INSTRUCT, REASON, DECODE: FINE-GRAINED SEMANTIC INSTRUCTION AND REASONING FOR EEG-TO-TEXT DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding open-vocabulary text from electroencephalography (EEG) remains challenging because of the low signal-to-noise ratio of EEG and the complexity of natural language. Although large language models can generate coherent text, limited understanding of EEG semantics can lead to excessive reliance on linguistic priors. During autoregressive generation, early prediction errors can also affect subsequent predictions. Existing approaches improve EEG–text alignment or introduce auxiliary semantic tasks, but their sentence-level learning objectives and coarse-grained prompts provide limited guidance about the specific content to express. To address these challenges, we propose SIRE (Semantic Instruction and Reasoning from EEG), an EEG-to-text framework that strengthens fine-grained semantic understanding and provides semantic guidance for generation. Task-specific question–answer instruction provides explicit learning targets for decoding evaluation aspects, judgments, factual categories, and semantic relations from EEG. Reasoning from EEG then adaptively selects semantic questions and integrates the predicted answers through chain-of-thought reasoning before sentence generation. These answers remain available throughout autoregressive decoding, allowing the model to use the predicted semantics alongside EEG representations and previously generated text. Experiments on ZuCo 1.0 and ZuCo 2.0 show that SIRE outperforms the compared baselines across multiple semantic decoding metrics under fully autoregressive generation. Ablations support the contributions of semantic instruction and reasoning, while input controls show that generation benefits from correctly paired EEG.

1 INTRODUCTION

Decoding natural language from electroencephalography (EEG) holds considerable promise. However, the low signal-to-noise ratio of EEG and the complexity of natural language make open-vocabulary decoding challenging. Some approaches retrieve keywords or semantic units from a candidate vocabulary set or a constructed semantic space, then use the decoded content for sentence reconstruction (Li et al., 2026; Yang et al., 2026). In parallel, EEG-conditioned generation methods use EEG representations as inputs to large language models, aiming to decode open-vocabulary text from EEG and generate coherent sentences (Wang & Ji, 2022; Duan et al., 2023; Liu et al., 2024; Wang et al., 2024).

Although large language models can generate coherent text conditioned on EEG signals, EEG-to-text decoding still faces many challenges. First, insufficient understanding of EEG semantics makes these models depend more heavily on their linguistic priors than on the EEG input. In fact, some models exhibit comparable decoding performance even when real EEG is replaced with noise (Jo et al., 2024; 2026). Second, large language models employ autoregressive decoding, in which early prediction errors can accumulate and affect subsequent predictions (Bengio et al., 2015). However, some previous work underestimated this problem by using teacher-forcing during evaluation, in which the model predicts each token using the preceding ground-truth tokens rather than its own predictions (Jo et al., 2026).

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.

To improve large language models' understanding of EEG semantics, GLIM uses multiple text variants to support semantic alignment between EEG and text representations (Liu et al., 2025). However, EEG–text alignment alone does not ensure that the language model learns fine-grained EEG semantics. On the other hand,SemKey introduces auxiliary tasks and uses the predicted semantic information as prompts for text generation(Wang et al., 2026). But hese semantic prompts also remain coarse-grained, providing limited guidance on the specific content to generate.

To address these challenges,we therefore propose SIRE (Semantic Instruction and Reasoning from EEG), a framework that enhances language models' understanding of EEG semantics and provides semantic guidance for text generation. Specifically, task-specific, fine-grained question–answer (QA) supervision helps the language model learn detailed semantic information from EEG (Wei et al., 2022a; Chung et al., 2024). Reasoning from EEG uses chain-of-thought (CoT) to integrate the decoded semantics and guide subsequent autoregressive text generation (Wei et al., 2022b; Zhang et al., 2024; Press et al., 2023). Experiments on ZuCo 1.0 and ZuCo 2.0 demonstrate that SIRE achieves state-of-the-art performance across multiple semantic decoding metrics under fully autoregressive generation.

Our contributions are fourfold:

- We propose SIRE, an EEG-to-text framework that enhances EEG semantic understanding and uses the decoded semantics to guide language decoding.
- We construct a task-specific, fine-grained QA set for semantic instruction tuning, enabling the language model to learn detailed semantic distinctions from EEG.
- We design Reasoning from EEG, using CoT to integrate fine-grained EEG semantics and guide subsequent autoregressive text generation.
- Experiments on the ZuCo 1.0 and ZuCo 2.0 datasets show that SIRE achieves state-of-the-art performance across multiple semantic decoding metrics, demonstrating the effectiveness of semantic instruction and reasoning for EEG-to-text decoding.

2 Related Work

2.1 EEG-to-Text Decoding

EEG-to-text decoding follows different modeling strategies. BrainMosaic and Brain-CLIPLM decode semantic units or keywords from constructed semantic spaces for sentence reconstruction (Li et al., 2026; Yang et al., 2026). Most studies condition large language models on EEG representations for open-vocabulary text generation, with representative methods such as EEG-to-Text (Wang & Ji, 2022) and DeWave (Duan et al., 2023). GLIM uses text data augmentation to strengthen EEG–text semantic alignment, helping the model learn semantic EEG representations (Liu et al., 2025). SemKey introduces auxiliary tasks and uses the predicted semantic information as decoding prompts to guide text generation (Wang et al., 2026).

Inspired by these approaches, we propose SIRE to strengthen the language model's understanding of fine-grained EEG semantics and provide semantic guidance for text generation. Task-specific QA instruction provides more detailed semantic learning targets, helping the language model better understand EEG semantics. Reasoning from EEG then integrates the decoded semantics through CoT before sentence generation to guide subsequent autoregressive decoding.

2.2 Instruction Tuning for EEG Understanding

Instruction tuning trains language models to follow natural-language task instructions (Wei et al., 2022a; Chung et al., 2024). LLaVA extends instruction tuning to visual understanding using image–text instruction data (Liu et al., 2023). Instruction-based learning has also been applied to EEG. NeuroLM uses multi-task instruction tuning for EEG analysis (Jiang et al., 2025), while LEAF uses task instructions to align EEG representations with textual label embeddings (Jiang et al., 2026). Thought2Text fine-tunes language models to generate image descriptions from EEG recorded while participants view images (Mishra et al., 2025).

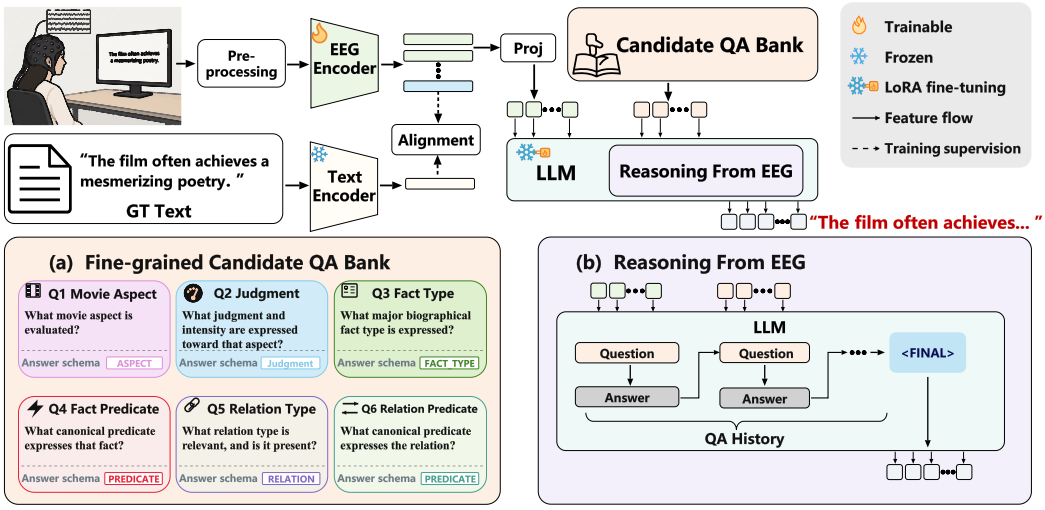

Figure 1: Overview of SIRE. The EEG encoder learns representations through EEG–text alignment, and a projector maps the sequence representations into the language-model embedding space. **(a)** The candidate QA bank defines six task-specific questions and their answer schemas. Annotated QA pairs provide fine-grained learning targets for semantic instruction tuning. **(b)** During Reasoning from EEG, the LLM adaptively selects and answers questions based on the EEG representations and preceding QA history. The resulting CoT integrates the decoded semantics and guides autoregressive sentence generation after the `<FINAL>` marker. The EEG encoder is trained during alignment and frozen during semantic instruction tuning.

