EEG2MM-BENCH: A UNIFIED BENCHMARK FOR EEG-TO-MULTIMODAL TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG-to-multimodal decoding aims to retrieve or reconstruct images, videos, and text from electroencephalography (EEG) signals. However, the transferability of pretrained EEG representations to these tasks remains unclear, while inconsistent evaluation protocols and strong generative priors complicate comparisons. We introduce EEG2MM-Bench, a unified benchmark comprising 11 existing datasets with 468 hours of EEG recordings. We evaluate 17 task-specific architectures and three EEG foundation models across multimodal retrieval and reconstruction under shared data splits and consistent evaluation protocols. Retrieval uses a common alignment and ranking protocol. Reconstruction combines original-system evaluations with unified modality-specific pipelines, comparing EEG encoders under consistent downstream architectures and training procedures with frozen pretrained generators. Gaussian-noise input controls further assess reconstruction sensitivity to EEG replacement. By jointly evaluating stimulus matching and content recovery, EEG2MM-Bench provides a common framework for characterizing the capabilities and limitations of pretrained and task-specific EEG representations across multimodal decoding tasks.

1 INTRODUCTION

Brain-signal decoding aims to recover perceptual and semantic information from neural activity and translate it into interpretable outputs (Tang et al., 2023; Benchetrit et al., 2024; Li et al., 2024; Liu et al., 2025). Electroencephalography (EEG) is accessible and non-invasive and has long been used for brain–computer communication and control (Wolpaw et al., 2002; Farwell & Donchin, 1988). With advances in deep learning, EEG-to-multimodal (EEG2MM) decoding has emerged to link EEG signals with multimodal content, including images, videos, and text. Given an EEG response, EEG2MM aims either to retrieve the corresponding multimodal stimulus or to reconstruct its content (Song et al., 2024; Liu et al., 2024; Wang & Ji, 2022a). By decoding rich visual and semantic content, EEG2MM connects the study of how such information is represented in the brain (Song et al., 2024; Zhou et al., 2024) with the long-term goal of content-level assistive communication.

Task-specific models are trained for individual EEG2MM tasks. For multimodal content retrieval, each model selects the best match from candidates in the target modality—image, video, or text—based on the EEG input (Song et al., 2024; Ye et al., 2022; Zhou et al., 2024). For multimodal content reconstruction, each model conditions the corresponding generator on EEG to recover the target content (Dai et al., 2026; Wang et al., 2024b; Liu et al., 2026). EEG foundation models offer an alternative by learning general EEG representations through large-scale pretraining. These representations can then be adapted separately to each retrieval or reconstruction task.

However, the effectiveness of EEG representations via large-scale pretraining for multimodal decoding tasks remains unclear. As shown in Figure 1, clinical recordings account for most of the EEG recording hours used to pretrain EEG foundation models, including LaBraM (Jiang et al., 2024), REVE (El Ouahidi et al., 2025), and LUNA (Döner et al., 2025). Clinical tasks assess disease or physiological states, whereas EEG2MM targets multimodal content, creating a substantial gap in task objectives. This gap, together with potential EEG distribution shifts arising from differences in populations, acquisition protocols, and experimental paradigms (Shi et al., 2026; El Ouahidi et al., 2025), makes effective transfer of pretrained EEG representations to EEG2MM uncertain. As indicated in recent work (Banville et al., 2026), the performance of pretrained models approaches

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.

saturation on certain clinical classification tasks, whereas stimulus decoding remains challenging. Based on these observations, a systematic benchmark is needed to characterize the applicability of pretrained representations across EEG2MM tasks.

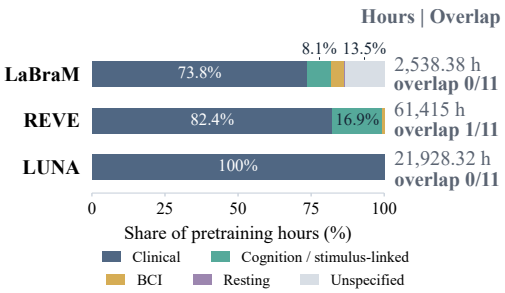


Figure 1: **Pretraining data composition.** Pretraining hours of LaBraM (Jiang et al., 2024), REVE (El Ouahidi et al., 2025), and LUNA (Döner et al., 2025) across clinical, cognition/stimulus-linked, BCI, resting-state, and unspecified recordings. We report their total hours and the identified source matches among our collected 11 EEG2MM datasets.

To comprehensively compare how well representations learned by EEG foundation models and task-specific models support multimodal decoding, evaluation should cover both retrieval and reconstruction across different output modalities. Successful global matching in retrieval (Song et al., 2024; Ye et al., 2022) does not guarantee the recovery of the fine-grained stimulus information required for reconstruction (Li et al., 2024; Dai et al., 2026). Reconstruction should also be evaluated under a controlled setting that allows the contributions of different EEG representations to be compared fairly. Existing methods differ in their adapters, generators, training procedures, and evaluation protocols, making it difficult to attribute differences in reported scores to the EEG representations themselves. Moreover, strong content priors in pretrained generators may produce plausible outputs despite weak dependence on EEG input. Prior work reports comparable EEG-to-text scores when EEG inputs are replaced with synthetic noise (Jo et al., 2026). A faithful benchmark therefore requires a common downstream reconstruction pipeline and noise-input comparisons.

Guided by these requirements, we introduce EEG2MM-Bench (Figure 2), a unified benchmark for evaluating EEG representations on multimodal retrieval and reconstruction tasks. We collect 11 existing EEG datasets covering image, video, and text to assess how pretrained EEG foundation models transfer to multimodal content decoding. Besides, we evaluate both retrieval and reconstruction in EEG2MM to examine whether EEG representations that support stimulus matching also enable faithful reconstruction. Specifically, we adopt a common downstream reconstruction pipeline within each modality to compare EEG representations under consistent generative conditions, complemented by noisy-input controls to assess EEG dependence. We also evaluate the original reconstruction pipelines to provide complementary comparisons of complete systems. Together, these designs support systematic and fair comparisons between EEG foundation models and task-specific models under shared data splits and evaluation protocols.

Overall, EEG2MM-Bench covers retrieval and reconstruction across image, video, and text using 11 existing datasets with 468 hours of EEG recordings. We standardize EEG–target pairing and data splits, and apply shared preprocessing to task-specific models within each dataset. Our results show that foundation models are competitive in retrieval but do not consistently outperform task-specific encoders in reconstruction. Our contributions are as follows:

- A unified benchmark built on 11 existing datasets for retrieval and reconstruction across image, video, and text, with consistent data splits and evaluation protocols.
- Controlled comparisons of EEG encoders under common downstream reconstruction pipelines, complemented by original-system evaluations and Gaussian-noise input.
- A systematic evaluation of 3 EEG foundation models and 17 task-specific architectures across EEG2MM retrieval and reconstruction tasks.

2 Related Work

EEG representation learning. EEG representation learning spans handcrafted spectral features, spatial filters, and covariance-based representations (Pfurtscheller & Lopes da Silva, 1999; Ramoser et al., 2000; Barachant et al., 2012), neural networks trained for specific downstream tasks, and models pretrained on large-scale EEG corpora. Task-specific networks learn spatio-temporal patterns from labeled EEG using convolutional architectures or convolution–attention hybrids and are commonly

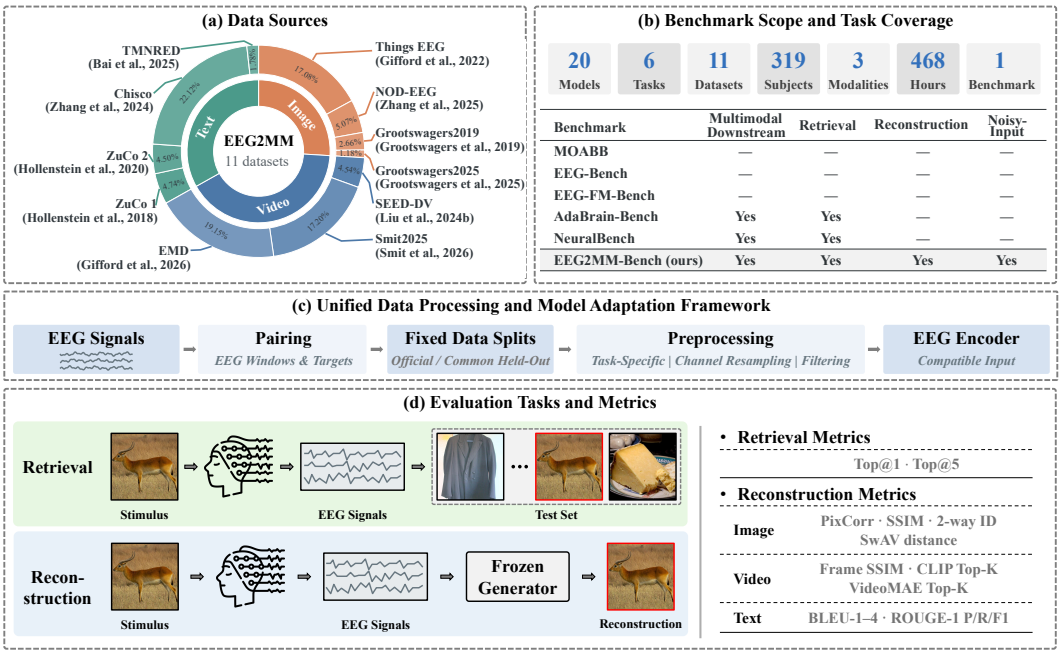


Figure 2: **Overview of EEG2MM-Bench.** The benchmark integrates image, video, and text datasets with consistent processing and evaluation protocols for EEG-based retrieval and reconstruction.

applied to tasks such as motor imagery and event-related potential decoding (Schirrmeister et al., 2017; Lawhern et al., 2018; Song et al., 2023). To obtain transferable representations, some pretrained methods predict masked latent features or discrete neural tokens (Kostas et al., 2021; Jiang et al., 2024), whereas others reconstruct masked EEG patches using architectures designed to model spatio-temporal structure or accommodate heterogeneous channel configurations (Wang et al., 2025; Döner et al., 2025; El Ouahidi et al., 2025). EEGPT further combines masked signal reconstruction with representation-level alignment (Wang et al., 2024a). EEG2MM-Bench evaluates how well these pretrained representations transfer to multimodal content decoding while accounting for their reported pretraining sources and overlap with downstream datasets.

