EEG2VIS-BENCH: DECODING STRUCTURED VISUALIZATION QUERIES FROM EEG DURING QUERY INTERPRETATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can brain activity recorded while a person reads an analytical question be decoded into the chart they want? The target is an executable visualization program, correct only when its chart type, aggregation, and filter conditions are all correct together. We introduce EEG2VIS-Bench, an EEG dataset of 3,177 trials from 23 data-analysis experts. In each trial, a participant reads a data table, then a question about it, then inspects the resulting chart; each question is paired with a target program derived from nvBench. Our baseline trains an EEG encoder to match a text language model's representations of the questions, after first pre-training it to reconstruct masked EEG from the table and chart segments. It predicts the exact full program for 27.68% of held-out trials, compared with 2.67% when each trial's EEG is replaced by a recording of the same length from another trial, so the gain comes from the signal itself rather than its duration. Accuracy drops to 15.57% for unseen participants and falls further for program structures absent from training.

1 INTRODUCTION

Non-invasive brain–computer interfaces support control through motor imagery and visually evoked responses (Pfurtscheller et al., 2010; Krepki et al., 2007; Värbu et al., 2022), as well as implicit interaction based on ongoing brain activity (Zander et al., 2014; Frey et al., 2014). Neural decoding of language, with both non-invasive and intracranial recordings, extends this setting to sequences, from typed and read text (Duan et al., 2023; Liu et al., 2024; Lévy et al., 2025) to speech (Moreira et al., 2025; Singh et al., 2025), but executable programs introduce additional requirements: their fields, operators, and logical constraints must be jointly correct. Visualization queries provide a concrete test of this structured prediction problem, linking an analytical request to a program whose components and rendered output can be evaluated.

We study the prediction of visualization programs from EEG recorded while participants read and interpret analytical queries. A controlled table–query–visualization sequence supplies both contextual stimuli and executable VisQuery targets derived from nvBench (Luo et al., 2021). This design makes it possible to examine which aspects of a compositional program are recoverable from query-stage EEG and how that prediction transfers across participants and program structures.

EEG2VIS-Bench contains 3,177 collected stimulus groups from 23 experts, with separate epochs for table reading, query interpretation, and chart evaluation. The query epoch is the inference input. Table and chart epochs support self-supervised pre-training, and a teacher–student objective aligns EEG representations with semantic representations of the queries. Grammar-constrained decoding connects these representations to executable outputs.

Our experiments compare the decoder with controls based on query surface features, duration, participant identity, and shuffled EEG. The matched-shuffle control replaces the recorded EEG with a length-compatible trial, testing the contribution of the trial's signal beyond its duration. We also test the contribution of contextual pre-training and evaluate transfer to held-out participants and unseen programs.

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.

Our contributions are: (1) an EEG-to-program task with executable visualization targets; (2) a three-stage dataset pairing query-interpretation EEG with compositional programs; and (3) a teacher–student baseline evaluated for signal dependence, pre-training, and transfer.

2 Related Work

EEG datasets. SEED (Zheng & Lu, 2015), DEAP (Koelstra et al., 2012), and SEED-VII (Jiang et al., 2025) pair EEG with emotion labels, while OpenBMI (Lee et al., 2019) includes motor-imagery recordings. SEED-V (Liu et al., 2022) adds eye movements to emotion recordings. Natural-reading datasets such as ZuCo (Hollenstein et al., 2018; 2020), ChineseEEG (Mou et al., 2024; Chen et al., 2025), and TMNRED (Bai et al., 2025) pair EEG with linguistic stimuli, and ZuCo features have been used to augment NLP models (Hollenstein et al., 2021). Our recording protocol uses analytical queries grounded in data tables and pairs each query epoch with the corresponding visualization program.

EEG generation. DeWave (Duan et al., 2023), EEG2Text (Liu et al., 2024), and contrastive EEG–text pre-training (Wang et al., 2024b) decode EEG into text, while DreamDiffusion (Bai et al., 2024) and BrainDreamer (Wang et al., 2026) generate images. Jo et al. (2024) show that EEG-to-text models can score similarly on noise inputs, which motivates the signal controls in Section 5.2. In EEG2VIS, the output is an executable specification: predicting the chart type alone is insufficient because the fields, aggregations, and logical clauses jointly determine the rendered chart.

EEG representation learning. Convolutional and attention-based encoders such as EEGNet (Lawhern et al., 2018) and EEG Conformer (Song et al., 2023), graph-based classifiers (Zhang et al., 2024), and pre-trained EEG models (Wang et al., 2024a; Zhou et al., 2025) are the main EEG encoder families; Craik et al. (2019) review earlier deep-learning approaches to EEG classification. Our student follows the masked pre-training approach on task-stage epochs of the same recordings.

Structured visualization queries. nvBench (Luo et al., 2021) supplies natural-language queries paired with visualization specifications across 105 domains. We derive the stimuli and target programs from these pairs and collect EEG during their interpretation. The resulting task retains an explicit reference program, enabling exact comparison of the complete prediction and its chart, aggregation, and clause components. Our grammar-constrained decoder draws on syntax-tree decoding for text-to-SQL (Yu et al., 2018; Liu et al., 2025).

3 The EEG-to-Visualization (EEG2VIS) Task

3.1 Task Formulation

The EEG2VIS task maps EEG recorded during query interpretation to an executable visualization program represented as a syntax tree Y.