These EEG applications mainly address signal analysis or image description. We apply instruction tuning to open-vocabulary sentence decoding from EEG recorded during reading. In SIRE, task-specific QA instruction teaches the language model to decode fine-grained EEG semantics and use this information to support subsequent reasoning and text generation.

3 Method

3.1 Framework Overview

Let $X \in \mathbb{R}^{T \times C}$ denote the input EEG signals, where T and C denote the numbers of time points and EEG channels, respectively. Let Y denote the corresponding sentence. As shown in Figure 1, after Pre-processing, the EEG Encoder produces a sequence representation Z and a global representation e. The sequence representation Z preserves EEG features across multiple positions for language decoding, while the global representation e is used for EEG–text contrastive alignment. The EEG representations are aligned with text representations extracted from the reference sentence (GT Text) by the frozen Text Encoder. The trainable projector (Proj) maps Z into the embedding space of the large language model (LLM) for subsequent reasoning and text generation.

The LLM receives the projected EEG representations together with the decoding instruction and the complete Candidate QA Bank. This bank contains six task-specific semantic questions and their answer schemas, as illustrated in Figure 1(a). During semantic instruction tuning, annotated QA pairs provide specific learning targets that help the LLM decode fine-grained semantic information from EEG.

During Reasoning from EEG (Figure 1(b)), the LLM adaptively selects and answers questions based on the EEG representations and the preceding QA History. Each predicted answer extends this history and informs subsequent question selection and answering. The resulting CoT integrates the decoded semantics before the LLM emits the `<FINAL>` marker and generates the final sentence autoregressively. The predicted answers in the QA History remain available throughout sentence generation, providing semantic guidance alongside the EEG representations.

Table 1: Task-specific questions used to construct the fine-grained QA set.

Reading context	Question	Semantic target
Sentiment reading	Q_1: What aspect of the movie is being evaluated?	Aspect
Sentiment reading	Q_2: What fine-grained judgment is expressed?	Judgment
Normal reading	Q_3: What type of factual information is expressed?	Fact type
Normal reading	Q_4: What action or predicate expresses the main fact?	Factual predicate
Task-specific reading	Q_5: What semantic relation is expressed in the sentence?	Relation type
Task-specific reading	Q_6: How is the relation expressed?	Relational predicate

3.2 EEG–Language Representation Alignment

We align EEG representations with text representations from the frozen Text Encoder to support subsequent semantic reasoning and text generation. For a batch of B EEG–text pairs, bidirectional contrastive learning brings matched global representations closer and separates unmatched pairs:

$$\mathcal{L}_{\mathrm{ctr}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ij})} + \log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ji})}\right]. \tag{1}$$

Here, $s_{ij} = \mathrm{sim}(e_i, t_j)/\kappa$, where e_i and t_j are the global EEG and text representations, sim denotes cosine similarity, and κ is a temperature parameter.

To also align the sequence features passed to the LLM, we use a representation regression loss $\mathcal{L}_{\mathrm{rep}}$. The sentence reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ encourages the EEG Encoder to retain information that the frozen LLM can use for text generation. The overall alignment objective is

$$\mathcal{L}_{\mathrm{align}} = \lambda_{\mathrm{ctr}}\mathcal{L}_{\mathrm{ctr}} + \lambda_{\mathrm{rep}}\mathcal{L}_{\mathrm{rep}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}}, \tag{2}$$

where the λ coefficients control the contribution of each loss.

3.3 Fine-Grained Semantic Instruction

To help the LLM learn fine-grained EEG semantics, we construct a task-specific QA set based on the reading contexts in ZuCo. The questions define semantic aspects relevant to sentence decoding, and their answers provide learning targets for the LLM used in Reasoning from EEG.

Task-specific question design. We define a shared bank of six questions, $\mathcal{Q} = \{Q_1, \ldots, Q_6\}$, covering evaluation, factual content, and semantic relations. For sentiment reading, the questions distinguish *what is being evaluated* from *the judgment expressed*. This separates, for example, a positive assessment of acting from a positive assessment of the plot. For normal reading, the questions identify *the type of factual information* and *the main predicate category*, such as education together with studying or graduating. For task-specific reading, the questions identify *the relation type* and *its associated predicate category*. Table 1 summarizes the question design.

Semantic answer spaces. For each question, we define a fixed set of answer labels and assign the same label to expressions with similar meanings. Evaluation aspects include categories such as acting, plot, and visual style, while judgments distinguish polarity and intensity. Factual answers describe categories such as education, employment, and achievements, together with their associated predicate categories. Relation answers follow the relation categories available in the dataset. The evaluation-aspect field permits multiple supported labels; the remaining fields use a single categorical answer. The complete label definitions are provided in Appendix A.2.

QA set construction. We construct the QA set using Qwen2.5-7B-Instruct for semantic annotation. For evaluation and factual questions, Qwen generates candidate answers from the original reference sentences according to the task-specific questions and predefined answer spaces. These candidates are checked against the sentence evidence and label definitions through candidate verification, automated consistency checks, and agent-assisted review of flagged cases. For relation questions, we use dataset-native relation labels. Their associated predicate categories are assigned

through a fixed mapping, so each predicate target is derived from its relation label. The annotation procedure uses training and validation material, with test sentences excluded. Annotation prompts, verification rules, and the full answer schema are documented in Appendix A.2.

From QA pairs to semantic instruction. Each annotated pair (Q_j, a^*_{ij}) supplies an intermediate learning target for the corresponding EEG observation X_i. During semantic instruction tuning, the LLM learns semantic answers together with the final sentence. Each training example uses the annotations associated with its reading context and receives the same complete question bank. The QA targets help the LLM learn how to answer semantic questions from EEG. During decoding, the LLM adaptively selects questions from this bank and uses the predicted answers for CoT reasoning and final text generation.

3.4 Reasoning from EEG

In Reasoning from EEG, the LLM adaptively selects semantic questions and predicts their answers based on the projected EEG representations and preceding QA history. The resulting question–answer CoT integrates the decoded semantics to guide subsequent autoregressive sentence generation.

EEG-conditioned context. A trainable projector Π maps the aligned EEG sequence representation Z into the language-model embedding space. The sequence representation preserves EEG features across multiple latent positions for subsequent decoding. We concatenate the projected sequence with the embeddings of a prompt P, which contains the decoding instruction and the complete question bank. The language-model encoder processes these inputs to produce a shared context for semantic reasoning and text generation:

$$H = \mathrm{Enc}_\theta\left([\mathrm{Emb}_\theta(P); \Pi(Z)]\right). \tag{3}$$

Here, θ denotes the LLM parameters, and Emb_θ and Enc_θ denote its token embedding layer and encoder, respectively. The encoder computes H once. The LLM attends to this shared context during both semantic prediction and final text generation, maintaining access to EEG information throughout decoding.

Adaptive CoT generation. We represent the CoT as $\tau = ((q_1, a_1), \ldots, (q_K, a_K))$, where q_k is the question selected at step k, a_k is its predicted answer, and K is the number of QA steps. At the start of reasoning and after each completed QA step, the LLM either selects an unused question from the shared question bank or terminates the CoT. If reasoning continues, the LLM selects q_k based on the EEG-conditioned context H and preceding QA history $\tau_{<k}$, then predicts its answer:

$$p_\theta(q_k, a_k \mid H, \tau_{<k}) = p_\theta(q_k \mid H, \tau_{<k})\, p_\theta(a_k \mid H, \tau_{<k}, q_k). \tag{4}$$

The selected question and its predicted answer are appended to the history, yielding $\tau_{\le k} = \tau_{<k} \oplus (q_k, a_k)$, where $\oplus$ denotes sequence concatenation. The LLM uses this updated history to select and answer subsequent questions.

Semantic guidance for text generation. Once the CoT terminates, the LLM generates the final sentence conditioned on both the EEG-conditioned context H and the predicted semantic history τ:

$$p_\theta(Y \mid H, \tau) = \prod_{n=1}^{N} p_\theta(y_n \mid H, \tau, y_{<n}), \tag{5}$$

where y_n denotes the n-th token of the sentence Y, and N is its token length. The semantic answers predicted in the CoT remain part of the decoder's prefix throughout sentence generation. They specify predicted aspects of sentence content, such as the evaluation aspect, factual category, or relation type. At each decoding step, the LLM can use these predictions together with the EEG representations and previously generated words. Even if an early sentence token is inaccurate, the semantic predictions remain available as additional context for later predictions. This design aims to reduce semantic drift caused by early generation errors.

