
GLIM	He was a member of the United States House of Representatives, the United States Senate...
Ours	During the Great Depression, the family moved to Ohio and rented a farm for the rest of their life...
GT	In 1920, he was nominated Vice President under Warren G.Harding.
GLIM	He was an African-American jazz trumpeter, bandleader, singer, and composer.
Ours	During the 1870s, he worked as a captain of the Army Air Forces as a colonel.
GT	Aldous Leonard Huxley (July 26, 1894 - November 22, 1963) was a British writer who emigrated to the United States.
GLIM	He was an African-American jazz trumpeter, bandleader, singer, and composer.
Ours	Libby Ryan (born June 9, 1916) is an American business executive and a former United States Secretary of Defense.
GT	Taylor was born with dual British and American citizenship.
GLIM	He was the first Whig Party governor and a populist Democrat.
Ours	Libby Carrie Fisher (born June 9, 1916) is an American, and a former United States Secretary of Defense.
GT	He is married to singer Chynna Phillips.
GLIM	He was born in Dublin, Ireland.
Ours	During this time, he married Patricia Arquette on April 8, 1953.
GT	He was then hired by the legal department of CBS subsidiary Columbia Records.
GLIM	He was a member of the Communist Party in the 1930s.
Ours	During the 1931 merger, he became the manager of the new employer of the Brookville plant in Alfred Harriman's company.
GT	Libby was a founding member of the Project for the New American Century.
GLIM	He was a member of the Federalist Party.
Ours	During his tenure as Minister of Agriculture, Henry Ford founded the Union Fire Company...
GT	Hanks is a Democrat and has supported many candidates, including Hilary Clinton...
GLIM	He was the youngest ex-governor in U.S. history.
Ours	Texas Republican James Douglas McKay, who was born on June 24, 1893, in Portland, Oregon...
GT	Ferrer had previously been married to Uta Hagen (1938-1948)...
GLIM	He was born in Cork, Ireland, the son of George H. W. Bush and Mary Lynne Morris.
Ours	Despite her poverty, Grace did not marry Whitney Corri Harris, an Irish Catholic...
GT	In 1962, Clampett created an animated version of the show called Beany and Cecil...
GLIM	Groucho Marx was a notorious snout killer in the Victorian era.
Ours	Libby Ward, the husband of Warren Buffett, has been married to Goldie Hawn...
GT	When Baldwin was young, he had a job as a busboy at famous New York City disco Studio 54.
GLIM	He was the son of Richard Bush and Mary Fairbanks.
Ours	During the year 1839, he was appointed to the board of the Legal Services Corporation.

Continued on next page

Table q – continued from previous page

Model	Generated Sentence
GT	Reagan later married actress Jane Wyman in 1940.
GLIM	He married Laura Welch, a librarian, in 1927.
Ours	During this period, he married Elizabeth Harris, an American actress, on August 10, 2002.
GT	Friedrich Wilhelm Bessel (July 22, 1784 - March 17, 1846) was a German mathematician...
GLIM	Hepburn was re-elected in 1960 but lost to Republican John Grisham.
Ours	Luca Guido Navier, an American rock musician, was born on May 16, 1965...
GT	She is married to cinematographer Christopher Duddy (married December 31, 1996 - present).
GLIM	He was born in Los Angeles to Italian immigrants.
Ours	During the following year, she married Michael Jackson, whom she married in 1938, and they later divorced.
GT	The movie is a blast of educational energy, as bouncy animation and catchy songs...
GLIM	It's a slam dunk that's so bad it's so good it's so good...
Ours	Compared to the earlier, better-looking, and more interesting, Love This Who? is a fun...
GT	Warm Water Under a Red Bridge is a quirky and poignant Japanese film that explores...
GLIM	The movie is a richly imagined and admirably mature work from a gifted director.
Ours	Leaving the movie a little bit shaky, but it's still a great story, with a great cast and a great story.
GT	The quirky drama touches the heart and the funnybone thanks to the energetic...
GLIM	This movie is more romantic, emotional, and satisfying than the original.
Ours	Despite its high-quality production, the movie is a little too funny and overly sweet...
GT	Intriguing and downright intoxicating.
GLIM	The movie is a dull and uninteresting experience.
Ours	... ... a solid, engaging and engaging film.
GT	Bad and baffling from the get-go.
GLIM	The movie is too long and overly complicated.
Ours	Despite the lack of a traditional Italian accent, the movie lacks a solid story...
GT	His nephew is the actor Jack Davenport.
GLIM	He was an American jazz musician.
Ours	an Italian actor played Cyrano in the French movie The Plot.
GT	A gratingly unfunny groaner littered with zero-dimensional, unlikable characters...
GLIM	It's not a bad movie, but it's not a great one either.
Ours	Despite a grand ambition, the movie is a poor one, with a poor score and a poor acting.
GT	... a haunting vision, with images that seem more like disturbing hallucinations.
GLIM	This movie is terrible and boring.
Ours	Despite its young, unimpressive characters, this movie feels like a poorly made satire...
GT	... pitiful, slapdash disaster.
GLIM	The movie is genuinely moving.
Ours	The movie looks bad and behaves badly in this bad movie.
GT	Another Best of the Year selection.
GLIM	It's a film that's a lot of fun.
Ours	... a genuinely uplifting film.

L AI USE STATEMENT

We used generative AI tools to assist with translation and to edit the manuscript to improve readability.

The authors reviewed all AI-assisted text. We checked translations and language edits against the original text for accuracy in terminology, numerical values, and meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

The architecture, training stages, hyperparameters, and loss weights of SEMKEY are given in Section 3 and Appendix C. The dataset, preprocessing, and data split (by unique sentence, with 17,908 / 2,200 / 2,227 samples for training / validation / test) are described in Appendix C.3. All evaluation metrics, including the retrieval protocol, are defined in Section 3.3; the noise and time-shuffling controls are described in Section 4.2 and Appendix D; and results over 7 random seeds are reported in Appendix A. Code will be released upon acceptance.