Input and Context. Let $E \in \mathbb{R}^{C\times T}$ denote a continuous preprocessed EEG recording for one participant, with C=31 channels and T time samples. During the data collection paradigm, a trial is constructed from a stimulus triple: a data *table* S_{table}, a natural-language *text query* S_{query} describing a desired chart, and the corresponding ground-truth *visualization* S_{vis}.

Using event markers, we segment the continuous EEG E into three task-aligned epochs per trial: X_{table} (table reading), X_{query} (query interpretation), and X_{vis} (visualization evaluation).

Output Target. We represent the displayed query as a structured VisQuery Y that specifies the intended chart. The detailed compositional grammar of Y is defined in Section 3.2.

Task Objective. Given the EEG epoch X_{query} recorded during query interpretation, the objective is to learn a decoding function f_{decode} to predict the corresponding structured output: $\hat{Y} = f_{\text{decode}}(X_{query})$.

The decoder takes X_{query} at inference. The flanking epochs X_{table} and X_{vis} support pre-training; stage-specific epochs also support the descriptive analyses.

3.2 Output Space: VisQuery as a Syntax Tree

To capture compositional visualization structure, we encode specifications derived from nvBench (Luo et al., 2021) using the following task-specific VisQuery grammar:

$$\textit{VisQuery} ::= \langle \textit{ChartType},\ \textit{Encodings},\ \textit{Clauses} \rangle \tag{1}$$

$$\textit{Encodings} ::= \{\textit{Encoding}\}^{+} \tag{2}$$

$$\textit{Encoding} ::= \langle \textit{Field},\ \textit{Role},\ \textit{Aggregation} \rangle \tag{3}$$

$$\textit{Aggregation} ::= \textit{None} \mid \textit{count} \mid \textit{sum} \mid \textit{avg} \mid \textit{min} \mid \textit{max} \tag{4}$$

$$\textit{Clauses} ::= \emptyset \mid \textit{Clause} \mid (\textit{Clause} \wedge \textit{Clause}) \mid (\textit{Clause} \vee \textit{Clause}) \tag{5}$$

$$\textit{Clause} ::= \langle \textit{Field},\ \textit{Op},\ \textit{Value} \rangle \tag{6}$$

ChartType (required) fixes the global visualization form (*e.g.,* bar, line, scatter). *Encodings* (at least one required) map data fields to visual roles (x-axis, y-axis, colour, size) with optional aggregation. *Clauses* (optional) encode logical constraints such as filters or ranges, and compose via Boolean connectives.

Illustrative query. Consider a table with fields `region`, `sales`, and `year`, and the request "Show total sales by region for 2024 as a bar chart." Its program selects a bar chart, maps `region` to the horizontal axis, maps `sales` to the vertical axis with `sum` aggregation, and applies the clause `year` $= 2024$:

$$\begin{aligned} Y = \langle \texttt{bar},\ \{&\langle \texttt{region}, x, \textit{None} \rangle, \langle \texttt{sales}, y, \texttt{sum} \rangle\}, \\ &\{\langle \texttt{year}, =, 2024 \rangle\} \rangle. \end{aligned} \tag{7}$$

A prediction that substitutes `avg` for `sum` has the correct chart type and clause but fails aggregation accuracy and full-program exact match. The rendered result then shows average rather than total sales. This example illustrates how the program metrics relate to chart content.

3.3 Evaluation Metrics

Full-query exact match. We measure whether the predicted VisQuery exactly matches the reference under canonical normalisation: $\text{Acc}_{\text{full}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{Y}^{(i)} \equiv Y^{(i)}]$.

Component accuracy. We report exact agreement separately for chart type, aggregation, and clause structure. Aggregation accuracy is computed on examples with a non-empty reference aggregation; chart and clause accuracy use the full test partition.

4 The EEG2VIS-Bench Dataset

4.1 Cognitive Paradigm and Stimuli

EEG2VIS-Bench employs a sequential three-stage design that mirrors natural data-analysis workflow with separately marked recording stages (Figure 1). **Task 1 (Table reading):** Participants read a data table to build a mental model (Johnson-Laird, 2010) of structure and content. **Task 2 (Query interpretation):** The corresponding natural-language query is displayed; participants comprehend its requirements and form a mental image of the expected chart. This stage provides the target EEG epoch X_{query}. **Task 3 (Visualization evaluation):** The ground-truth chart is shown for comparison against the mental representation. Stage markers align each EEG epoch with its stimulus; the flanking stages provide contextual EEG for pre-training (Section 5).

Twenty-three healthy, right-handed adults (12M/11F; mean age 23.3 ± 2.0) participated. All had professional experience in data analysis; 91.3% held postgraduate degrees. Non-native English speakers had IELTS reading $\geq$7.5. Participants were recruited for their experience interpreting analytical queries and visualizations. The protocol was approved by the institutional ethics commission; all participants provided written informed consent and were compensated at $30/hour.