Joint semantic instruction tuning. We jointly optimize semantic prediction and sentence generation in the same LLM. For each training example, we construct a QA sequence τ_i^* from its valid annotations and append the reference sentence Y_i to form the target sequence R_i^*. During training, the LLM predicts each target token from the shared context H_i and the preceding reference tokens.

Within each minibatch, let Ω_{sem} contain the token positions for question selection, semantic answers, and CoT control. Let Ω_{text} contain the positions for the final-sentence segment and sequence termination. For either set Ω, the loss is

$$\mathcal{L}_\Omega = -\frac{1}{\max(1, |\Omega|)} \sum_{(i,u)\in\Omega} \log p_\theta \left(r_{i,u}^* \mid H_i, r_{i,<u}^* \right), \tag{6}$$

where $r_{i,u}^*$ is the u-th token of R_i^*, and $r_{i,<u}^*$ denotes its preceding reference tokens. We average each loss over its corresponding target tokens before applying the weights in the joint objective:

$$\mathcal{L}_{\text{SIRE}} = \gamma \left[\mathcal{L}_{\Omega_{\text{text}}} + \lambda_{\text{sem}} \mathcal{L}_{\Omega_{\text{sem}}} \right]. \tag{7}$$

Here, λ_{sem} controls the weight of semantic prediction relative to sentence generation, and γ scales the overall loss. During optimization, we update the EEG projector, LoRA adapters, and structural-token parameters. The EEG representation network and the LLM's base parameters remain frozen (Hu et al., 2022).

Autoregressive inference. During the inference, the LLM integrates the decoded semantics through Reasoning from EEG to guide subsequent sentence generation. The QA pairs and final sentence are decoded as a single autoregressive sequence. Each prediction is conditioned on H and previously generated tokens. Decoding constraints ensure a valid QA format and prevent repeated questions. No ground-truth semantic answers or reference-text prefixes are provided, and only the final sentence is evaluated.

4 Experiments

4.1 Experimental Setup

Datasets and preprocessing. We conduct experiments on ZuCo 1.0 and ZuCo 2.0 (Hollenstein et al., 2018; 2020). The data cover three reading tasks: sentiment reading, normal reading, and task-specific reading. Each sample pairs a sentence-level EEG recording with its corresponding English sentence. Before encoding, we discard EEG segments containing non-finite values or lasting less than 0.5 s or more than 10 s. We remove the empty channel in the provided recordings, retaining 104 EEG channels, and resample the signals from 500 Hz to 128 Hz. Each segment is then zero-padded to 1,280 temporal samples, corresponding to a duration of 10 s. The channel dimension is also zero-padded to 128 positions to obtain a fixed input shape. A binary temporal validity mask distinguishes recorded samples from padded positions. The data are partitioned into training, validation, and test sets with a ratio of 70%/10%/20% at the sentence level, with all recordings associated with the same sentence assigned to the same partition. Dataset statistics and further preprocessing details are provided in the appendix.

Baselines. We compare SIRE with EEG-to-Text (Wang & Ji, 2022), DeWave (Duan et al., 2023), GLIM (Liu et al., 2025), and SemKey (Wang et al., 2026), covering different EEG representation learning and semantic guidance strategies. We also include Brain-CLIPLM (Yang et al., 2026), which uses EEG-derived semantic anchors and retrieved corpus sentences for generation. SIRE without LoRA is evaluated to assess the contribution of language-model adaptation. All methods are evaluated without ground-truth text prefixes.

Evaluation metrics. We evaluate the final decoded sentences using two groups of metrics. For *semantic alignment and content recovery*, **N-way semantic matching accuracy** ($N \in \{2, 4, 10, 24\}$) measures how often a generated sentence is correctly matched to its reference among N candidates. **Content-word recall** measures the proportion of reference content words present in the generated text. **Fréchet distance (FD)** measures the difference between generated and reference sentence distributions in semantic space, with lower values indicating closer distributions.

For *lexical overlap and generation diversity*, **BLEU-1–4** (Papineni et al., 2002) measure word and phrase overlap with reference sentences. **Distinct-1/2** (Li et al., 2016) measure the proportions of unique unigrams and bigrams in the generated corpus, with higher values indicating greater lexical diversity. **Self-BLEU** (Zhu et al., 2018) measures overlap among generated sentences, with lower values indicating less repetition. **Head Entropy** measures the entropy of sentence-opening token sequences, with higher values indicating more varied sentence openings. Formal definitions and evaluation settings are provided in Appendix B.2.

Implementation and decoding settings are provided in Appendix B.3.

4.2 Main Results

Semantic alignment and content recovery. Table 2 shows that SIRE achieves the best results across all four N-way settings, content-word recall, and Fréchet distance (FD). Compared with SemKey, SIRE increases 24-way accuracy from 11.02% to 14.85% and nearly doubles content-word recall from 0.0168 to 0.0326. These improvements indicate that the generated sentences better distinguish the target meaning from competing references and recover more informative reference words. In SIRE, fine-grained QA instruction provides explicit learning targets for evaluation aspects, judgments, factual categories, and semantic relations. These targets teach the language model more specific distinctions about sentence content than broad semantic attributes alone provide.

Reasoning from EEG makes the predicted semantics available before sentence generation. Subsequent decoding steps can therefore refer to this semantic context alongside EEG representations and the generated text prefix, with the aim of limiting semantic drift caused by early prediction errors. The improvements in semantic matching and content recovery show that the complete framework better preserves target content, while the lowest FD indicates closer generated and reference distributions in semantic space. Section 4.3 examines the contributions of semantic instruction and CoT guidance separately.

Lexical overlap and generation diversity. Table 3 evaluates lexical agreement and diversity for the same generated sentences. SIRE obtains the highest BLEU-1 and BLEU-4 scores, while GLIM leads on BLEU-2 and BLEU-3. SIRE's BLEU-4 increases from SemKey's 1.69 to 2.18, indicating better agreement with reference phrases. SIRE also achieves the highest Distinct-1 and Head Entropy, together with lower Self-BLEU than SemKey, indicating more varied vocabulary and sentence openings with less repetition. Fine-grained QA instruction teaches the language model to recover content-relevant distinctions from EEG, while Reasoning from EEG integrates the predicted semantics to guide subsequent generation. The gains in both reference agreement and diversity suggest that this semantic guidance supports content recovery while reducing repetitive outputs.

Brain-CLIPLM achieves the highest Distinct-2 and the lowest Self-BLEU. Its retrieval-augmented generation supplies corpus sentences as additional context (Yang et al., 2026), which may contribute to more varied wording. Without retrieving corpus sentences at inference, SIRE achieves higher semantic matching, content-word recall, and BLEU scores than Brain-CLIPLM. These results distinguish output diversity from target-content recovery: SIRE maintains competitive diversity while better preserving the meaning associated with the EEG input.

4.3 Ablation Studies

Table 4 examines QA granularity, CoT guidance, and LoRA adaptation. All variants share the same aligned EEG encoder, language backbone, data split, and evaluation protocol.

Fine-grained QA instruction improves 24-way accuracy from 11.07% to 14.11% even without a CoT prefix during generation, showing the benefit of detailed semantic learning targets. Adding CoT further improves all six metrics, supporting the value of using the decoded semantics to guide text generation. With CoT and LoRA enabled, fine-grained QA also outperforms coarse-grained QA across all six metrics, highlighting the importance of semantic granularity. LoRA adaptation brings further improvements, while SIRE without LoRA still outperforms SemKey on these metrics, showing that the gains do not depend solely on language-model adaptation.

A comparison of fixed-order and adaptive QA generation is provided in Appendix B.6.

Table 2: Semantic alignment and content recovery on ZuCo under autonomous generation. N-way accuracies are reported as percentages. C. Recall denotes content-word recall and is reported as a ratio. Fréchet distance is reported without percentage scaling. Arrows indicate preferred directions; bold denotes the best result in each column.