EEG-to-multimodal decoding. EEG-to-multimodal decoding includes retrieval and reconstruction. Retrieval methods align EEG and stimulus representations in a shared space or aggregate word-level semantic predictions to identify the corresponding stimulus (Song et al., 2024; Ye et al., 2022; Zhou et al., 2024). Reconstruction methods instead convert EEG representations into conditioning signals for modality-specific generative models (Wang & Ji, 2022b; Duan et al., 2023; Liu et al., 2025; Li et al., 2025; Lan et al., 2023; Li et al., 2024; Kneeland et al., 2026; Fu et al., 2025; Liu et al., 2024; Zhou et al., 2026; Liu et al., 2026; Singh et al., 2026). Some studies compare recorded EEG with Gaussian-noise inputs (Liu et al., 2025; Li et al., 2025), while others analyze which spatial and temporal signal components support decoding (Singh et al., 2026). Because existing systems differ in their encoders, adapters, generators, and training procedures, their reported scores cannot directly isolate the contribution of EEG representations. EEG2MM-Bench therefore evaluates selected methods in their original reconstruction pipelines and separately compares EEG encoders within a common downstream pipeline for each reconstruction modality.

EEG benchmarks. Existing EEG benchmarks address standardized BCI evaluation, transfer across subjects and datasets, clinical applications, and the transferability of pretrained EEG representations. MOABB standardizes evaluation across EEG datasets and BCI paradigms (Jayaram & Barachant, 2018), while BEETL focuses on cross-subject and cross-dataset transfer (Wei et al., 2022). EEG-Bench evaluates pretrained models on clinical applications (Kastrati et al., 2025), whereas EEG-FM-Bench and AdaBrain-Bench cover broader EEG tasks and adaptation settings (Xiong et al., 2026; Wu et al., 2025). NeuralBench further evaluates heterogeneous EEG tasks and documents overlap between pretraining corpora and downstream datasets, but formulates stimulus-content decoding as

retrieval rather than reconstruction (Banville et al., 2026). More focused efforts examine controlled EEG–image retrieval, joint EEG–fMRI audiovisual reconstruction, and teacher-forcing-free EEG-to-text evaluation (Zhang et al., 2026a; Gao et al., 2025; Zhang et al., 2026b). Complementary work further investigates how noise inputs and teacher-forced evaluation affect reported EEG-to-text scores (Jo et al., 2026). Taken together, these efforts do not jointly evaluate retrieval and reconstruction across multiple output modalities. EEG2MM-Bench bridges this gap by evaluating both tasks under consistent data splits and evaluation protocols. For reconstruction, a common downstream pipeline within each modality enables controlled comparisons of EEG representations, while comparisons with Gaussian-noise inputs assess whether performance benefits from recorded EEG.

3 EEG2MM Benchmark

EEG2MM-Bench evaluates EEG-to-multimodal retrieval and reconstruction across 11 datasets using fixed per-dataset splits and a common evaluation framework. This section defines the tasks, dataset organization, evaluation pipelines, and metrics. Model selection and experimental configurations are described in Section 4.1.

3.1 Task Definition and Overview

Task formulation. For modality $m \in \{\mathrm{I}, \mathrm{V}, \mathrm{T}\}$, denoting image, video, and text, respectively, let $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ be an EEG–target pair. Here, $\boldsymbol{x}_i$ denotes the EEG input after dataset-specific preprocessing, and $\boldsymbol{y}_i$ is the associated image, video, or annotated text span. The pairing is determined by the dataset's acquisition paradigm and annotations. At test time, reconstruction receives the EEG input without its paired target. Retrieval additionally has access to a candidate test set, whereas the identity of the correct candidate is withheld. Reference targets are used for scoring.

Tasks. Retrieval ranks a test set $\mathcal{G} = \{\boldsymbol{y}^{(j)}\}_{j=1}^{N_\mathrm{c}}$ of N_c candidate targets according to their correspondence with an EEG input. Reconstruction generates an output $\hat{\boldsymbol{y}}_i$ conditioned on $\boldsymbol{x}_i$: an RGB image, a sequence of video frames, or a text sequence. Retrieval therefore evaluates candidate selection, whereas reconstruction evaluates the generated content against the paired target.

Overview. All retrieval baselines use a common alignment and ranking protocol. Reconstruction is evaluated in two settings: *Original* retains each method's reconstruction system, while *Unified* compares EEG encoders within a shared modality-specific pipeline. We additionally evaluate a Gaussian-noise input baseline, replacing EEG inputs with same-shaped tensors whose entries are sampled independently from $\mathcal{N}(0, 1)$.

3.2 Datasets and Processing

Dataset selection. The EEG2MM inventory comprises 11 image, video, and text source datasets, with their sources and task assignments summarized in Table A.1.

Sample construction and preprocessing. Each sample pairs an EEG time window with an image, a video segment, or an annotated text span. All task-specific models evaluated on the same dataset use identical preprocessing as detailed in Table C.2. For foundation models, preprocessing follows the pretraining configuration of the corresponding checkpoint. We apply the required channel or montage mapping, resampling, filtering, normalization, and input-format adaptations to match that configuration, while preserving EEG–target correspondence and data splits.

Data splits and statistics. We use the official splitting strategy for datasets with predefined train/validation/test splits and a common splitting policy for datasets without an official split. The detailed splitting strategy is summarized in Table C.1, and we provide a statistical summary of all datasets in Table A.1.

3.3 Evaluation Pipelines

Retrieval. We adopt the EEG-to-target retrieval formulation of NeuralBench (Banville et al., 2026). The frozen, modality-specific target encoders $F^{(m)}$ are DINOv2-Giant (Oquab et al., 2024) for images, V-JEPA2 (Assran et al., 2025) for videos, and GPT-2 (Radford et al., 2019) for text. The EEG

encoder $E_{\boldsymbol{\theta}}$ and output projection $P_{\boldsymbol{\psi}}$ produce $\hat{\boldsymbol{z}}_i = P_{\boldsymbol{\psi}}(E_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$, while $\boldsymbol{z}_i = F^{(m)}(\boldsymbol{y}_i)$ is the fixed target embedding. For task-specific networks, the output projection is implemented by their adapted output layer; for foundation models, a linear projection is appended. We optimize an EEG-to-target contrastive objective, following the alignment principle of CLIP (Radford et al., 2021):

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s(\hat{\boldsymbol{z}}_i, \boldsymbol{z}_i)/\tau)}{\sum_{j=1}^{B} \exp(s(\hat{\boldsymbol{z}}_i, \boldsymbol{z}_j)/\tau)}, \tag{1}$$

where B is the training batch size, $s(\cdot,\cdot)$ is cosine similarity, and $\tau > 0$ is the temperature. This is the EEG-to-target direction of the contrastive loss; gradients update the EEG encoder and projection, while $F^{(m)}$ remains frozen.

At evaluation, candidates are ranked by $s(\hat{\boldsymbol{z}}_i, F^{(m)}(\boldsymbol{y}^{(j)}))$ over the full test set, which contains one entry per distinct target ID. Test set size and target-identity rules are fixed across model comparisons.

Reconstruction. The *Original* setting retains each selected method's encoder architecture, adaptation modules, generative components, and training stages, subject to the benchmark's EEG initialization policy, data splits, and evaluation metrics. The *Unified* setting places an EEG encoder from a foundation model or a task-specific reconstruction method into a common pipeline within each modality. *Original* results compare complete systems; *Unified* results compare encoders while controlling downstream design. Encoder initialization and trainable parameter scopes are recorded separately in Section 4.1.

Each unified pipeline comprises a replaceable EEG encoder, a linear adapter $A_{\boldsymbol{\phi}}$, and a modality-specific pretrained generator $G^{(m)}$:

$$\boldsymbol{h}_i = E_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \qquad \boldsymbol{c}_i = A_{\boldsymbol{\phi}}(\boldsymbol{h}_i), \qquad \hat{\boldsymbol{y}}_i = G^{(m)}(\boldsymbol{c}_i). \tag{2}$$

Here, $\boldsymbol{h}_i$ is the EEG representation and $\boldsymbol{c}_i$ is the generation condition. We omit the generator's internal sampling randomness from the notation for clarity. Parameters and representation shapes are modality-specific. Pretrained target encoders and generative components remain frozen, with the same checkpoints used across EEG encoder comparisons within each dataset–task configuration. During training, paired targets provide supervision. Figure 3 summarizes the retrieval and reconstruction pipelines; the three reconstruction panels instantiate this common architecture.