We derive stimuli from nvBench (Luo et al., 2021), which pairs natural-language queries with executable visualization specifications across 105 domains. These specifications supply the VisQuery targets. Stimuli were curated from nvBench by removing entries with incomplete groups, ambiguous

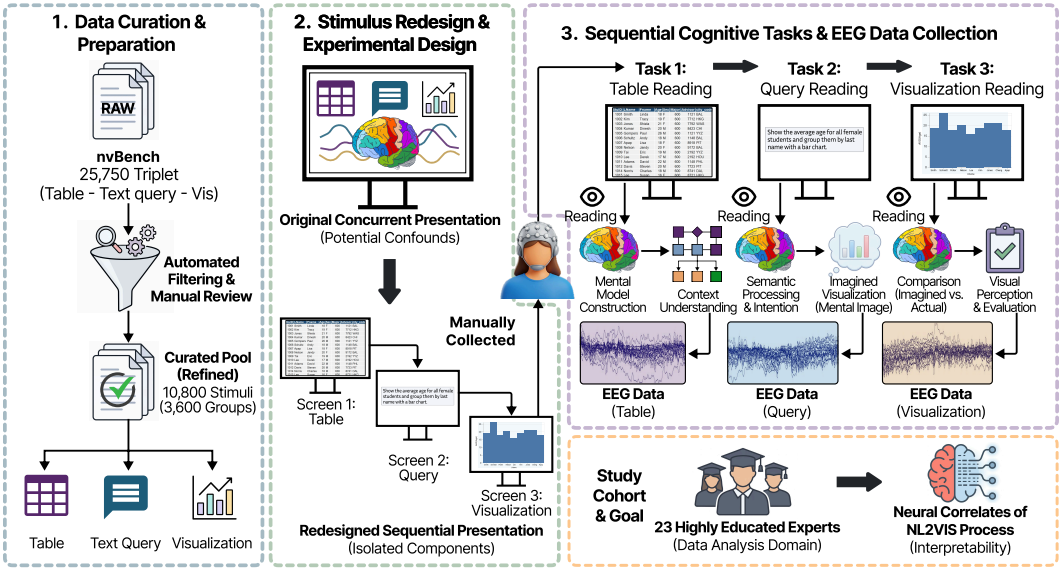


Figure 1: **EEG2VIS-Bench data collection overview.** EEG collection workflow: stimulus triples are presented sequentially; experts read the table, interpret the query, then evaluate the chart, with stage-aligned markers.

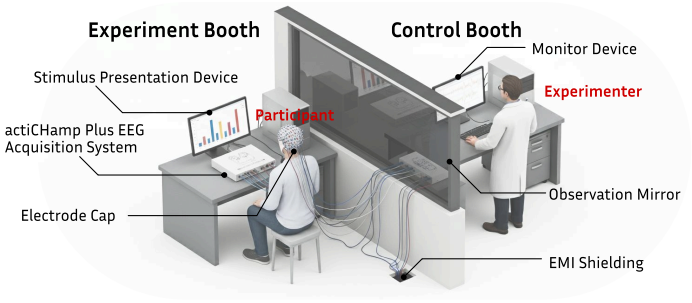


Figure 2: Two-room experimental setup with EMI shielding.

instructions, or factual errors, yielding 3,600 table–query–visualization triples (10,800 individual stimuli). Each triple was reviewed by two annotators to ensure the query unambiguously specifies a single visualization.

4.2 EEG Acquisition and Neurophysiological Signatures

The participant was seated in an EMI-shielded Experiment Booth equipped with the stimulus monitor, EEG amplifier, and electrode cap. The experimenter monitored from an adjacent Control Booth via a one-way observation mirror. Participants were seated 90 cm from a 24-inch 2K monitor. Appropriately sized EEG caps were fitted with conductive gel to maintain electrode impedances below 20 kΩ, checked every 30 minutes.

EEG was recorded using a 32-channel actiCHamp Plus system (Brain Products GmbH, 2026) in an EMI-shielded laboratory (Figure 2), sampled at 1000 Hz (0.1–100 Hz bandpass). Stimuli were presented via PsychoPy (Peirce et al., 2019); self-paced progression emitted cyclic markers (mk1/mk2/mk3) for segmentation. Each participant completed 60–202 stimulus groups (180–606 trials) in 2–3 hours, depending on individual pace and session availability. Offline preprocessing comprised re-referencing (Fz $\rightarrow$ Tp9/Tp10 average), 0.3–40 Hz bandpass filtering, ICA-based ocular artifact removal (Onton et al., 2006; Langlois et al., 2010; Fatourechi et al., 2007), and marker-based epoching (Figure 3).

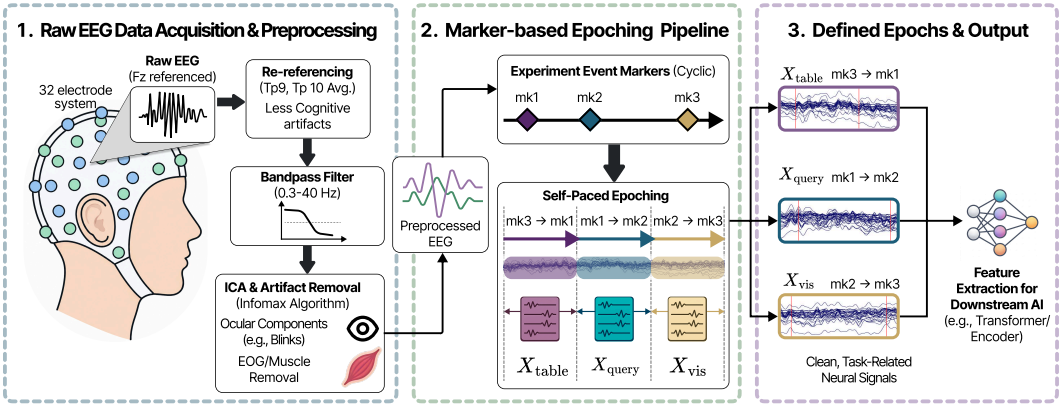


Figure 3: **EEG preprocessing pipeline.** Raw EEG is cleaned via re-referencing, bandpass filtering, and ICA artifact removal, then segmented into self-paced epochs.