	Semantic matching (%)				Content recovery	Distribution
Method	2-way ↑	4-way ↑	10-way ↑	24-way ↑	C. Recall ↑	Fréchet distance ↓
EEG-to-Text	49.96	24.97	10.06	4.20	0.0181	1.53
DeWave	50.04	25.04	10.03	4.18	0.0178	1.64
GLIM	66.29	42.08	21.43	10.21	0.0133	0.6945
SemKey	68.98	43.51	22.61	11.02	0.0168	0.7148
Brain-CLIPLM	57.81	33.35	15.50	7.18	0.0149	0.8628
SIRE w/o LoRA	69.70	45.60	25.03	12.52	0.0248	0.4807
SIRE	**73.38**	**50.50**	**28.36**	**14.85**	**0.0326**	**0.3951**

Table 3: Lexical overlap and generation diversity for the same decoded outputs evaluated in Table 2. BLEU and Self-BLEU are reported as percentages. Dist-n denotes Distinct-n and is reported as a ratio. H. Ent. denotes sentence-prefix entropy in bits. Arrows indicate preferred directions; bold denotes the best result in each column.

	Lexical overlap (%)				Generation diversity			
Method	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	Dist-1 ↑	Dist-2 ↑	Self-BLEU (%) ↓	H. Ent. ↑
EEG-to-Text	20.28	6.18	1.62	0.51	0.0274	0.0547	98.83	0.09
DeWave	22.25	7.29	1.62	0.35	0.0373	0.0618	99.40	0.00
GLIM	26.04	**10.56**	**3.97**	1.97	0.0356	0.0958	91.35	2.16
SemKey	23.14	7.39	2.01	1.69	0.0176	0.0542	86.33	2.27
Brain-CLIPLM	8.66	1.97	0.35	0.11	0.0348	**0.1404**	**69.37**	3.49
SIRE w/o LoRA	23.91	8.75	2.81	1.20	0.0327	0.1056	79.02	3.31
SIRE	**26.85**	10.16	3.96	**2.18**	**0.0411**	0.1368	75.16	**4.00**

Table 4: Component ablations of SIRE on ZuCo 1.0 and ZuCo 2.0 under autoregressive generation. None denotes no QA instruction. Coarse QA uses sentiment, topic, length, and surprisal; Fine QA uses our task-specific semantic targets. ✓ and ✗ indicate enabled and disabled components, respectively. The last row is full SIRE. N-way accuracies are percentages, C. Recall is a ratio, and Fréchet distance is unscaled. Bold denotes the best result in each metric.

Configuration			N-way accuracy (%)					
QA instruction	CoT	LoRA	2-way ↑	4-way ↑	10-way ↑	24-way ↑	C. Recall ↑	Fréchet distance ↓
None	✗	✓	68.79	44.35	23.06	11.07	0.0233	0.5924
Fine	✗	✓	71.06	48.35	27.09	14.11	0.0273	0.4528
Coarse	✓	✓	71.06	46.93	25.80	12.71	0.0239	0.4367
Fine	✓	✗	69.70	45.60	25.03	12.52	0.0248	0.4807
Fine	✓	✓	**73.38**	**50.50**	**28.36**	**14.85**	**0.0326**	**0.3951**

4.4 EEG Dependency Analysis

Input controls. We evaluate three input conditions to test whether SIRE uses EEG information associated with the target sentence (Jo et al., 2024; Liu et al., 2025). **Real EEG** uses correctly paired recordings. **Shuffled EEG** exchanges complete recordings across different sentence identities, with each recording's validity mask moving with it. **Gaussian noise** replaces valid EEG samples with independent Gaussian noise drawn from distributions matching the recording's channel-wise means and standard deviations. Valid lengths, masks, and padding remain unchanged in the noise condition. We keep the checkpoint, prompts, reference sentences, candidate pools, and decoding settings fixed. EEG representations and the CoT are regenerated for each condition without additional training. Full details are provided in Appendix B.4.

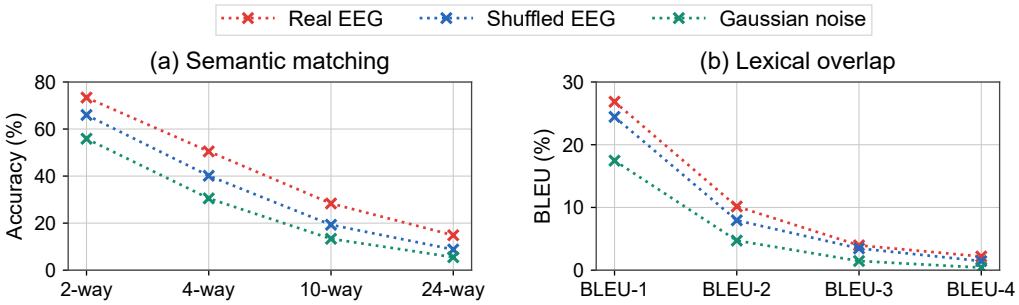

Figure 2: EEG dependency of SIRE under fully autoregressive generation. **(a)** N-way semantic matching accuracy. **(b)** BLEU scores. All scores are reported as percentages; higher is better. Real-EEG results are reproduced from Tables 2 and 3.

Dependence on sentence-specific EEG information. Figure 2 shows that correctly paired EEG yields the best semantic matching and lexical overlap. Shuffling preserves real EEG recordings but breaks their correspondence with the target sentences, lowering scores across both metric groups. This decline indicates that SIRE benefits from correctly paired EEG during generation. Gaussian noise produces the lowest scores, but noise replacement also changes the input distribution. Shuffling therefore provides the more direct test of sentence-specific EEG dependence. These findings support SIRE's use of EEG information during text generation, while the ablation results in Table 4 demonstrate the contributions of semantic instruction and CoT guidance.

5 Discussion

Limitations. Although SIRE improves semantic decoding on ZuCo 1.0 and ZuCo 2.0, our evaluation remains limited to the ZuCo dataset. The semantic questions and answer categories are also designed around its reading tasks. Further evaluation is therefore needed on datasets with different recording conditions and task requirements. In particular, the current experiments do not establish whether the same QA design can provide effective semantic guidance for sentence content beyond the evaluated reading contexts.

Future work. Future work will evaluate SIRE on additional EEG–text datasets to examine its generalizability. For datasets with different task requirements, we will adapt the questions and answer categories to the semantic information relevant to sentence decoding. We will also evaluate the full framework alongside its ablated variants and repeat the EEG input controls. These experiments will assess whether the gains from semantic instruction and CoT guidance persist across datasets and whether generation continues to benefit from correctly paired EEG. This will help distinguish the transferable benefits of the framework from those associated with its task-specific QA design.

Conclusion. We proposed SIRE for EEG-to-text decoding. Task-specific QA instruction provides fine-grained learning targets that help the LLM understand EEG semantics. Reasoning from EEG integrates the decoded semantics to guide subsequent autoregressive sentence generation. Experiments on ZuCo 1.0 and ZuCo 2.0 demonstrate improvements in semantic decoding under fully autoregressive generation. Ablations support the contributions of semantic instruction and CoT guidance, while input controls indicate that correctly paired EEG improves generation. These findings support learning fine-grained EEG semantics and using them as guidance for language decoding.

AI USE STATEMENT

We used generative AI tools to assist with language editing, including grammar correction, wording refinement, and improvements to clarity and readability. The authors take responsibility for the final manuscript.

ETHICS STATEMENT

This study uses existing EEG recordings from ZuCo 1.0 and ZuCo 2.0 (Hollenstein et al., 2018; 2020), without collecting new recordings or recruiting participants. EEG-to-text decoding raises concerns about neural-data privacy and the interpretation of generated text. Our evaluation concerns sentences presented during reading; it does not establish the ability to infer unrestricted mental content. Decoded sentences may contain errors and should not be treated as reliable evidence of an individual's thoughts or intentions. Future applications should require informed consent, protect access to neural data, and undergo validation for their intended population and setting.