For reconstruction, let $\boldsymbol{u}_i^{(m)}$ denote the fixed supervision target in the generator's conditioning space. The corresponding embedding regression objective is

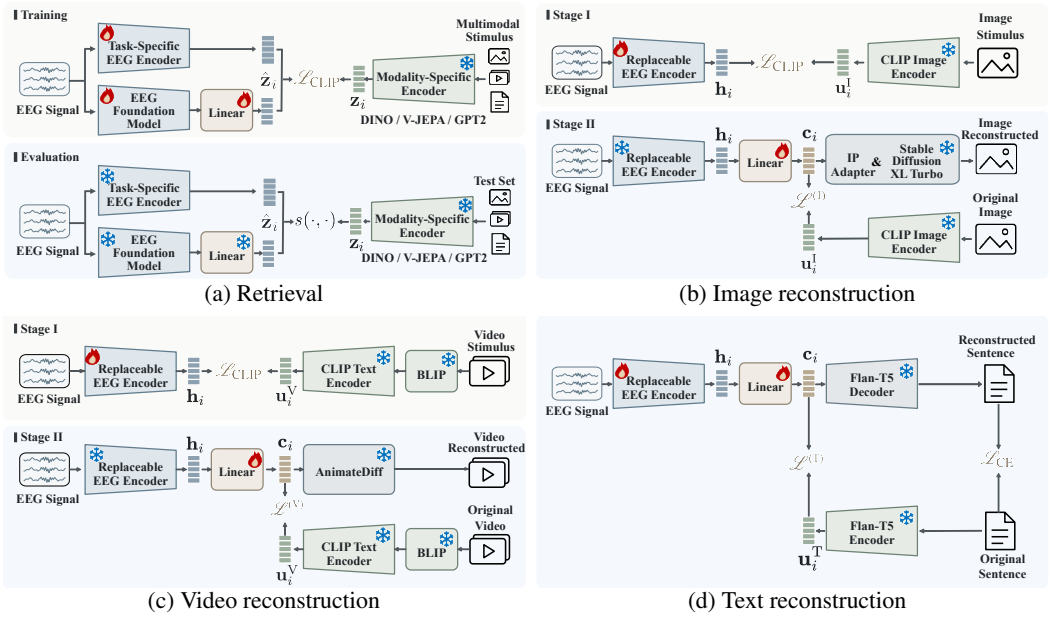
$$\mathcal{L}^{(m)} = \frac{1}{B}\sum_{i=1}^{B} \left\| \boldsymbol{c}_i - \boldsymbol{u}_i^{(m)} \right\|_2^2, \; m \in \{\text{I}, \text{V}, \text{T}\}. \tag{3}$$

The shapes of adapter outputs and supervision targets match. The modality-specific conditions and training procedures are:

Image reconstruction. The condition $\boldsymbol{c}_i$ predicts the paired image's embedding $\boldsymbol{u}_i^{(\text{I})}$ from a frozen CLIP image encoder (Radford et al., 2021). It is passed through frozen IP-Adapter (Ye et al., 2023) to condition Stable Diffusion XL Turbo (Sauer et al., 2024), as shown in Figure 3b. Stage I trains the EEG encoder with Eq. (1), setting $\hat{\boldsymbol{z}}_i = \boldsymbol{h}_i$ and $\boldsymbol{z}_i = \boldsymbol{u}_i^{(\text{I})}$. Stage II freezes the EEG encoder and trains only the linear adapter with $\mathcal{L}^{(\text{I})}$.

Video reconstruction. We use a captioning procedure based on frozen BLIP (Li et al., 2022) image captioning to derive a text description of the paired video. A frozen CLIP text encoder (Radford et al., 2021) maps this description to $\boldsymbol{u}_i^{(\text{V})}$. The EEG-derived condition $\boldsymbol{c}_i$ predicts this text representation and conditions the frozen AnimateDiff generator (GUO et al., 2024) to produce a sequence of frames (Figure 3c). Stage I trains the EEG encoder with Eq. (1), setting $\hat{\boldsymbol{z}}_i = \boldsymbol{h}_i$ and $\boldsymbol{z}_i = \boldsymbol{u}_i^{(\text{V})}$. Stage II freezes the encoder and trains only the linear adapter with $\mathcal{L}^{(\text{V})}$. This objective supervises caption-level content; temporal coherence relies on the pretrained motion prior, without explicit frame-wise supervision of the paired video's dynamics.

Text reconstruction. The condition $\boldsymbol{c}_i$ is aligned with the reference text representation $\boldsymbol{u}_i^{(\text{T})}$ from a frozen Flan-T5 encoder (Chung et al., 2024) and supplies cross-attention memory to the frozen Flan-T5 decoder (Figure 3d). We jointly train the EEG encoder and linear adapter using $\mathcal{L}_{\text{CE}} + \lambda_{\text{T}}\mathcal{L}^{(\text{T})}$,



(a) Retrieval (b) Image reconstruction

(c) Video reconstruction (d) Text reconstruction

Figure 3: **Unified evaluation pipelines**. We train the EEG encoder (and a linear layer) to align with the embedding of the paired stimulus under a modality-specific encoder for retrieval. Reconstruction uses a replaceable EEG encoder, a linear adapter, and a frozen generator. Image and video pipelines first train the EEG encoder to predict target embeddings, then freeze it and train the adapter. The text pipeline jointly trains the EEG encoder and adapter. Pretrained target encoders and generative components remain frozen.

where $\mathcal{L}_{\text{CE}} = -\frac{1}{\sum_{i=1}^{B} N_i} \sum_{i=1}^{B} \sum_{t=1}^{N_i} \log p_{\text{T5}}(r_{i,t} \mid r_{i,<t}, \boldsymbol{c}_i)$ is the next-token prediction loss of language models, $r_{i,t}$ is reference token t, $w_{i,<t}$ is its reference prefix, N_i is the number of reference tokens excluding padding, p_{T5} is the frozen decoder's conditional token distribution, and $\lambda_{\text{T}} \geq 0$ weights embedding regression. Teacher forcing is used for the training loss only. At test time, the decoder generates autoregressively from $\boldsymbol{c}_i$ and its own previously generated tokens.

Gaussian-noise input baseline. For reconstruction, we replace each paired EEG input at evaluation with a synthetic tensor of the same shape, whose entries are sampled independently from the standard normal distribution $\mathcal{N}(0, 1)$, while keeping the trained model and remaining inference pipeline unchanged. Following the noise-based analysis of Jo et al. (2026), this comparison measures how reconstruction performance changes when recorded EEG is replaced by Gaussian noise. Comparable scores suggest limited measured benefit from recorded EEG over the Gaussian-noise baseline under the tested protocol. A decrease indicates sensitivity to the replacement, but does not by itself establish use of stimulus-specific information, because the replacement also changes the input distribution.

3.4 Evaluation Metrics

Retrieval metrics. Top@1 and Top@5 report the fraction of EEG queries for which a correct target appears among the first one or five ranked candidates, respectively. Each result is accompanied by the full test set size M.

Reconstruction metrics. Image reconstruction is evaluated with PixCorr and SSIM (Wang et al., 2004) for pixel-level and structural correspondence, two-way identification using AlexNet (Krizhevsky et al., 2012) layers 2 and 5, Inception (Szegedy et al., 2016), and CLIP (Radford et al., 2021) features for feature-level correspondence, and SwAV (Caron et al., 2020) correlation distance as an additional representation-level measure. Video reconstruction uses mean frame-wise SSIM and a CLIP-based classifier (Radford et al., 2021) with an ImageNet (Deng et al., 2009) label vocabulary for frame-level N-way Top-K accuracy, and VideoMAE (Tong et al., 2022) classification accuracy on Kinetics-400 (Kay et al., 2017) for clip-level semantic classification. Text reconstruction

uses BLEU-1 through BLEU-4 (Papineni et al., 2002) and ROUGE-1 (Lin, 2004) precision, recall, and F1 to compare generated text with its paired reference. The evaluation metrics are detailed in Appendix B.

4 EXPERIMENTS

We evaluate EEG-to-image, EEG-to-video, and EEG-to-text retrieval and reconstruction using the source inventory in Appendix A; raw results are listed in Appendix D. Within each dataset–task configuration, compared models use the same samples, split, target identities, and scoring protocol.

4.1 EXPERIMENTAL SETUP

Dataset-specific preprocessing is detailed in Section 3.2. Retrieval uses the full test set of distinct stimulus IDs. Repeated EEG trials remain separate queries while sharing their target identity.