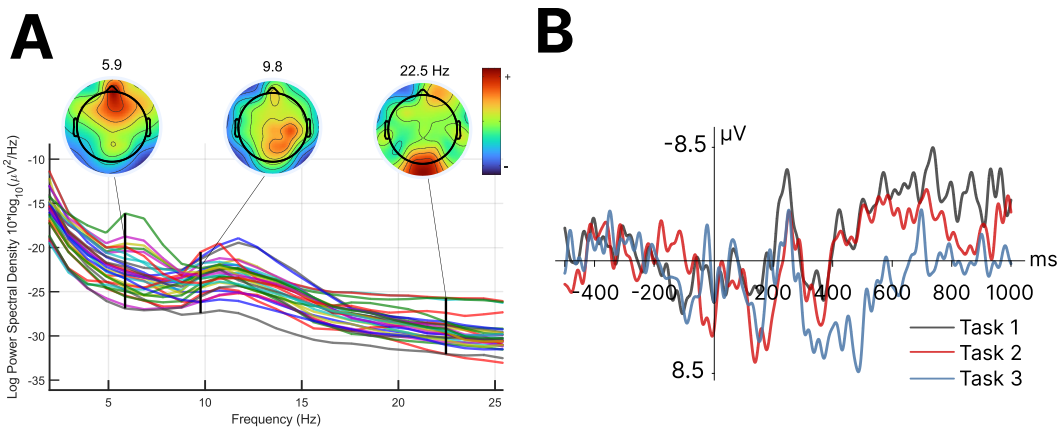


Figure 4: Power spectral density (left) and event-related potential waveforms (right) across table reading, query interpretation, and chart evaluation.

Model inputs contain 31 channels at 200 Hz after Fp1 exclusion, with epochs cropped or padded to 15 s. Approximately 46 participant-hours were recorded, yielding approximately 28 hours of retained epochs.

Stage-wise timing and neural activity. Task 2 (query interpretation) had a mean duration of 13.99 s (SD=9.04), compared with 12.75 s (SD=9.51) for table reading and 5.02 s (SD=4.35) for chart evaluation. A repeated-measures ANOVA found a main effect of task stage, $F(2, 44) = 84.14$, $p < 0.001$, $\eta_p^2 = 0.793$.

Figure 4 shows the power spectra and event-related potentials across the three stages. Query interpretation elicited a sustained negative-going potential beginning approximately 300 ms after stimulus onset, with an N400-like profile consistent with semantic processing (Kutas & Federmeier, 2011). The recordings also show an alpha peak near 10 Hz, fronto-midline theta during query interpretation, and beta activity across stages. Appendix B gives the corresponding scalp topographies.

4.3 Distribution and Complexity

The corpus contains 3,177 collected stimulus groups and 9,531 stage-specific epochs. The output space contains more than 2,000 distinct canonical programs, with chart types including bar, line, scatter, pie, area, and grouped or stacked bar. Programs use six aggregation options, 1–4 encodings, and zero to five or more clauses.

Table 1: Fixed data partitions. Each group contains one epoch per recording stage.

Partition	Stimulus groups	EEG epochs
Training	2,267	6,801
Validation	423	1,269
Test	487	1,461
Total	3,177	9,531

The fixed partition uses random seed 42 (Table 1). All three epochs from a stimulus group stay in the same partition. Participants are shared across these partitions; Section 5.3 reports a separate evaluation on held-out participants.

5 EXPERIMENTS

5.1 TEACHER–STUDENT DECODER

The Qwen-3 teacher (Yang et al., 2025) encodes textual stimuli as semantic representations V_{text}. Following teacher–student distillation (Hinton et al., 2015), an EEG Transformer student maps query-stage EEG X_{query} to V_{eeg} and is trained with semantic-alignment and task-prediction losses. Before this stage, the student learns to reconstruct masked EEG from table and chart epochs. We denote the resulting model Qwen-3-PT. Training uses 8× NVIDIA RTX 4090 GPUs.

The prediction heads produce posteriors over chart type, aggregation, clauses, and ordering.

Contextual pre-training. The training-set epochs X_{table} and X_{vis} supply a reconstruction objective for the EEG student. We mask 20% of contiguous blocks of approximately 100 ms. Let $\mathcal{M}$ contain the masked channel–time entries and let $\hat{X}$ denote the reconstruction. The loss averages squared error over these entries:

$$\mathcal{L}_{\text{pre}} = \frac{1}{|\mathcal{M}|} \sum_{(c,t)\in\mathcal{M}} (\hat{X}_{c,t} - X_{c,t})^2. \tag{8}$$

The learned encoder parameters initialize query-stage training. The pre-training ablation varies which contextual stages supply these epochs.