REPRODUCIBILITY STATEMENT

Section 3 describes the SIRE framework and its learning objectives. Appendix B.1 documents the dataset partitions and EEG preprocessing, while Appendix A.2 provides the QA schemas, annotation prompts, and verification rules. Training configurations, random seeds, checkpoint selection, and autoregressive decoding settings are detailed in Appendix B.3. Metric definitions and evaluation procedures are provided in Appendix B.2, and the shuffled-EEG and Gaussian-noise protocols are specified in Appendix B.4. Our code is at `https://anonymous.4open.science/r/SIRE-297B/README.md`

REFERENCES

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL `https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html`.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. URL `https://www.jmlr.org/papers/v25/23-0870.html`.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. DeWave: Discrete encoding of EEG waves for EEG to text translation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9907–9918. Curran Associates, Inc., 2023. doi: 10.52202/075280-0432. URL `https://proceedings.neurips.cc/paper_files/paper/2023/hash/1f2fd23309a5b2d2537d063b29ec1b52-Abstract-Conference.html`.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5:180291, 2018. doi: 10.1038/sdata.2018.291. URL `https://www.nature.com/articles/sdata2018291`.

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, Marseille, France, May 2020. European Language Resources Association. URL `https://aclanthology.org/2020.lrec-1.18/`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Con-*

ference on Learning Representations, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Io9yFt7XH7.

Muyun Jiang, Shuailei Zhang, Zhenjie Yang, Mengjun Wu, Weibang Jiang, Zhiwei Guo, Wei Zhang, Rui Liu, Shangen Zhang, Yong Li, Yi Ding, and Cuntai Guan. LEAF: Language-EEG aligned foundation model for brain-computer interfaces. *arXiv preprint arXiv:2509.24302v2*, 2026. URL https://arxiv.org/abs/2509.24302v2.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are EEG-to-text models working?, 2024. URL https://arxiv.org/abs/2405.06459.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Evaluating EEG-to-text models through noise-based performance analysis. *Scientific Reports*, 16(1):350, 2026. doi: 10.1038/s41598-025-29587-x. URL https://www.nature.com/articles/s41598-025-29587-x.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119. Association for Computational Linguistics, 2016. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Jiahe Li, Junru Chen, Fanqi Shen, Jialan Yang, Jada Li, Zhizhang Yuan, Baowen Cheng, Meng Li, and Yang Yang. Assembling the mind's mosaic: Towards EEG semantic intent decoding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8OgJ2uhiu8.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

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. doi: 10.1109/BIGDATA62323.2024.10825980. URL https://doi.org/10.1109/BIGDATA62323.2024.10825980.

Xiaozhao Liu, Dinggang Shen, and Xihui Liu. Learning interpretable representations leads to semantically faithful EEG-to-text generation, 2025. URL https://arxiv.org/abs/2505.17099.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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, Inga Zhuravleva, Adelyn Tu-Chan, Karunesh Ganguly, Gopala K. Anumanchipalli, and Edward F. Chang. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620:1037–1046, 2023. doi: 10.1038/s41586-023-06443-4. URL https://www.nature.com/articles/s41586-023-06443-4.

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, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.207. URL https://aclanthology.org/2025.findings-naacl.207/.

Akshaj Murhekar, Christina Liu, Abhijit Mishra, Shounak Roychowdhury, and Jacek Gwizdka. SENSE: Efficient EEG-to-text via privacy-preserving semantic retrieval. *arXiv preprint arXiv:2603.17109*, 2026. URL https://arxiv.org/abs/2603.17109.

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. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/v21/20-074.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. Enhancing EEG-to-text decoding through transferable representations from pre-trained contrastive EEG-text masked autoencoder. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7278–7292. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.393. URL https://aclanthology.org/2024.acl-long.393/.

Yuchen Wang, Haonan Wang, Yu Guo, Honglong Yang, and Xiaomeng Li. Escaping the BLEU trap: A signal-grounded framework with decoupled semantic guidance for EEG-to-text decoding, 2026. URL https://arxiv.org/abs/2603.03312.

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

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=gEZrGCozdqR.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022b. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. *Nature*, 620: 1031–1036, 2023. doi: 10.1038/s41586-023-06377-x. URL https://www.nature.com/articles/s41586-023-06377-x.

Xiaoli Yang, Huiyuan Tian, Yurui Li, Jianyu Zhang, Shijian Li, and Gang Pan. Brain-CLIPLM: Semantic compression for EEG-to-text decoding. *Frontiers in Neuroscience*, 20:1899770, July 2026. doi: 10.3389/fnins.2026.1899770. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2026.1899770/full.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=y1pPWFVfvR.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1097–1100. Association for Computing Machinery, 2018. doi: 10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.

A METHOD IMPLEMENTATION DETAILS

A.1 EEG REPRESENTATION ALIGNMENT

A conditional EEG Transformer produces temporal hidden states, and a query-based merger uses $L = 96$ learnable queries to obtain a fixed-length latent sequence. Let $V_i \in \mathbb{R}^{96\times128}$ denote the merged EEG states. A linear alignment layer projects these states to the Flan-T5 representation width $d = 1024$:

$$Z_i = V_i W_a + b_a, \qquad W_a \in \mathbb{R}^{128\times1024}, \qquad Z_i \in \mathbb{R}^{96\times1024}. \tag{8}$$

A learnable global query aggregates Z_i through multi-head attention:

$$e_i = \mathrm{MHA}(q_g, Z_i, Z_i), \qquad e_i \in \mathbb{R}^{1024}. \tag{9}$$

The sequence representation Z_i retains EEG features across multiple latent positions, while the global representation e_i is used for EEG–text contrastive alignment. The EEG projector receives the unnormalized sequence representation returned by the aligner. Normalization is applied only within the alignment objectives.

EEG–text alignment uses a frozen Flan-T5 model for linguistic supervision. Let $t_i \in \mathbb{R}^d$ denote the global text representation and let $Z_i^{\mathrm{text}} \in \mathbb{R}^{L\times d}$ denote the dimension-matched sequence-level linguistic representation. The representation network is optimized with three objectives. First, the bidirectional batch-wise contrastive objective in Eq. 1 aligns the global EEG and text representations. Second, sequence-level correspondence is encouraged by

$$\mathcal{L}_{\mathrm{rep}} = \mathrm{MSE}\left(\mathrm{Norm}(Z_i), \mathrm{Norm}(Z_i^{\mathrm{text}})\right). \tag{10}$$

Third, the frozen T5 decoder provides EEG-conditioned language reconstruction supervision through $\mathcal{L}_{\mathrm{rec}}$. The alignment objective used in our implementation is

$$\mathcal{L}_{\mathrm{align}} = 0.5\mathcal{L}_{\mathrm{ctr}} + 0.7\mathcal{L}_{\mathrm{rep}} + 0.5\mathcal{L}_{\mathrm{rec}}. \tag{11}$$

The sentiment, topic, length, and surprisal auxiliary heads are not instantiated in this route. The resulting EEG representation network is frozen and kept in evaluation mode throughout semantic instruction tuning.

A.2 QA ANSWER SPACES AND ANNOTATION

The six questions in Table 1 share a fixed answer schema. Q_1 permits one to three supported aspect labels; the other questions use one categorical answer. Labels normalize related expressions rather than recovering arbitrary words or complete event structures. Tables 5–7 give the label definitions. The relation and relational-predicate columns in Table 7 also specify the deterministic mapping from Q_5 to Q_6.

Table 5: Answer labels for evaluation aspects and judgments.

Label	Definition
Q_1**: Aspect**	
`OVERALL_QUALITY`	Overall evaluation of the movie when no narrower supported aspect dominates. Do not add merely because another aspect is selected.
`STORY_PLOT`	Narrative premise, story events, plot logic, twists, or story coherence.
`ACTING_PERFORMANCE`	Quality of an actor, cast, performance, delivery, or portrayal as acting skill.
`CHARACTERIZATION`	Character construction, development, motivation, depth, or relationships; not the actor performance itself.
`DIRECTION_FILMMAKING`	Direction, staging, filmmaking choices, control of tone, or overall directorial craft.
`WRITING_DIALOGUE`	Screenplay, dialogue, narration, verbal humor, prose, or quality of the written lines.

Table 5 (continued)

Label	Definition
VISUAL_STYLE	Cinematography, imagery, animation, production design, editing as visual craft, or visual effects.
MUSIC_SOUND	Score, songs, soundtrack, audio, voice sound, or sound design.
PACING_STRUCTURE	Tempo, duration, repetition, sequencing, narrative organization, or structural flow.
THEMATIC_EMOTIONAL_IMPACT	Themes, ideas, message, originality, emotional force, resonance, or audience impact.
Q_2: Fine judgment	
VERY_NEGATIVE	Explicitly extreme condemnation or exceptionally poor evaluation, supported by strong intensifiers or superlatives.
NEGATIVE	Clearly unfavorable evaluation without evidence of extreme intensity.
MIXED_OR_NEUTRAL	Balanced praise and criticism, ambivalence, or no clearly supported evaluative polarity.
POSITIVE	Clearly favorable evaluation without evidence of exceptional intensity.
VERY_POSITIVE	Explicitly exceptional praise or outstanding evaluation, supported by strong intensifiers or superlatives.