The comparison set includes LaBraM (Jiang et al., 2024), REVE (El Ouahidi et al., 2025), and LUNA (Döner et al., 2025) as foundation EEG encoders. Retrieval also includes ShallowFBCSP-Net and Deep4Net (Schirrmeister et al., 2017), EEGNet (Lawhern et al., 2018), BDTCN (Gemein et al., 2020), ATCNet (Altaheri et al., 2023), EEGConformer (Song et al., 2023), SimpleConvTimeAgg (El Ouahidi et al., 2023), and CTNet (Zhao et al., 2024). For reconstruction, ATM (Li et al., 2024), ENIGMA (Kneeland et al., 2026), and AVDE (Dai et al., 2026) are evaluated for images; EEG2Video (Liu et al., 2024), EEGVid (Singh et al., 2026), and MindCine (Zhou et al., 2026) for videos; and EEG2Text (Wang & Ji, 2022a), RAG-based EEG-to-Text (R-EEG2Text) (Collautti et al., 2026), and GLIM (Liu et al., 2025) for text. The three foundation encoders enter retrieval and the *Unified* reconstruction pipelines. Besides AVDE, the task-specific reconstruction methods are evaluated both as complete *Original* systems and through their EEG encoders in the corresponding *Unified* pipeline. AVDE is only evaluated as complete *Original* systems, since it leverages EEG foundation model as the encoder.

Foundation EEG encoders, including those used in task-specific reconstruction methods, are initialized from their released checkpoints, whereas all task-specific EEG models are randomly initialized. Both model groups follow the same training recipe within each dataset–task configuration for retrieval and *Unified* reconstruction, using the pipelines defined in Section 3.3. The same EEG encoder initialization policy applies to task-specific models in the *Original* setting. Their remaining components and training stages follow the method-specific procedures described there. The detailed training recipes are summarized in Table C.3.

Noisy Baselines. For the noisy reconstruction control, each EEG input is replaced at evaluation by a tensor of the same shape with every element sampled independently from $\mathcal{N}(0, 1)$. The trained model and remaining inference pipeline are unchanged, as detailed in Section 3.3.

4.2 COMPARISONS AND RESULTS

Retrieval is strongly dataset dependent (Figure 4; full values are in Table D.1). LaBraM is the strongest foundation encoder (21.80/40.60%), while SimpleConvTimeAgg is the strongest task-specific baseline (19.66/38.44%). LaBraM ranges from 60.12/89.50% on Things EEG to 0.08/0.39% on NOD-EEG. REVE leads video retrieval (9.29/29.72%), whereas LUNA leads text retrieval (3.51/13.49%), showing that transfer gains depend on dataset and modality.

Reconstruction results reveal complementary strengths (Figure 5). ATM has the highest mean image PixCorr/SSIM (0.287/0.247), while ENIGMA has the strongest semantic scores (AlexNet-2/5 = 0.645/0.676; CLIP = 0.625). EEGVid gives the best clean video performance (Video-2-way/40-way = 0.847/0.278; Frame-2-way/40-way = 0.823/0.289; SSIM = 0.370). For text, R-EEG2Text achieves the highest average N-way identification scores among Original systems evaluated on ZuCo 1 and ZuCo 2; GLIM-Enc. leads BLEU-1/2 (0.230/0.077), and LUNA leads Dist-1/Dist-2 diversity.

For image and text settings with paired controls, Gaussian-noise replacement can reduce semantic identification scores toward chance (e.g., GLIM-Enc. 2-way 0.602→0.501 and ENIGMA CLIP

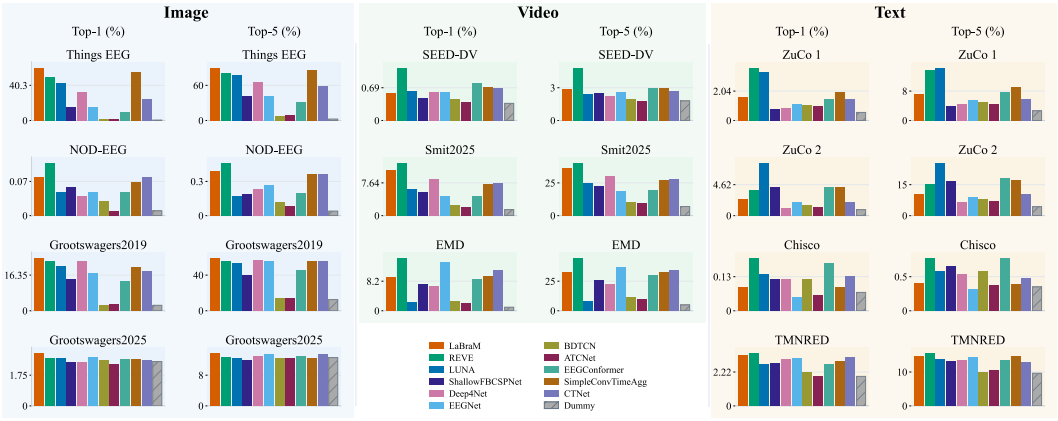


Figure 4: **Retrieval performance across datasets.** Top@1 and Top@5 accuracy are reported for each dataset; higher values indicate better performance.

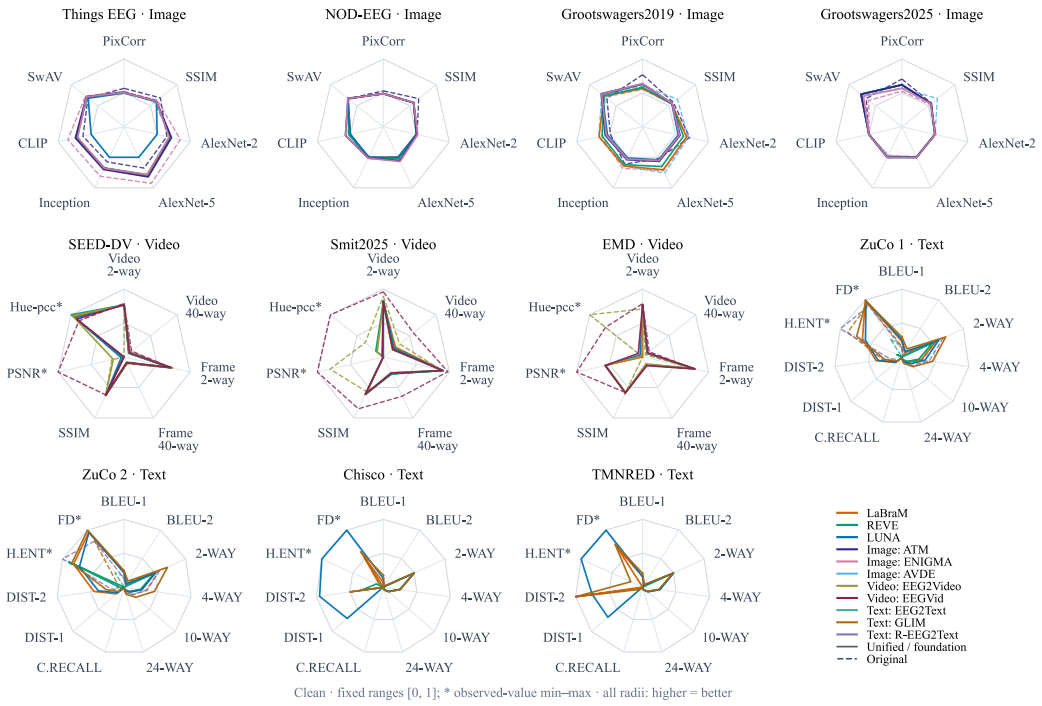


Figure 5: **Reconstrauction performance across datasets.** Metrics without * are normalized to $[0, 1]$ using theoretical bounds, while metrics marked with * are min–max normalized over observed values. All metrics are oriented such that higher is better.

0.625→0.483; Figure 6). Pixel-level scores can remain stable or increase, showing that these metrics do not consistently favor recorded EEG over Gaussian-noise inputs. Clean–noisy gaps characterize sensitivity to input replacement, but do not by themselves establish the use of stimulus-specific information.

5 Discussion

Do pretrained EEG representations transfer effectively to multimodal decoding? A foundation model achieves the highest observed retrieval Top-1 on all datasets, although the leading model varies. This suggests that the gap between pretraining objectives and multimodal decoding does not

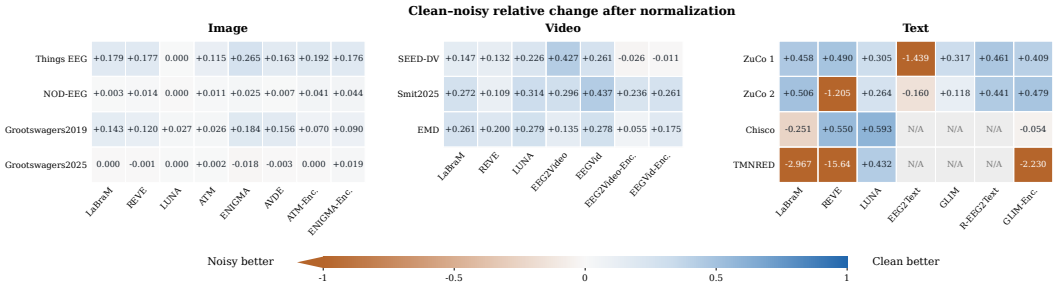


Figure 6: **Clean and noisy reconstruction performance.** Each cell reports the average relative performance difference across normalized metrics. Positive values (blue) favor recorded EEG, while negative values (orange) favor Gaussian-noise inputs. Small or negative gains for several method–dataset pairs suggest that recorded EEG does not consistently improve reconstruction performance under the evaluated protocol. N/A indicates that the corresponding method does not support Chinese-language datasets.

preclude useful adaptation for stimulus matching. However, this advantage does not consistently extend to reconstruction: under the Unified text pipeline on ZuCo 2, GLIM-Enc. achieves a 2-WAY score of 0.70, whereas all three foundation models remain near the dummy baseline of 0.50. Overall, pretrained EEG models are competitive for multimodal decoding, but their relative advantages remain task dependent.