Teacher–student objective. After pre-training on table and chart epochs, the student is fine-tuned with a composite objective:

$$\mathcal{L}_{\text{total}} = \alpha\, \mathcal{L}_{\text{align}}(V_{eeg}, V_{text}) + \beta\, \mathcal{L}_{\text{task}}(\text{Head}(V_{eeg}),\, Y). \tag{9}$$

Here, $\mathcal{L}_{\text{align}}$ measures the distance between the EEG and teacher representations, and $\mathcal{L}_{\text{task}}$ is the classification loss for the target program components. The weights α and β balance the two objectives. The teacher supplies semantic targets during training; inference uses the trained EEG student and its prediction heads.

Grammar-constrained decoding. The grammar-constrained reranker selects the highest-scoring valid program:

$$\hat{Y} = \arg\max_{Y\in\mathcal{Y}_{\text{valid}}} P_{\text{neural}}(Y \mid X_{query}) \cdot P_{\text{prior}}(Y \mid \text{schema}). \tag{10}$$

Here, P_{neural} scores a program from the EEG predictions, and P_{prior} is a schema-conditioned program prior over the valid output space $\mathcal{Y}_{\text{valid}}$.

5.2 HELD-OUT EVALUATION AND SIGNAL CONTROLS

Model selection uses the 423 validation groups, and evaluation uses the fixed 487-group test partition. Table 2 compares Qwen-3-PT with feature-based controls and EEG substitutions.

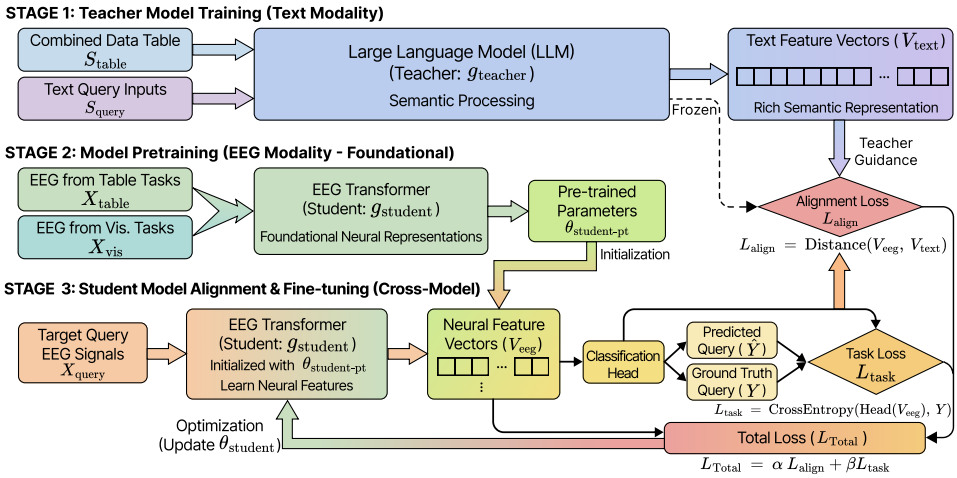


Figure 5: Teacher–student framework. The text teacher supplies semantic targets; the EEG student learns from table and chart epochs before query-stage alignment and prediction.

Table 2: Held-out results (percent). Chart, clause, and full exact match use 487 examples; aggregation uses the 473 examples with non-empty gold aggregation. Gaussian EEG, both shuffle controls, and Qwen-3-PT report five-seed mean ± SD; the other controls are single-run or deterministic.

Method	Chart	Agg.	Clause	Full exact
Majority	31.83	39.96	3.08	0.00
Query surface	52.36	41.86	43.33	2.26
Duration	51.54	34.25	52.98	2.87
Duration + padding	58.73	36.79	54.83	3.08
Participant prior	48.67	34.67	37.37	1.03
Session prior	45.79	35.10	45.38	1.64
Gaussian EEG	33.31 ± 1.71	31.08 ± 1.43	31.09 ± 2.37	0.82 ± 0.21
Cross-participant shuffle	32.11 ± 1.30	29.94 ± 1.08	32.16 ± 1.83	0.21 ± 0.15
Matched shuffle	47.76 ± 1.78	38.86 ± 2.12	38.85 ± 1.64	2.67 ± 0.32
Qwen-3-PT	80.57 ± 1.33	54.93 ± 1.53	58.97 ± 1.19	27.68 ± 2.05

Feature-based controls. Query-surface features are token, character, and punctuation counts. The duration control uses Task-2 reading time; duration plus padding additionally includes the padding pattern. Participant and session priors condition predictions on the corresponding recording identities.

Matched shuffle. All EEG-substitution controls use the same five trained checkpoints and preserve each target's valid-length mask and zero padding. Matched shuffle selects a length-compatible donor epoch from the same participant and truncates it to the target length. This supplies donors for 484 of the 487 test trials; the remaining three use length-compatible donors from other participants.

Other EEG substitutions. Cross-participant shuffle uses donor EEG from other participants. Gaussian EEG replaces the channel-normalized signal with zero-mean, unit-variance noise.

Qwen-3-PT exceeds matched shuffle by 25.01 percentage points in full exact match. This difference supports a contribution from trial-specific EEG beyond duration and padding. Duration-based controls also predict chart type and clauses above the majority baseline, but rarely recover the full program.