Table 6: Answer labels for factual content and predicates.

Label	Definition
Q_3: Fact type	
IDENTITY_ORIGIN	Identity, classification, nationality, provenance, name, or personal origin.
EDUCATION	Schooling, degree, study, academic training, or graduation.
EMPLOYMENT	Employment, employer, workplace, professional engagement, or being hired.
ROLE_POSITION	Occupation, job title, office, rank, membership role, or formal position.
ACHIEVEMENT_AWARD	Award, honor, decoration, victory, recognized accomplishment, or notable achievement.
FOUNDING_CREATION	Founding an organization, creating or inventing something, or initiating an institution or work.
LOCATION_MOVEMENT	Residence, location, travel, visit, migration, arrival, departure, or geographic movement.
FAMILY_RELATIONSHIP	Marriage, spouse, parentage, kinship, ancestry, or other family relationship.
POLITICAL_AFFILIATION	Political party, ideology, campaign, faction, or explicit political affiliation or support.
LIFE_EVENT_OR_OTHER	Birth, death, age, date, or another supported factual frame that fits none of the defined classes.
Q_4: Factual predicate	
BE_IDENTIFY	Copular or identifying relation such as be, become, be known as, originate, or be from.
BORN_DIE	Birth, death, lifespan, or age-at-life-event predicate.
WORK_EMPLOY	Work for, employ, hire, practice a profession, or leave employment.
STUDY_GRADUATE	Study, attend, train, earn a degree, or graduate.
SERVE_HOLD_ROLE	Serve as, hold office, act as, be appointed, or occupy a formal role.
FOUND_CREATE	Found, establish, create, invent, form, or launch.
WIN_RECEIVE_AWARD	Win, receive, earn, be awarded, honor, or decorate.
WRITE_PUBLISH_PERFORM	Write, publish, compose, direct a work, appear, act, or perform.
MOVE_LIVE_VISIT	Move, live, reside, travel, visit, arrive, or leave a place.
MARRY_FAMILY_RELATE	Marry, divorce, be a spouse or relative, descend from, or have a family relation.

Table 6 (continued)

Label	Definition
JOIN_SUPPORT_AFFILIATE	Join, belong to, affiliate with, support, advocate, or participate in an organization or cause.
OTHER	A supported main predicate that cannot be mapped to any defined predicate class; do not use when evidence is insufficient.

Table 7: Relation labels, associated predicates, and their meanings. The middle column is deterministically derived from the first.

Q_5: Relation	Q_6: Predicate	Meaning
AWARD	WIN_RECEIVE	Award, honor, medal, prize, or decoration relation.
EDUCATION	STUDY_GRADUATE	Education, study, institution, degree, or graduation relation.
EMPLOYER	WORK_EMPLOY	Employer or employment-organization relation.
FOUNDER	FOUND_ESTABLISH	Founder or organization-establishment relation.
JOB_TITLE	SERVE_HOLD_ROLE	Occupation, office, rank, role, or job-title relation.
NATIONALITY	BE_FROM_NATIONALITY	Nationality, citizenship, or explicit national-origin relation.
POLITICAL_AFFILIATION	JOIN_SUPPORT_AFFILIATE	Political party, faction, ideology, or affiliation relation.
VISITED	MOVE_VISIT_LIVE	Visit, residence, travel, or location-movement relation.
WIFE	MARRY_SPOUSE	Spouse or marriage relation, regardless of surface gender wording.
NO_RELATION	OTHER_NONE	Control sentence or sentence without a supported target relation from the dataset-native ontology.

Annotation prompts. Qwen2.5-7B-Instruct receives the original sentence, its reading context, and the corresponding frozen answer schema. Candidate annotation and verification use separate requests. The candidate-generation system instruction is:

> You are annotating ORIGINAL ZuCo sentences for EEG semantic supervision. Do not rewrite the sentence or infer unsupported facts. Choose labels only from the supplied ontology. If evidence is insufficient, set valid=false. Return strict JSON only.

The verification system instruction is:

> You are a strict verifier, not a second generator and not a voter. Treat the candidate annotation as an untrusted claim. Determine from the ORIGINAL sentence and frozen ontology whether each candidate slot is supported. You may confirm, correct, reject, or mark it uncertain; there is no requirement to preserve the candidate. Return strict JSON only.

Each request includes the slot descriptions, label definitions, selection rules, and original sentence. Verification additionally receives the candidate annotation. Responses use a JSON structure with a label value, a validity flag, confidence, and supporting text evidence for each slot.

Verification rules. The pipeline checks schema membership, field validity, agreement between annotation and verification, and whether cited evidence occurs in the original sentence. Disagreement, uncertain decisions, low confidence, or unsupported evidence triggers further review. Dataset-native labels provide the relation targets, and the associated predicates follow the mapping in Table 7. Agent-assisted review resolves flagged cases against the sentence and the fixed definitions; it is not an independent human annotation study. The annotation schema and semantic targets are constructed from training and validation material, excluding test sentences.

Supervised trajectories. Sentiment-reading targets contain the valid Q_1 and Q_2 annotations, normal-reading targets contain Q_3 and Q_4, and task-specific-reading targets contain Q_5 and Q_6. The supervised order follows these pairs, and invalid or empty annotations are omitted. Every input receives the same complete question bank. During inference, the decoder selects individual questions from that bank rather than receiving a prescribed pair.

A.3 EEG Conditioning and Sequence Structure

The EEG projector maps only the sequence representation Z_i into the LLM embedding space. The global representation e_i is used for contrastive alignment and is not appended to the language-model input. Let L_{P_i} denote the number of tokens in the decoding prompt P_i. The projected EEG sequence and encoder input are

$$\widetilde{Z}_i = \Pi(Z_i) \in \mathbb{R}^{L\times d}, \qquad U_i = \left[\mathrm{Emb}_\theta(P_i); \widetilde{Z}_i\right] \in \mathbb{R}^{(L_{P_i}+L)\times d}. \tag{12}$$

The shared encoder context is

$$H_i = \mathrm{Enc}_\theta(U_i). \tag{13}$$

The prompt contains the decoding instruction and the complete six-question bank. Ground-truth semantic answers and the reference sentence are not included. The encoder is evaluated once per example; generated QA pairs remain in the causal decoder prefix rather than being re-encoded after every step.

The structured output follows the logical form

$$\mathrm{COT} \to (q_1, a_1) \to \cdots \to (q_K, a_K) \to \mathrm{FINAL} \to Y, \tag{14}$$

where K is the number of inquiry steps generated for the example. In the implementation, this structure is serialized with explicit control and field markers. A representative target is

```
<COT>
<SELECT><Q1></SELECT><ASPECT>ACTING_PERFORMANCE</ASPECT>
<SELECT><Q2></SELECT><FINE_JUDGMENT>POSITIVE</FINE_JUDGMENT>
</COT><FINAL>The performances are engaging.</FINAL>
```

The closing `</COT>` marker acts as the stop action for semantic inquiry; no additional `<STOP>` token is introduced. The implementation registers 24 indivisible structural tokens, comprising COT/SELECT/FINAL markers, six question identifiers, and six answer-field marker pairs.

A.4 Trainable Parameters and Token Losses

The EEG representation network and pretrained FLAN-T5 backbone remain frozen during semantic instruction tuning. The trainable parameters are the EEG projector, LoRA adapters on the T5 attention projections, and small input/output updates for the newly introduced structural tokens. For an adapted projection, LoRA (Hu et al., 2022) takes the form

$$W = W_0 + \frac{\alpha}{r} BA, \qquad A \in \mathbb{R}^{r\times d_{\mathrm{in}}}, \quad B \in \mathbb{R}^{d_{\mathrm{out}}\times r}, \tag{15}$$

where W_0 is frozen. We use rank $r = 8$, scale $\alpha = 16$, and dropout 0.1 on the query and value projections.