Does strong retrieval performance translate into effective reconstruction? Our results show partial consistency between retrieval and reconstruction, but also demonstrate that agreement in relative ranking does not guarantee successful content recovery. On Things EEG, the ordering LaBraM $>$ REVE $>$ LUNA is preserved from retrieval Top 1 to reconstruction CLIP two-way identification. However, LUNA achieves a retrieval Top 1 while its reconstruction CLIP score is essentially matching the baseline. Thus, a model can retain its relative position across tasks without retaining a meaningful advantage over the reconstruction baseline. Overall, retrieval captures an important aspect of decoding capability, but cannot replace direct evaluation of reconstruction.

Do stronger reconstruction systems necessarily contain better EEG encoders? Our results show that stronger complete-system performance does not necessarily imply a better EEG encoder. On Things EEG, ENIGMA outperforms ATM in CLIP two-way identification under their Original pipelines (0.85 vs. 0.62), but this ordering reverses under the Unified pipeline (0.70 for ENIGMA-Enc. vs. 0.74 for ATM-Enc.). This reversal suggests that system-level advantages depend partly on downstream design rather than encoder quality alone. Reporting both settings therefore helps distinguish complete-system performance from encoder utility under shared downstream conditions.

6 Conclusion

We presented EEG2MM-Bench, a unified benchmark covering image, video, and text retrieval and reconstruction across 11 existing datasets and 468 hours of EEG recordings. The benchmark evaluates 17 task-specific architectures and three EEG foundation models under consistent protocols. Retrieval and reconstruction provide complementary assessments of stimulus matching and content recovery. Original-system evaluations characterize complete decoding systems, while unified reconstruction pipelines compare EEG encoders under controlled downstream conditions. Gaussian-noise input controls additionally characterize reconstruction sensitivity to EEG replacement. Our evaluation shows that EEG foundation models are competitive in retrieval but do not consistently outperform task-specific encoders in reconstruction. Together, these components support systematic comparisons of pretrained and task-specific EEG representations, highlighting the importance of task coverage, downstream design, and input sensitivity in evaluating multimodal decoding. Future work could extend evaluation to broader cross-subject and cross-dataset settings and explore pretraining strategies better aligned with multimodal content decoding. These directions would help connect advances in EEG representation learning with more generalizable retrieval and reconstruction capabilities.

AI USE STATEMENT

In this work, we used large language model tools only to assist with implementing the experimental methods in code and to polish the manuscript for readability. We did not use generative AI tools for any other required-disclosure task. All AI-assisted code and manuscript content were manually reviewed and verified by the authors, including checks of the code's correctness. The authors take full responsibility for the final content, claims, and artifacts of this paper.

ETHICS STATEMENT

This work uses only publicly available EEG datasets, and all dataset sources and associated publications are cited in Appendix A. The study is a computational benchmark: we did not recruit participants, collect new human data, conduct interventions, or interact with human participants. We followed the access conditions and usage requirements of the original dataset providers.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the benchmark throughout the paper and appendices. Appendix A records the dataset sources. Section 3.2 describes EEG–stimulus sample construction, the preprocessing procedures, and the split policy used to avoid leakage. Section 3.3 defines the common retrieval protocol, target encoders, *Original* and *Unified* reconstruction pipelines, trainable and frozen components, and modality-specific training objectives. Section 4.1 records the model inventory, initialization rules, training recipes, and controls. Together, these descriptions make the data sources, preprocessing, model comparisons, training procedures, and controls reproducible.

ACKNOWLEDGMENTS

Use unnumbered third-level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. *IEEE Transactions on Industrial Informatics*, 19(2):2249–2258, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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.

Hubert Banville, Stéphane d'Ascoli, Simon Dahan, Jérémy Rapin, Marlène Careil, Yohann Benchetrit, Jarod Lévy, Saarang Panchavati, Antoine Ratouchniak, Elisa Cascardi, et al. NeuralBench: A unifying framework to benchmark neuroAI models. *arXiv preprint arXiv:2605.08495*, 2026.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: Toward real-time reconstruction of visual perception. In *ICLR*, 2024.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

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

Enrico Collautti, Xiaopeng Mao, Luca Tonin, Stefano Tortora, and Sadasivan Puthusserypady. RAG-based EEG-to-text translation using deep learning and LLMs. *arXiv preprint arXiv:2605.17503*, 2026.

Sicheng Dai, Hongwang Xiao, Shan Yu, and Qiwei Ye. Autoregressive visual decoding from EEG signals. In *ICLR*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: efficient and topology-agnostic foundation model for EEG signal analysis. In *NeurIPS*, 2025.

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 *NeurIPS*, 2023.

Yassine El Ouahidi, Vincent Gripon, Bastien Pasdeloup, Ghaith Bouallegue, Nicolas Farrugia, and Giulia Lioi. A strong and simple deep learning baseline for BCI MI decoding. *arXiv preprint arXiv:2309.07159*, 2023.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG-adapting to any setup with large-scale pretraining on 25,000 subjects. In *NeurIPS*, 2025.

Lawrence A. Farwell and Emanuel Donchin. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, 1988.

Honghao Fu, Zhiqi Shen, Jing Jih Chin, and Hao Wang. BrainVis: Exploring the bridge between brain and visual signals via image reconstruction. In *ICASSP*, 2025.

Jianxiong Gao, Yichang Liu, Baofeng Yang, Jianfeng Feng, and Yanwei Fu. CineBrain: A large-scale multi-modal brain dataset during naturalistic audiovisual narrative processing. *arXiv preprint arXiv:2503.06940*, 2025.

Lukas A. W. Gemein, Robin T. Schirrmeister, Patryk Chrabąszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of EEG pathology. *NeuroImage*, 220:117021, 2020.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Alessandro T Gifford, Pablo Oyarzo, Anne W Zonneveld, Christina Sartzetaki, Iris IA Groen, and Radoslaw M Cichy. A large dataset of human EEG responses to short naturalistic videos for studying dynamic visual event processing. *arXiv preprint arXiv:2608.28768*, 2026.

Tijl Grootswagers, Amanda K. Robinson, and Thomas A. Carlson. The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188:668–679, 2019.

Tijl Grootswagers, Genevieve L. Quek, Zhen Zeng, and Manuel Varlet. Human infant EEG recordings for 200 object images presented in rapid visual streams. *Scientific Data*, 12:407, 2025.

Yuwei GUO, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo DAI. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

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.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *LREC*, 2020.

Vinay Jayaram and Alexandre Barachant. MOABB: trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *ICLR*, 2024.

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.

Ard Kastrati, Josua B"urki, Jonas Lauer, Cheng Xuan, Raffaele Iaquinto, and Roger Wattenhofer. EEG-Bench: A benchmark for EEG foundation models in clinical applications. In *NeurIPS Workshop*, 2025.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and Mustafa Suleyman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Reese Kneeland, Wangshu Jiang, Ugo Bruzadin Nunes, Paul Steven Scotti, Arnaud Delorme, and Jonathan Xu. ENIGMA: EEG-to-image in 15 minutes using less than 1% of the parameters. *arXiv preprint arXiv:2602.10361*, 2026.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

Yu-Ting Lan, Kan Ren, Yansen Wang, Wei-Long Zheng, Dongsheng Li, Bao-Liang Lu, and Lili Qiu. Seeing through the brain: Image reconstruction of visual perception from human brain signals. *arXiv preprint arXiv:2308.02510*, 2023.

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.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *NeurIPS*, 2024.

Jilong Li, Zhenxi Song, Jiaqi Wang, Meishan Zhang, Honghai Liu, Min Zhang, and Zhiguo Zhang. BrainECHO: Semantic brain signal decoding through vector-quantized spectrogram reconstruction for whisper-enhanced text generation. In *Findings of ACL*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL Workshop*, 2004.

Junxiang Liu, Junming Lin, Jie Zhou, Wei Xiong, Jiangtong Li, Jie Li, Jie Zhuang, and Hongfei Ji. DynaMind: Reconstructing dynamic visual scenes from EEG by aligning temporal dynamics and multimodal semantics to guided diffusion. In *Findings of CVPR*, 2026.

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

Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. EEG2Video: Towards decoding dynamic visual perception from EEG signals. In *NeurIPS*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.

Gert Pfurtscheller and Fernando H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, 1999.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Herbert Ramoser, Johannes Müller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8 (4):441–446, 2000.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, 2024.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

Shengle Shi, Yinglu Song, Yong Wang, Jiaqing Xiao, Xinpeng Lin, Heng Wang, Zhibin Zhao, Pengyu Wang, Zitong Lu, Xiaoli Li, and He Chen. Toward general-purpose foundation models for electroencephalography: A unified data registry. *Brain-X*, 4(1):e70046, 2026.

Prajwal Singh, Anupam Sharma, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. Beyond reconstruction: What EEG-to-video decoding actually recovers. *arXiv preprint arXiv:2505.21385v4*, 2026.