5.3 Pre-Training and Generalization

Table 3 compares pre-training on table epochs, chart epochs, and their combination. The ablations match update budgets and downstream training schedules.

Joint pre-training improves full exact match by 4.81 percentage points. Pre-training on either table or chart epochs alone produces little change relative to the model trained directly on query epochs.

Table 3: Pre-training ablation on the fixed test split, five-seed mean ± SD.

Pre-training	Full exact (%)
None	22.87 ± 1.73
X_{table} only	23.24 ± 1.76
X_{vis} only	23.04 ± 1.76
$X_{table} + X_{vis}$	27.68 ± 2.05

Table 4: Test-set exact match by training-program overlap (seed 0).

Test examples	Correct / total	Exact match (%)
Seen program	120 / 275	43.64
Unseen program	15 / 212	7.08
Of which: unseen template	14 / 209	6.70

For participant transfer, we retrain the full pipeline in 23 leave-one-subject-out folds. Each held-out participant is excluded from pre-training, training, and validation. Full exact match is $15.57 \pm 8.52\%$: the mean and SD across participants after averaging five seeds within each fold.

Table 4 separates test examples by overlap with training programs. A template is the normalized clause structure with chart type and aggregation removed. Exact match falls sharply on unseen programs, including those with unseen templates.

5.4 Render-Level Evaluation

We sample 150 predictions from the seed-0 test outputs before manual inspection, stratifying by chart type and clause complexity with at least five examples per chart type and three per complexity level.

Two graduate annotators, separate from the EEG participants and blind to model identity, view predicted and reference charts in randomized left/right order, together with the query and canonical programs. A third annotator adjudicates disagreements. The rubric is:

Score	Criterion
5	Identical charts
4	Minor cosmetic differences
3	One incorrect component
2	At least two errors
1	Unrecognizable chart

Before adjudication, Krippendorff's α is 0.86, exact agreement is 81.3%, and within-one agreement is 100%.

Table 5 reports the adjudicated ratings alongside program exact match on the same outputs. Render ratings distinguish cosmetic changes from errors in chart content.

Table 5: Evaluation of 150 held-out predictions.

Measure	Percentage
Rating ≥ 4: identical or minor cosmetic differences	73.3
Rating ≥ 3: at most one incorrect component	93.3
Program exact match	27.33

6 DISCUSSION

Program structure and transfer. Chart type is recovered more accurately than aggregation or clauses. Full-program exact match requires these choices to be correct together, making composition a substantial source of error. The further drop on unseen programs shows that performance depends strongly on training-program coverage. Participant transfer also reduces accuracy, even with contextual pre-training.

Recording conditions. The recordings capture reading, comprehension, and visualization-related processing during displayed analytical queries. The cohort consists of data-analysis experts with strong English reading proficiency. These conditions define the population and interaction setting represented by the benchmark.

Data stewardship. EEG reuse should respect participant consent and protect against re-identification and non-consensual monitoring. Demographics are reported at cohort level.

7 CONCLUSION

EEG2VIS-Bench pairs EEG recorded during analytical query interpretation with executable visualization programs. The teacher–student decoder benefits from joint table and chart pre-training and outperforms matched-shuffle controls. Transfer to new participants and unseen programs is the main difficulty exposed by the evaluations. Dataset and code: `https://anonymous.4open.science/r/EEG2VIS-benchmark-1142/`.

AI USE STATEMENT

AI tools were used to assist with writing and polishing the manuscript, including drafting textual revisions and checking references.

REPRODUCIBILITY STATEMENT

Section 4.3 specifies data partitions; Section 4.2 describes EEG input processing. Section 5.3 gives the pre-training and participant-transfer protocols. Section 5.1 defines the training objective and decoder, Section 5.2 specifies the controls, and Section 5.4 details human annotation.

REFERENCES

Yanru Bai, Qi Tang, Ran Zhao, Hongxing Liu, Shuming Zhang, Mingkun Guo, Minghan Guo, Junjie Wang, Changjian Wang, Mu Xing, Guangjian Ni, and Dong Ming. TMNRED, a Chinese language EEG dataset for fuzzy semantic target identification in natural reading environments. *Scientific Data*, 12:701, 2025. doi: 10.1038/s41597-025-05036-2.

Yunpeng Bai, Xintao Wang, Yan-Pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. DreamDiffusion: High-quality EEG-to-image generation with temporal masked signal modeling and CLIP alignment. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 472–488. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72751-1_27.

Brain Products GmbH. actiCHamp Plus (actiCHamp series). `https://www.brainproducts.com/solutions/actichamp/`, 2026. Gilching, Germany. Accessed September 26, 2026.

Sitong Chen, Beiqianyi Li, Cuilin He, Dongyang Li, Mingyang Wu, Xinke Shen, Song Wang, Xuetao Wei, Xindi Wang, Haiyan Wu, and Quanying Liu. An EEG dataset for multimodal semantic alignment and neural decoding during reading and listening. *Scientific Data*, 13:148, 2025. doi: 10.1038/s41597-025-06466-8.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, 16(3):031001, 2019.

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, 2023. doi: 10.52202/075280-0432.

Mehrdad Fatourechi, Ali Bashashati, Rabab K Ward, and Gary E Birch. EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology*, 118(3):480–494, 2007.