The serialized target is partitioned into three mutually exclusive token masks. M_{sel} contains inquiry-control and question-selection tokens, M_{ans} contains answer-field and answer-content tokens, and M_{final} contains the final-sentence segment and sequence termination. We define

$$M_{\mathrm{sem}} = M_{\mathrm{sel}} \cup M_{\mathrm{ans}}. \tag{16}$$

For a minibatch $\mathcal{B}$, the token-normalized loss is

$$\mathcal{L}(M) = -\frac{\sum_{i\in\mathcal{B}} \sum_u M_{iu} \log p_\theta(r^*_{iu} \mid H_i, r^*_{i,<u})}{\max\left(1, \sum_{i\in\mathcal{B}} \sum_u M_{iu}\right)}. \tag{17}$$

The semantic instruction-tuning objective used in the experiments is

$$\mathcal{L}_{\mathrm{SIRE}} = 1.5\left[\mathcal{L}(M_{\mathrm{final}}) + 0.3\mathcal{L}(M_{\mathrm{sem}})\right]. \tag{18}$$

An empty mask contributes zero loss. Separately normalizing the semantic and final segments prevents their different token counts from implicitly fixing their relative contribution.

The structural-token input and output updates are maintained as small FP32 parameter tables while the pretrained vocabulary parameters remain frozen. This allows the newly registered control tokens to acquire task-specific representations without unfreezing the full embedding or language-model head.

A.5 Autonomous Decoding

At inference, the model receives the same complete six-question bank for every example. Generation begins with the COT segment. At each valid inquiry boundary, the model may either continue by emitting a new selection or stop by closing the COT segment. If it continues, only previously unused question identifiers are legal. After a question is selected, its associated answer field is generated before the decoder can return to the next inquiry boundary. The final sentence is generated only after the COT segment is closed.

We use grammar-constrained greedy decoding to enforce structural validity, question–field consistency, and non-repetition of questions. The constraints do not supply semantic answers or choose among multiple legal questions on the model's behalf. The complete generated sequence shares a maximum budget of 192 new tokens. Ground-truth semantic answers and reference-text prefixes are absent from inference inputs. Only the parsed FINAL segment is used for language-decoding metrics; the generated inquiry trajectory is retained for analysis.

B Experimental Details

B.1 Dataset Partition and EEG Preprocessing

We combine the sentence-level EEG–text pairs from ZuCo 1.0 and ZuCo 2.0 (Hollenstein et al., 2018; 2020). Sentence identities are assigned to training, validation, and test partitions with proportions of 70%, 10%, and 20%, respectively, using a split seed of 42. Repeated recordings of a sentence share its partition. The resulting observation counts are reported in Table 8; they count EEG–text pairs rather than unique sentences.

Table 8: Numbers of EEG–text pairs after preprocessing.

	Training	Validation	Test
EEG–text pairs	10,999	1,672	3,206

Sentence segments with non-finite values or durations outside 0.5–10 seconds are excluded. After removing the empty channel in the released sentence-level arrays, 104 recorded channels remain. These signals are resampled from 500 Hz to 128 Hz. Each segment is zero-padded to 1,280 temporal samples and 128 channel positions. In the notation of Section 3.1, the preprocessed tensor is $\widetilde{X} \in \mathbb{R}^{1,280\times 128}$, with temporal samples along the first dimension and channels along the second. X denotes the original EEG input before this preprocessing. The implementation uses the same $[\mathrm{time}, \mathrm{channel}]$ order. A temporal mask records the valid sample positions. Fine-grained semantic annotations are attached to the training and validation partitions; test reference texts are not used to supply semantic answers during generation.

EEG–text alignment and semantic instruction tuning use the same precomputed EEG cache. The cached preprocessing includes an adjacent-difference operation along the final, channel dimension of the stored $[\mathrm{time}, \mathrm{channel}]$ array. Baseline-specific input adaptation follows each method's implementation.

B.2 Evaluation Protocol and Metric Definitions

Sentence-level semantic matching. We use the official SemKey candidate-pool construction (Wang et al., 2026). For each $N \in \{2, 4, 10, 24\}$, a trial samples N test observations without replacement. Their reference sentences form the candidate pool, and each of the N generated sentences is used as a query. Query and reference embeddings are normalized, and references are ranked by cosine similarity. A query is correct when the highest-ranked reference is the one paired with that observation. We average the query accuracies over 10,000 trials for each N. The evaluation seed is 42 and is reset independently for each value of N.

Sampling is performed over test observations without deduplicating their reference texts. The same sentence encoder and candidate-pool procedure are used for model comparison. Reference candidates are introduced only after generation for scoring. All 3,206 test observations are retained, including cases with malformed intermediate output or empty predictions. Only the parsed final sentence is submitted to the sentence evaluator.

Content recovery and distributional alignment. Content-word recall is computed after lowercasing, tokenizing, and removing English stopwords and non-alphanumeric tokens. For each reference containing at least one retained token, recall is the multiset overlap between predicted and reference content words divided by the reference content-word count. We average these per-example recalls.

Fréchet distance compares the means and covariance matrices of the generated and reference sentence embeddings from `all-mpnet-base-v2`, following the SemKey evaluator (Wang et al., 2026). It measures distributional alignment rather than correctness of individual sentence pairs.

Lexical overlap. We compute BLEU-1–4 (Papineni et al., 2002) using the SemKey evaluator's eight reference variants for each target sentence. The score for each order is the mean of the per-sentence scores, rather than corpus BLEU. The same reference construction is used across methods.

Generation diversity. Distinct-1 and Distinct-2 (Li et al., 2016) are the numbers of unique unigrams and bigrams divided by their respective total counts across the generated corpus.

Self-BLEU (Zhu et al., 2018) averages BLEU-4 scores obtained by treating each sampled output as the hypothesis and the other sampled outputs as references. Following the evaluator, at most 1,000 outputs are sampled for this calculation.

Head Entropy is the base-2 Shannon entropy of the empirical distribution of the first two tokens of generated sentences; outputs shorter than two tokens are excluded from this metric. All three diversity measures use lowercased NLTK tokenization. Lower Self-BLEU and higher Distinct or Head Entropy indicate more diverse outputs, which we assess together with semantic and content recovery metrics.

Reporting scales. BLEU, N-way accuracies, and Self-BLEU are reported as percentages. Content-word recall and Distinct-1/2 are reported as ratios. Head Entropy is reported in bits, and Fréchet distance is reported without percentage scaling.

B.3 Implementation Details

We use Flan-T5-large (Chung et al., 2024) as the LLM.

Representation alignment. We train the EEG representation network through contrastive alignment, representation regression, and EEG-conditioned sentence reconstruction, while keeping the LLM frozen. Training uses Adam with a batch size of 72, a peak learning rate of 2×10^{-4}, and a minimum learning rate of 10^{-5}. The cosine learning-rate schedule is configured for 50 epochs, including 15 warmup epochs, with training seed 42. The EEG Transformer uses six input blocks, six output blocks, eight attention heads, and a hidden width of 128. The loss weights for contrastive alignment, representation regression, and sentence reconstruction are 0.5, 0.7, and 0.5, respectively. We use the EEG representation checkpoint saved after the 38th training epoch for semantic instruction tuning.

Semantic instruction tuning. After alignment, we freeze the EEG representation network and perform semantic instruction tuning using AdamW (Loshchilov & Hutter, 2019) and LoRA (Hu et al., 2022). LoRA is applied to the query and value projections of the T5 attention layers. We update the EEG projector, LoRA adapters, and structural-token parameters while keeping the LLM's base parameters frozen. These trainable components use separate parameter groups, with the learning rates listed in Table 9. Structural-token updates use zero weight decay. The learning-rate schedule uses three warmup epochs followed by cosine decay to a minimum learning rate of 10^{-6}. The instruction-tuning seed is 648. We select the checkpoint with the lowest total validation loss; test scores are not used for checkpoint selection.

Table 9: Semantic instruction-tuning settings.

Setting	Value
Language backbone	Flan-T5-large
Optimizer	AdamW
Epochs / batch size	25 / 24
LoRA rank / scale / dropout	8 / 16 / 0.1
EEG projector learning rate	2×10^{-4}
LoRA learning rate	10^{-5}
Structural-token learning rate	10^{-3}
Weight decay (non-token groups)	0.01
Semantic loss weight λ_{sem}	0.3
Overall loss scale γ	1.5

Decoding settings. Inference uses deterministic greedy autoregressive decoding with one beam and no sampling. The maximum number of generated tokens is 192, with a minimum of three new tokens. The repetition penalty is 1.1, and repeated trigrams are disallowed. These limits apply to the complete generated sequence, including the semantic QA pairs and the final sentence.