Sophie Smit, Almudena Ramirez-Haro, Manuel Varlet, Denise Moerel, Genevieve L. Quek, and Tijl Grootswagers. A comprehensive EEG dataset for investigating visual touch perception. *Scientific Data*, 13:381, 2026.

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.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *ICLR*, 2024.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

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

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

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 *NeurIPS*, 2024a.

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 *ACL*, 2024b.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *ICLR*, 2025.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *AAAI*, 2022a.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *AAAI*, 2022b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Xiaoxi Wei, A. Aldo Faisal, Moritz Grosse-Wentrup, Alexandre Gramfort, Sylvain Chevallier, Vinay Jayaram, Camille Jeunet, Stylianos Bakas, Siegfried Ludwig, Konstantinos Barmpas, et al. 2021 BEETL competition: Advancing transfer learning for subject independence and heterogenous EEG data sets. In *NeurIPS Competitions and Demonstrations Track*, pp. 205–219, 2022.

Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

Jiamin Wu, Zichen Ren, Junyu Wang, Pengyu Zhu, Yonghao Song, Mianxin Liu, Qihao Zheng, Lei Bai, Wanli Ouyang, and Chunfeng Song. AdaBrain-Bench: Benchmarking brain foundation models for brain-computer interface applications. *arXiv preprint arXiv:2507.09882*, 2025.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-Bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of EEG foundation models. *arXiv preprint arXiv:2508.17742v3*, 2026.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Zesheng Ye, Lina Yao, Yu Zhang, and Sylvia Gustin. See what you see: Self-supervised cross-modal retrieval of visual stimuli from brain activity. *arXiv preprint arXiv:2208.03666*, 2022.

Guohao Zhang, Ming Zhou, Shuyi Zhen, Shaohua Tang, Zheng Li, and Zonglei Zhen. A large-scale MEG and EEG dataset for object recognition in naturalistic scenes. *Scientific Data*, 12:857, 2025.

Kaifan Zhang, Lihuo He, Yuqi Ji, Junjie Ke, Lukun Wu, Tianhao You, and Xinbo Gao. EEG-EditBench: Probing visual information in EEG-image retrieval models with controlled image edits. *arXiv preprint arXiv:2607.27857*, 2026a.

Zihan Zhang, Xiao Ding, Yu Bao, Yi Zhao, Xia Liang, Bing Qin, and Ting Liu. Chisco: An EEG-based BCI dataset for decoding of imagined speech. *Scientific Data*, 11:1265, 2024.

Zihan Zhang, Yu Bao, Xiao Ding, Tianyi Jiang, and Kai Xiong. Is EEG-to-text feasible in real-world scenarios? an in-depth analysis using a neuropsychology-inspired benchmark. In *ACL*, 2026b.

Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. CTNet: A convolutional transformer network for EEG-based motor imagery classification. *Scientific Reports*, 14(1):20237, 2024.

Jinzhao Zhou, Yiqun Duan, Ziyi Zhao, Yu-Cheng Chang, Yu-Kai Wang, Thomas Do, and Chin-Teng Lin. Towards linguistic neural representation learning and sentence retrieval from electroencephalogram recordings. In *BCIMM Workshop*, 2024.

Tian-Yi Zhou, Xuan-Hao Liu, Bao-Liang Lu, and Wei-Long Zheng. MindCine: Multimodal EEG-to-video reconstruction with large-scale pretrained models. In *ICASSP*, 2026.

Table A.1: **Datasets included in the benchmark, their task assignments, and dataset statistics.** Recon. denotes reconstruction, and Retr. denotes retrieval. Channels describe the reported acquisition setup; subjects count EEG contributors in the cited studies.

Modality	Dataset	Task	Channels	Subjects	Hours
Image	Things EEG (Gifford et al., 2022)	Recon., Retr.	63	10	79.91
	NOD-EEG (Zhang et al., 2025)		62	80	23.70
	Grootswagers2019 (Grootswagers et al., 2019)		63	16	12.44
	Grootswagers2025 (Grootswagers et al., 2025)		32	42	5.52
Video	SEED-DV (Liu et al., 2024)	Recon., Retr.	62	20	21.23
	Smit2025 (Smit et al., 2026)		64	80	80.47
	EMD (Gifford et al., 2026)		127	6	89.6
Text	ZuCo 1 (Hollenstein et al., 2018)	Recon., Retr.	105	12	22.16
	ZuCo 2 (Hollenstein et al., 2020)		105	18	21.05
	Chisco (Zhang et al., 2024)		122	5	103.50
	TMNRED (Bai et al., 2025)		30	30	8.33

A Dataset Sources and Statistics

Table A.1 summarizes the source references, supported tasks, and statistics of all selected datasets, including the numbers of EEG channels and participants and the recording duration in hours.

B Evaluation Metrics

This appendix gives the scoring definitions used by the benchmark. Higher values are better except for SwAV correlation distance.

B.1 Stimulus retrieval

For EEG query i, let $\hat{\boldsymbol{z}}_i$ be its predicted embedding and let $\boldsymbol{z}_j$ be the frozen target embedding for gallery stimulus $j \in \mathcal{G}$. Candidates are ranked by cosine similarity with deterministic stimulus-ID tie breaking. If $g(i)$ is the paired stimulus ID,

$$\text{Top@}K = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[g(i) \in \text{TopK}_{j\in\mathcal{G}} s(\hat{\boldsymbol{z}}_i, \boldsymbol{z}_j)\right], \qquad K \in \{1, 5\}. \tag{4}$$

Repeated trials remain separate queries but share a gallery ID. The full test gallery and its size are reported; Top@5 is omitted when fewer than five distinct stimuli are available (Banville et al., 2026).

B.2 Image reconstruction

Let I_i and $\hat{I}_i$ be the paired reference and reconstruction after the same resizing and intensity normalization. PixCorr is the Pearson correlation between their flattened RGB values; constant images are invalid. SSIM compares local luminance, contrast, and structure (Wang et al., 2004):

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{5}$$

where $C_1 = (0.01L)^2$, $C_2 = (0.03L)^2$, and L is the fixed intensity range. Window scores are averaged spatially and then over channels.

For two-way identification, each reference is paired with one fixed distractor from the same test set. With a frozen feature extractor f,

$$2\text{way}_f = \frac{1}{N}\sum_{i=1}^{N} q\Big(\rho(f(\hat{I}_i), f(I_i)), \rho(f(\hat{I}_i), f(I_{j(i)}))\Big), \tag{6}$$

where $q(a, b) = 1$ for $a > b$, $1/2$ for a tie, and 0 otherwise. We compute this separately for AlexNet layers 2 and 5 (Krizhevsky et al., 2012), Inception (Szegedy et al., 2016), and CLIP features (Radford et al., 2021); chance is 50%. SwAV correlation distance is $N^{-1}\sum_i[1 - \rho(f_{\text{SwAV}}(\hat{I}_i), f_{\text{SwAV}}(I_i))]$, for which lower is better (Caron et al., 2020).

B.3 Video reconstruction

Reference and reconstructed videos are sampled at the same timestamps and resized identically. Frame SSIM averages valid frame scores within each video and then across videos. Frame-level N-way Top-K uses a frozen CLIP-based ImageNet classifier (Deng et al., 2009); video-level Top-K uses a frozen VideoMAE classifier (Tong et al., 2022) trained on Kinetics-400 (Kay et al., 2017). We report N, K, the class vocabulary and mapping, timestamps, and checkpoints. A classifier score is marked N/A when target classes cannot be mapped unambiguously and is not treated as an incorrect reconstruction.

B.4 Text reconstruction

Generated and reference sentences use the same fixed dataset-specific tokenizer. For $n \in \{1, 2, 3, 4\}$, sentence-level BLEU-n uses clipped n-gram precision, uniform weights, and the brevity penalty (Papineni et al., 2002):

$$\text{BLEU-}n(h_i, r_i) = \text{BP}(h_i, r_i)\exp\left(\frac{1}{n}\sum_{k=1}^{n}\log p_k\right). \tag{7}$$

An empty hypothesis or any required zero precision receives zero (no smoothing). BLEU scores are averaged over sentences and reported on a 0–100 scale. For ROUGE-1 (Lin, 2004), clipped unigram overlap gives $P_i = m_i/|h_i|$, $R_i = m_i/|r_i|$, and $F_{1,i} = 2P_iR_i/(P_i + R_i)$, with zero for a vanishing denominator. Precision, recall, and F1 are averaged separately and reported on the same scale.

C Benchmark Setup

This appendix records the dataset split in Table C.1, EEG preprocessing in Table C.2, and unified training protocol used by the benchmark setup file in Table C.3.

Table C.1: **Dataset splits.** Test targets are disjoint from training targets unless noted.

Modality	Dataset	Split
Image	Things EEG	Train/test on 1,654/200 concepts; concept-disjoint. Average test trials over 80 repetitions; retain all training repetitions.
	NOD-EEG	Train/test on 800/200 classes; class-disjoint. Average test trials within class across sessions.
	Grootswager2019	Train/test on 180/20 stimuli; stimulus-disjoint. Average test stimuli over 80 repetitions (40 sequences × 2 speeds); retain all training repetitions.
	Grootswagers2025	Train/test on 180/20 stimuli; stimulus-disjoint. Average test stimuli across presentation blocks; retain all training repetitions.
Video	SEED-DV	Seven blocks × 200 clips (40 concepts × 5 repetitions). Train on blocks 1–6 (1,200 trials), test on block 7 (200); block-disjoint.
	Smit2025	90 videos, each shown once across 8 blocks. Train on 7 blocks (630 trials), test on 1 (90); block-disjoint.
	EMD	1,102 videos: IDs 1–1,000 train (6 repetitions), 1,001–1,102 test (24); stimulus-disjoint. Retain all training repetitions; use the first test presentation (102 trials).