Jérémy Frey, Christian Mühl, Fabien Lotte, and Martin Hachet. Review of the use of electroencephalography as an evaluation method for human-computer interaction. In *Proceedings of the International Conference on Physiological Computing Systems (PhyCS)*, pp. 214–223. SCITEPRESS, 2014. doi: 10.5220/0004708102140223.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. NIPS 2014 Deep Learning Workshop.

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.

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 (LREC)*, pp. 138–146. European Language Resources Association, 2020.

Nora Hollenstein, Cedric Renggli, Benjamin Glaus, Maria Barrett, Marius Troendle, Nicolas Langer, and Ce Zhang. Decoding EEG brain activity for multi-modal natural language processing. *Frontiers in Human Neuroscience*, 15:659410, 2021. doi: 10.3389/fnhum.2021.659410.

Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. SEED-VII: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 16(2):969–985, 2025. doi: 10.1109/TAFFC.2024.3485057.

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

Philip N. Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010. doi: 10.1073/pnas.1012933107.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. doi: 10.1109/T-AFFC.2011.15.

Roman Krepki, Gabriel Curio, Benjamin Blankertz, and Klaus-Robert Müller. Berlin brain–computer interface—the HCI communication channel for discovery. *International Journal of Human-Computer Studies*, 65(5):460–477, 2007. doi: 10.1016/j.ijhcs.2006.11.010.

Marta Kutas and Kara D. Federmeier. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62:621–647, 2011. doi: 10.1146/annurev.psych.093008.131123.

Dominic Langlois, Sylvain Chartier, and Dominique Gosselin. An introduction to independent component analysis: InfoMax and FastICA algorithms. *Tutorials in Quantitative Methods for Psychology*, 6(1):31–38, 2010.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. *GigaScience*, 8(5):giz002, 2019. doi: 10.1093/gigascience/giz002.

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

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, 2024. doi: 10.1109/BigData62323.2024.10825980.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2022. doi: 10.1109/TCDS.2021.3071170.

Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, and Yuyu Luo. A survey of text-to-SQL in the era of LLMs: Where are we, and where are we going? *IEEE Trans. Knowl. Data Eng.*, 37(10):5735–5754, 2025. doi: 10.1109/TKDE.2025.3592032.

Yuyu Luo, Nan Tang, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. Synthesizing natural language to visualization (NL2VIS) benchmarks from NL2SQL benchmarks. In *SIGMOD '21: International Conference on Management of Data*, pp. 1235–1247. ACM, 2021. doi: 10.1145/3448016.3457261.

João Pedro Carvalho Moreira, Vinícius Rezende Carvalho, Eduardo Mazoni Andrade Marçal Mendes, Aria Fallah, Terrence J. Sejnowski, Claudia Lainscsek, and Lindy Comstock. An open-access EEG dataset for speech decoding: Exploring the role of articulation and coarticulation. *Scientific Data*, 12:1017, 2025. doi: 10.1038/s41597-025-05187-2.

Xinyu Mou, Cuilin He, Liwei Tan, et al. ChineseEEG: A Chinese linguistic corpora EEG dataset for semantic alignment and neural decoding. *Scientific Data*, 11:550, 2024. doi: 10.1038/s41597-024-03398-7.

Julie Onton, Marissa Westerfield, Jeanne Townsend, and Scott Makeig. Imaging human EEG dynamics using independent component analysis. *Neuroscience & Biobehavioral Reviews*, 30(6):808–822, 2006. doi: 10.1016/j.neubiorev.2006.06.007.

Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203, 2019. doi: 10.3758/s13428-018-01193-y.

Gert Pfurtscheller, Brendan Z. Allison, Clemens Brunner, Günther Bauernfeind, Teodoro Solis-Escalante, Reinhold Scherer, Thorsten O. Zander, Gernot Mueller-Putz, Christa Neuper, and Niels Birbaumer. The hybrid BCI. *Frontiers in Neuroscience*, 4:30, 2010. doi: 10.3389/fnpro.2010.00003.

Aditya Singh, Tessy Thomas, Jinlong Li, Greg Hickok, Xaq Pitkow, and Nitin Tandon. Transfer learning via distributed brain recordings enables reliable speech decoding. *Nature Communications*, 16:8749, 2025. doi: 10.1038/s41467-025-63825-0.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

Kaido Värbu, Naveed Muhammad, and Yar Muhammad. Past, present, and future of EEG-based BCI applications. *Sensors*, 22(9):3331, 2022. doi: 10.3390/s22093331.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pp. 39249–39280, 2024a. doi: 10.52202/079017-1239.

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 (ACL)*, pp. 7278–7292, 2024b.

Ling Wang, Chen Wu, and Lin Wang. BrainDreamer: Reasoning-coherent and controllable image generation from EEG brain signals via language guidance. *IEEE Transactions on Multimedia*, pp. 1–11, 2026. doi: 10.1109/TMM.2026.3718223.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025.

Tao Yu, Michihiro Yasunaga, Kai Yang, Rui Zhang, Dongxu Wang, Zifan Li, and Dragomir Radev. Syntaxsqlnet: Syntax tree networks for complex and cross-domain text-to-SQL task. In *Proceedings of EMNLP*, 2018.

Thorsten O. Zander, Jonas Brönstrup, Romy Lorenz, and Laurens R. Krol. *Towards BCI-Based Implicit Control in Human–Computer Interaction*, pp. 67–90. Springer London, 2014. doi: 10.1007/978-1-4471-6392-3_4.