B.4 Protocol for EEG Input Controls

Shared evaluation conditions. The input-control evaluation uses the selected SIRE checkpoint without further training. Input substitution is performed after preprocessing and immediately before the EEG representation network. The reference sentence, decoding instruction, question bank, and other non-EEG inputs remain associated with the original test observation. All EEG representations and the CoT are regenerated from the substituted input; semantic predictions obtained from the original EEG are not reused. The three conditions use the same test population, reference candidate pools, and decoding configuration.

Real EEG. Let $\widetilde{X}_i$ denote the preprocessed recording and M_i its valid EEG support. The support identifies recorded channels and valid temporal samples, excluding padded positions. The real-input condition evaluates $\widetilde{X}_i$ against its paired reference sentence Y_i.

Shuffled EEG. Let u_i denote the original sentence identity. We construct a one-to-one permutation π of the test observations satisfying $u_{\pi(i)} \neq u_i$. This excludes both self-pairing and substitution by another recording of the same sentence. To reduce contextual mismatches, the permutation is constructed within groups defined by dataset version, participant, and reading context.

Each EEG tensor moves together with its validity mask:

$$\left(\widetilde{X}_i^{\text{shuf}}, M_i^{\text{shuf}}\right) = \left(\widetilde{X}_{\pi(i)}, M_{\pi(i)}\right). \tag{19}$$

The substituted recording is evaluated against the original target Y_i, rather than the donor's reference sentence. The temporal and channel structure within each EEG observation remains intact. The permutation is constructed over the test set before batching.

Gaussian noise. For each recorded channel c, let μ_{ic} and σ_{ic} denote the mean and standard deviation of $\widetilde{X}_i(t, c)$ over its valid temporal samples. We draw independent samples $\epsilon_{itc} \sim \mathcal{N}(0, 1)$ and

construct

$$\widetilde{X}_i^{\text{noise}}(t,c) = \begin{cases} \mu_{ic} + \sigma_{ic}\epsilon_{itc}, & M_i(t,c) = 1, \\ \widetilde{X}_i(t,c), & M_i(t,c) = 0. \end{cases} \tag{20}$$

The Gaussian sampling distribution therefore matches the channel-wise mean and variance of the real input, while replacing its temporal and cross-channel dependence with independent samples. Statistics are computed only over valid samples, excluding padding. The original validity mask and padded values are preserved, and zero-variance channels remain constant. Noise is introduced after preprocessing so that subsequent normalization does not alter the intended scale matching.

Reproducibility and interpretation. The proposed protocol uses seed 42 for input construction, with separate random generators for the EEG permutation, Gaussian samples, and metric candidate pools. The perturbations are fixed before inference and are not selected according to their evaluation scores.

We primarily examine changes in N-way semantic matching and BLEU scores relative to Real EEG. FD depends on the corpus-level distribution rather than individual EEG–sentence pairings: if shuffling simply permutes the generated sentences, FD remains unchanged even when correspondence is lost. Gaussian noise serves as a complementary control, since a performance drop under noise alone can also reflect sensitivity to an unfamiliar input distribution. These experiments assess EEG input dependence; they do not directly measure autoregressive error propagation.

B.5 Effect of LoRA Rank

We compare SIRE with LoRA ranks $r \in \{4, 8, 16\}$. The results for the default setting, $r = 8$, are reproduced from Tables 2 and 3.

Table 10: Semantic alignment and content recovery of SIRE with different LoRA ranks. N-way accuracies are reported as percentages. C. Recall denotes content-word recall and is reported as a ratio. Fréchet distance is reported without percentage scaling. Arrows indicate preferred directions; bold denotes the best result in each column.

	Semantic matching (%)				Content recovery	Distribution
LoRA rank	2-way ↑	4-way ↑	10-way ↑	24-way ↑	C. Recall ↑	Fréchet distance ↓
$r = 4$	72.68	49.45	27.63	14.28	0.0264	0.3974
$r = 8$	**73.38**	**50.50**	**28.36**	**14.85**	**0.0326**	0.3951
$r = 16$	72.46	50.41	28.28	14.70	0.0292	**0.3647**

Table 11: Lexical overlap and generation diversity of SIRE with different LoRA ranks. BLEU and Self-BLEU are reported as percentages. Dist-n denotes Distinct-n and is reported as a ratio. H. Ent. denotes sentence-prefix entropy in bits. Arrows indicate preferred directions; bold denotes the best result in each column.

	Lexical overlap (%)				Generation diversity			
LoRA rank	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	Dist-1 ↑	Dist-2 ↑	Self-BLEU (%) ↓	H. Ent. ↑
$r = 4$	25.58	9.35	2.99	1.28	0.0406	0.1433	75.10	4.08
$r = 8$	**26.85**	**10.16**	**3.96**	**2.18**	0.0411	0.1368	75.16	4.00
$r = 16$	24.53	8.04	2.37	0.94	**0.0463**	**0.1636**	**73.01**	**4.43**

Tables 10 and 11 show that $r = 8$ achieves the highest scores across all four N-way settings, content-word recall, and BLEU-1 through BLEU-4. Increasing the rank to $r = 16$ improves generation diversity and reduces Fréchet distance, but does not improve semantic matching or lexical agreement with the target sentences. Within the tested range, a larger LoRA rank therefore does not consistently improve decoding performance: $r = 8$ provides stronger agreement with individual references, whereas $r = 16$ produces more diverse outputs and a closer corpus-level semantic distribution.

B.6 Fixed-Order and Adaptive QA Generation

We compare SIRE with a fixed-order QA variant that answers the two questions associated with each reading context in their prescribed order before generating the final sentence. These pairs are Q_1 followed by Q_2 for sentiment reading, Q_3 followed by Q_4 for normal reading, and Q_5 followed by Q_6 for task-specific reading. Both variants use fine-grained QA instruction and LoRA adaptation. SIRE selects unused questions from the complete question bank based on the EEG-conditioned context and preceding QA history, and determines when to end the QA sequence. The SIRE results below are reproduced from Tables 2 and 3.

Table 12: Semantic alignment and content recovery with fixed-order and adaptive QA generation. Both variants use LoRA. N-way accuracies are percentages, C. Recall is a ratio, and Fréchet distance is unscaled. Arrows indicate preferred directions; bold denotes the best result in each column.

	Semantic matching (%)				Content recovery	Distribution
QA policy	2-way ↑	4-way ↑	10-way ↑	24-way ↑	C. Recall ↑	Fréchet distance ↓
Fixed order	71.98	48.53	26.43	12.89	0.0318	0.5065
Adaptive (SIRE)	**73.38**	**50.50**	**28.36**	**14.85**	**0.0326**	**0.3951**

Table 13: Lexical overlap and generation diversity for the outputs evaluated in Table 12. BLEU and Self-BLEU are percentages; Dist-n denotes Distinct-n and is reported as a ratio. H. Ent. denotes sentence-prefix entropy in bits. Arrows indicate preferred directions; bold denotes the best result in each column.

	Lexical overlap (%)				Generation diversity			
QA policy	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	Dist-1 ↑	Dist-2 ↑	Self-BLEU (%) ↓	H. Ent. ↑
Fixed order	**27.09**	**10.53**	3.85	1.50	0.0235	0.0741	90.00	2.55
Adaptive (SIRE)	26.85	10.16	**3.96**	**2.18**	**0.0411**	**0.1368**	**75.16**	**4.00**

Table 12 shows higher semantic matching accuracy with adaptive QA generation across all four N-way settings. In particular, 24-way accuracy increases from 12.89% to 14.85%, a gain of 1.96 percentage points, while Fréchet distance decreases from 0.5065 to 0.3951. Content-word recall increases modestly from 0.0318 to 0.0326. These results indicate better discrimination among candidate references and closer generated and reference distributions in semantic space, with a smaller gain in content-word recovery.

Table 13 shows higher BLEU-1 and BLEU-2 for the fixed-order variant, whereas SIRE achieves higher BLEU-3 and BLEU-4. SIRE also produces more diverse outputs, with higher Distinct-1/2 and sentence-prefix entropy and lower Self-BLEU. Together, these results support the adaptive QA policy for semantic matching and generation diversity in this comparison. The policies differ in question selection and termination as well as ordering, so this comparison does not isolate the effect of question order alone.