Continued on next page

Table C.1: **Dataset splits.** Test targets are disjoint from training targets unless noted. (Continued)

Modality	Dataset	Split
Text	ZuCo 1	Sentence-disjoint across train/validation/test. Fixed-seed stratified split (about 8:1:1); all 12 subjects contribute.
	ZuCo 2	Same sentence-disjoint split; 12/18 subjects.
	Chisco	Same sentence-disjoint split; 2/5 subjects.
	TMNRED	Same sentence-disjoint split; 12/30 subjects.

Table C.2: **EEG preprocessing.** Learned-normalization statistics are fitted on the training split.

Modality	Dataset	Preprocessing
Image	Things EEG	0.1–100 Hz band-pass; keep 63 electrodes. Epoch 0–1000 ms and baseline-correct over 200 ms pre-stimulus. Downsample to 250 Hz. Fit multivariate noise normalization on train; apply to train/test.
	NOD-EEG	Use released cleaned continuous signal; no re-filtering. Drop HEO/VEO. Epoch $[-100, 800]$ ms and baseline-correct over 100 ms pre-stimulus. Fit multivariate noise normalization on train; apply to train/test.
	Grootswager2019	Keep the online Cz reference; no re-reference. Apply a 0.1–100 Hz Hamming-window FIR band-pass. Epoch $[-100, 1000]$ ms after the non-target and downsample to 250 Hz. Fit multivariate noise normalization on train; apply to train/test.
	Grootswagers2025	Apply a 0.5–40 Hz band-pass and average reference. Interpolate flat channels and log-variance outliers with spherical splines; reject the 2% largest peak-to-peak trials. Epoch $[-100, 800]$ ms, downsample to 250 Hz, and fit multivariate noise normalization on train.
Video	SEED-DV	Use released continuous signal; no re-reference. Keep 200 Hz and take 2 s from onset. Normalize in-model with train-fitted statistics.
	Smit2025	Apply a 0.1–100 Hz band-pass. Take 0.9 s from onset at 200 Hz, then zero-pad to 1 s. Normalize in-model with train-fitted statistics.
	EMD	Use upstream 0.1–40 Hz signal with Fz reference. Downsample to 200 Hz and take 2 s from onset. Normalize in-model with train-fitted statistics.
Text	ZuCo 1	Use released artifact-cleaned continuous signal. Extract sentence segments at 128 Hz; truncate/zero-pad to 1,280 points (10 s).
	ZuCo 2	Same preprocessing, acquisition, and cleaning pipeline as ZuCo 1.
	Chisco	Use released bad-channel rejection, interpolation, 1–250 Hz filtering, and ICA. Drop VEO/HEO/Trigger; retain 122 EEG channels. Downsample to 128 Hz with polyphase anti-aliasing; truncate/zero-pad to 1,280 points.
	TMNRED	Use upstream bad-channel rejection, average reference, 0.5–80 Hz band-pass, and ICA. Keep EEG channels only. Downsample to 128 Hz with polyphase anti-aliasing; truncate/zero-pad to 1,280 points.

Table C.3: **Unified training recipes.** Task-specific original systems are outside this table.

Modality	Stage	Recipe
Image	EEG encoder	Train ATM-S from scratch; initialize LaBraM/REVE/LUNA from official weights and fine-tune end to end. Regress to frozen CLIP ViT-H/14 embeddings with $0.9\,\mathrm{MSE} + 0.1\,\mathrm{InfoNCE}$, scaling the total loss by 10. AdamW, lr 3×10^{-4}, batch 64, 40 epochs.
	CLIP projection	Replace ATM's Diffusion Prior with $\mathrm{Linear}(1024, 1024)$ into frozen CLIP image space. Optimize MSE with Adam, lr 1×10^{-3}, batch 1024, 40 epochs; cosine decay with 500-step warmup.
	Image generation	Frozen SDXL-Turbo + IP-Adapter (ViT-H), conditioned on the predicted CLIP embedding. Use 4 sampling steps, guidance 0, IP-Adapter scale 1, and 1 image/test condition.
	Global setting	One global seed for splitting, initialization, data order, and generation noise. Retrieval reuses the encoder with a contrastive objective.
Video	EEG encoder	Replace only the EEG2Video encoder (default original; alternatives LaBraM/REVE/LUNA/EEGVid temporal encoders); share other stages. Train encoder + projection head end to end toward frozen CLIP text embeddings; output $[B, 77, 768]$. MSE with Adam, lr 5×10^{-4}, batch 32, 40 epochs, cosine annealing.
	Latent adapter	Freeze stage-1 parameters; train only a per-token $\mathrm{Linear}(768, 768)$ adapter. Adam, lr 5×10^{-4}, 40 epochs, 500-step warmup.
	Video generation	Frozen SD 1.5 + AnimateDiff v2, conditioned on $[B, 77, 768]$. Use 25 steps, guidance 7.5, generate 16 frames, return 6.
	Global setting	Train one parameter set per subject. Use one global seed for splitting, initialization, data order, and generation noise.

Continued on next page

Table C.3: **Unified training recipes.** Task-specific original systems are outside this table. (Continued)

Modality	Stage	Recipe
Text	EEG encoder	GLIM with domain prompts disabled. Default GLIM temporal encoder; full replacements LaBraM-base/LUNA-base/REVE-base. Initialize from official weights, fine-tune end to end, and adaptively pool to 96 time steps.
	EEG–text alignment	Cross-attention pooling compresses EEG and frozen text hidden states to single embeddings. Use symmetric in-batch contrastive alignment: mean cross-entropy over the cosine matrix in both directions.
	Text decoder	Frozen encoder–decoder LM: flan-t5-large (English), mt0-large (Chinese). Train with LM cross-entropy while freezing LM weights. Expand each trial to 8 rewritten variants; use at most 1 variant/sentence per batch.
	Global setting	Optimize equal-weight alignment + LM cross-entropy. Adam, lr 1×10^{-4}, zero weight decay, no schedule, effective batch 576, 200 epochs. Use one global seed for splitting, initialization, data order, and noise realization.

D EXPERIMENTAL RESULTS

This appendix contains the unaggregated result tables.

Table D.1: **Raw results for EEG-to-stimulus retrieval across all datasets (`FM`: foundation model).** Values are reported as clean/noisy.

Dataset	Baseline	Top1 ↑	Top5 ↑
Things EEG	LaBraM (FM)	60.12	89.50
	REVE (FM)	49.93	81.83
	LUNA (FM)	43.37	78.28
	ShallowFBCSPNet	15.68	41.82
	Deep4Net	33.20	66.33
	EEGNet	15.03	41.65
	BDTCN	1.60	7.77
	ATCNet	1.73	8.97
	EEGConformer	9.28	31.72
	SimpleConvTimeAgg	55.98	85.95
	CTNet	24.93	58.22
	Dummy	0.50	2.50
NOD-EEG	LaBraM (FM)	0.08	0.39
	REVE (FM)	0.11	0.46
	LUNA (FM)	0.05	0.17
	ShallowFBCSPNet	0.06	0.19
	Deep4Net	0.04	0.23
	EEGNet	0.05	0.27
	BDTCN	0.03	0.12
	ATCNet	0.01	0.08
	EEGConformer	0.05	0.20
	SimpleConvTimeAgg	0.07	0.36
	CTNet	0.08	0.36
	Dummy	0.01	0.04
Grootswagers2019	LaBraM (FM)	24.04	58.83
	REVE (FM)	22.73	56.09
	LUNA (FM)	20.31	53.70
	ShallowFBCSPNet	14.22	40.10
	Deep4Net	22.60	56.56
	EEGNet	17.24	55.23
	BDTCN	2.73	13.85
	ATCNet	2.79	14.30
	EEGConformer	13.75	45.83
	SimpleConvTimeAgg	19.95	55.13
	CTNet	17.94	55.89
	Dummy	2.50	12.50
Grootswagers2025	LaBraM (FM)	2.98	13.66
	REVE (FM)	2.68	12.76
	LUNA (FM)	2.68	12.50
	ShallowFBCSPNet	2.48	11.80
	Deep4Net	2.50	12.82
	EEGNet	2.74	13.34
	BDTCN	2.60	12.42
	ATCNet	2.36	12.36
	EEGConformer	2.64	12.82
	SimpleConvTimeAgg	2.62	12.34
	CTNet	2.60	13.52
	Dummy	2.52	12.58
SEED-DV	LaBraM (FM)	0.57	2.85
	REVE (FM)	1.10	4.80
	LUNA (FM)	0.61	2.46
	ShallowFBCSPNet	0.47	2.51
	Deep4Net	0.59	2.21
	EEGNet	0.60	2.57

Continued on next page

Table D.1: **Raw results for EEG-to-stimulus retrieval across all datasets (FM: foundation model).** Values are reported as clean/noisy. (Continued)