Kaiyuan Zhang, Ziyi Ye, Qingyao Ai, Xiaohui Xie, and Yiqun Liu. GNN4EEG: A benchmark and toolkit for electroencephalography classification with graph neural network. In *Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion)*, pp. 612–617, 2024. doi: 10.1145/3675094.3678475.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A Implementation Notes

Optional retrieval. An optional retrieval cache selects a stored training program by EEG-embedding cosine similarity; the evaluations reported here use the decoder predictions.

Text-input references. Query-only Qwen-8B reaches 10.68% full exact match; supplying both table and query yields 26.69%.

B Scalp Topographies

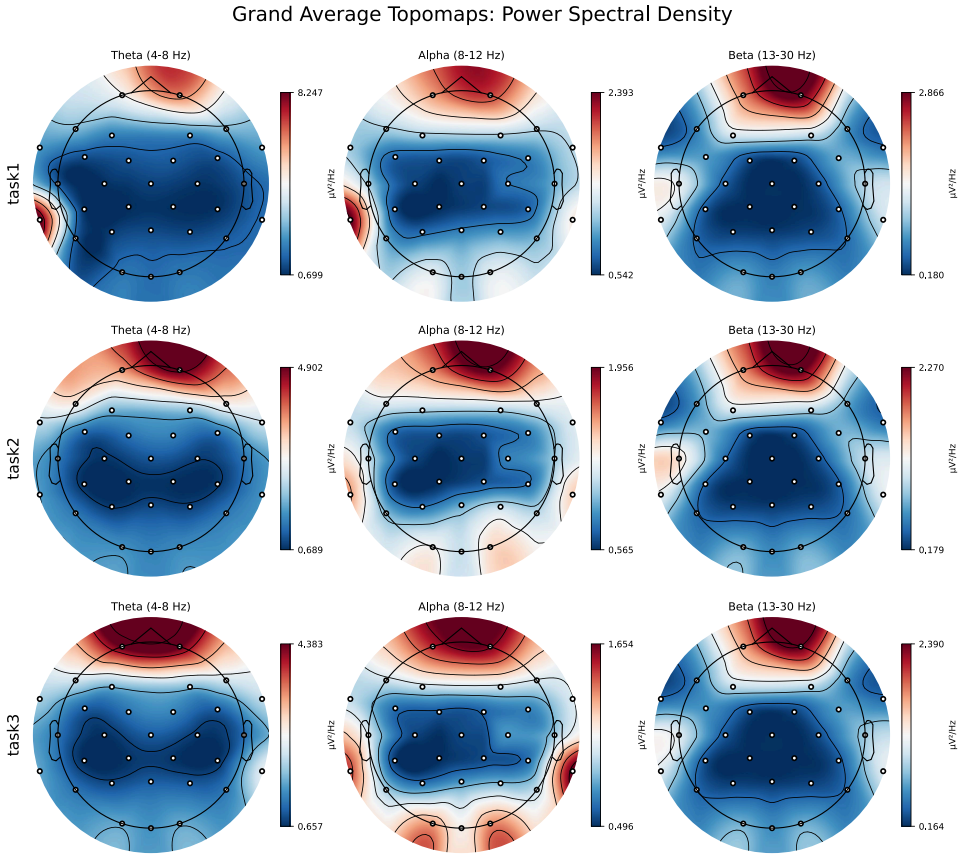


Figure 6: Scalp power topographies for theta, alpha, and beta bands across the three task stages.

C Datasheet for EEG2VIS-Bench

Following the datasheet framework (Gebru et al., 2021), we provide a datasheet for the EEG2VIS-Bench dataset.

Motivation. EEG2VIS-Bench was created to enable research on decoding structured visualization queries during query interpretation from non-invasive EEG signals, using data visualization as a concrete testbed.

Composition. The dataset contains 32-channel acquired EEG from 23 participants across 9,531 stage-specific epochs (3,177 per task stage), paired with structured VisQuery annotations derived from nvBench (Luo et al., 2021). Each trial includes raw EEG, preprocessed EEG, event markers, and the corresponding stimulus triple (table, query, visualization).

Collection process. EEG was recorded using a 32-channel actiCHamp Plus system at 1000 Hz in an EMI-shielded laboratory. The experimental protocol followed a three-stage paradigm (Section 4.1) with self-paced progression. All participants provided written informed consent and were compensated at $30/hour. The study was approved by the institutional ethics commission.

Preprocessing. EEG is acquired at 32 channels and 1,000 Hz. Model inputs use 31 channels at 200 Hz after Fp1 exclusion, with epochs cropped or padded to 15 seconds. Preprocessing: re-referencing (Tp9/Tp10), bandpass filtering (0.3–40 Hz), ICA-based artifact removal, and marker-based epoching.

Uses. Primary: benchmarking structured visualization-query decoding from EEG. Secondary: studying query-interpretation EEG, EEG representation learning, and brain–computer interfaces.

Distribution. The dataset, code, and baselines will be released at `https://anonymous.4open.science/r/EEG2VIS-benchmark-1142/` under a CC-BY-NC 4.0 licence for academic use.

Maintenance. The authors will maintain the dataset and respond to issues via the project repository.