Dataset	Baseline	Top1 ↑	Top5 ↑
	BDTCN	0.46	1.99
	ATCNet	0.39	1.77
	EEGConformer	0.79	2.95
	SimpleConvTimeAgg	0.71	2.95
	CTNet	0.68	2.69
	Dummy	0.36	1.79
Smit2025	LaBraM (FM)	10.56	35.98
	REVE (FM)	12.23	40.03
	LUNA (FM)	6.25	24.62
	ShallowFBCSPNet	5.44	22.29
	Deep4Net	8.50	30.42
	EEGNet	4.55	18.57
	BDTCN	2.38	10.38
	ATCNet	2.08	9.62
	EEGConformer	4.61	19.76
	SimpleConvTimeAgg	7.44	27.03
	CTNet	7.62	27.83
	Dummy	1.39	6.94
EMD	LaBraM (FM)	9.26	32.35
	REVE (FM)	14.54	44.34
	LUNA (FM)	2.40	8.33
	ShallowFBCSPNet	7.24	25.49
	Deep4Net	6.75	22.06
	EEGNet	13.56	37.09
	BDTCN	2.72	11.49
	ATCNet	1.96	9.59
	EEGConformer	8.77	30.17
	SimpleConvTimeAgg	9.59	32.35
	CTNet	11.22	34.42
	Dummy	0.98	4.90
ZuCo 1	LaBraM (FM)	1.60	7.33
	REVE (FM)	3.64	13.77
	LUNA (FM)	3.34	14.27
	ShallowFBCSPNet	0.82	3.84
	Deep4Net	0.88	4.45
	EEGNet	1.16	5.62
	BDTCN	1.11	5.04
	ATCNet	1.00	4.41
	EEGConformer	1.52	7.81
	SimpleConvTimeAgg	1.98	9.21
	CTNet	1.46	5.90
	Dummy	0.47	2.31
ZuCo 2	LaBraM (FM)	2.42	10.55
	REVE (FM)	3.75	15.37
	LUNA (FM)	7.83	25.40
	ShallowFBCSPNet	4.25	16.90
	Deep4Net	1.06	6.53
	EEGNet	2.00	8.93
	BDTCN	1.60	8.19
	ATCNet	1.28	7.02
	EEGConformer	4.30	18.06
	SimpleConvTimeAgg	4.33	17.19
	CTNet	2.06	10.57
	Dummy	0.34	3.24
Chisco	LaBraM (FM)	0.09	0.40
	REVE (FM)	0.20	0.77
	LUNA (FM)	0.14	0.58
	ShallowFBCSPNet	0.12	0.65
	Deep4Net	0.12	0.53
	EEGNet	0.05	0.31

Continued on next page

Table D.1: **Raw results for EEG-to-stimulus retrieval across all datasets (`FM`: foundation model).** Values are reported as clean/noisy. (Continued)

Dataset	Baseline	Top1 ↑	Top5 ↑
	BDTCN	0.12	0.58
	ATCNet	0.06	0.38
	EEGConformer	0.18	0.77
	SimpleConvTimeAgg	0.09	0.39
	CTNet	0.13	0.47
	Dummy	0.09	0.42
TMNRED	LaBraM (FM)	3.30	14.63
	REVE (FM)	3.43	15.47
	LUNA (FM)	2.74	13.69
	ShallowFBCSPNet	2.81	13.02
	Deep4Net	3.06	13.45
	EEGNet	3.13	14.29
	BDTCN	2.16	10.00
	ATCNet	1.95	10.58
	EEGConformer	2.72	13.28
	SimpleConvTimeAgg	2.89	14.50
	CTNet	3.15	12.90
	Dummy	1.93	9.64

Table D.2: **Raw results for image reconstruction.** For task-specific methods, `Enc.` denotes the EEG encoder evaluated in the unified pipeline, while the unsuffixed method uses its original pipeline. Foundation models use the unified pipeline. Clean results are shown on the first row and noisy results on the following row; Dummy is reported once because it has no noisy result.

Dataset	Baseline	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
Things EEG	LaBraM	0.0166/	0.2106/	0.7069/	0.7953/	0.6769/	0.7254/	0.5899/
		-0.0206	0.2810	0.5027	0.5011	0.5079	0.5005	0.6650
	REVE	0.0106/	0.2114/	0.6934/	0.7850/	0.6740/	0.7164/	0.5998/
		-0.0130	0.2757	0.4934	0.4955	0.5057	0.4949	0.6738
	LUNA	-0.0045/	0.2223/	0.5001/	0.4999/	0.4993/	0.5001/	0.6441/
		-0.0046	0.2223	0.5000	0.4998	0.4988	0.4997	0.6447
	ATM	0.1363/	0.3712/	0.6482/	0.6761/	0.5753/	0.6236/	0.6339/
		0.1094	0.3737	0.5250	0.5120	0.4887	0.5078	0.6494
	ENIGMA	0.0404/	0.2603/	0.8515/	0.9265/	0.8101/	0.8529/	0.5414/
		-0.0021	0.2557	0.4896	0.5019	0.5089	0.4746	0.7048
	AVDE	0.0551/	0.1489/	0.7160/	0.8153/	0.6679/	0.7197/	0.6170/
		0.0329	0.2793	0.5214	0.5463	0.4749	0.4933	0.6589
	ATM-Enc.	0.0136/	0.2194/	0.7182/	0.8206/	0.6994/	0.7390/	0.5711/
		-0.0241	0.2974	0.4932	0.4977	0.5004	0.4984	0.6511
	ENIGMA-Enc.	0.0110/	0.2245/	0.6816/	0.7828/	0.6790/	0.7023/	0.5780/
		-0.0181	0.2844	0.4962	0.4943	0.4971	0.4998	0.6614
	Dummy	0.1065	0.3845	0.5000	0.5000	0.5000	0.5000	0.6480

Continued on next page

Table D.2: **Raw results for image reconstruction.** For task-specific methods, `Enc.` denotes the EEG encoder evaluated in the unified pipeline, while the unsuffixed method uses its original pipeline. Foundation models use the unified pipeline. Clean results are shown on the first row and noisy results on the following row; Dummy is reported once because it has no noisy result. (Continued)

Dataset	Baseline	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
NOD-EEG	LaBraM	-0.0222/	0.1445/	0.4988/	0.4992/	0.5013/	0.5028/	0.6580/
		-0.0237	0.1423	0.5005	0.4990	0.4986	0.4958	0.6596
	REVE	-0.0221/	0.1442/	0.5050/	0.5210/	0.5042/	0.5234/	0.6578/
		-0.0225	0.1427	0.5011	0.5031	0.4991	0.5009	0.6604
	LUNA	-0.0207/	0.1466/	0.4999/	0.4998/	0.4997/	0.5000/	0.6525/
		-0.0208	0.1466	0.4998	0.5004	0.5006	0.5000	0.6521
	ATM	0.0618/	0.3360/	0.5109/	0.4992/	0.5033/	0.4979/	0.6684/
		0.0030	0.3346	0.5033	0.5085	0.4867	0.4979	0.6596
	ENIGMA	-0.0035/	0.1561/	0.5241/	0.5630/	0.4857/	0.5307/	0.6914/
		0.0086	0.1534	0.5171	0.4992	0.4951	0.5015	0.7248
	AVDE	-0.0104/	0.1541/	0.5144/	0.5247/	0.4860/	0.5058/	0.6836/
		-0.0118	0.1657	0.4759	0.4891	0.5242	0.5039	0.6686
	ATM-Enc.	-0.0232/	0.1383/	0.5115/	0.5545/	0.5136/	0.5736/	0.6585/
		-0.0278	0.1442	0.5004	0.5026	0.5013	0.4965	0.6736
	ENIGMA-Enc.	-0.0206/	0.1461/	0.5189/	0.5692/	0.5146/	0.5756/	0.6604/
		-0.0037	0.1566	0.5047	0.5067	0.4983	0.5038	0.7132
	Dummy	0.0537	0.3484	0.5000	0.5000	0.5000	0.5000	0.6626
Grootswager2019	LaBraM	0.1306/	0.0979/	0.6653/	0.7046/	0.6421/	0.6604/	0.5342/
		0.1613	0.0939	0.4824	0.4911	0.5248	0.4982	0.5654
	REVE	0.1678/	0.0997/	0.6189/	0.6508/	0.6201/	0.6132/	0.5376/
		0.1573	0.0983	0.5002	0.4975	0.4947	0.5033	0.5646
	LUNA	0.2603/	0.1022/	0.5400/	0.5405/	0.5109/	0.5332/	0.4680/
		0.2603	0.1027	0.5135	0.5063	0.4967	0.5043	0.4568
	ATM	0.5384/	0.1607/	0.7132/	0.5237/	0.6105/	0.4895/	0.6498/
		0.4693	0.1613	0.5711	0.5105	0.6500	0.4974	0.6438
	ENIGMA	0.1061/	0.0870/	0.6868/	0.7158/	0.6816/	0.6211/	0.5884/
		0.0612	0.0600	0.5079	0.5526	0.4447	0.4553	0.7607
	AVDE	0.2151/	0.3112/	0.6921/	0.7579/	0.6395/	0.6658/	0.4554/
		0.1944	0.3838	0.5579	0.5263	0.4263	0.4895	0.5087
	ATM-Enc.	0.2675/	0.1137/	0.5681/	0.5661/	0.5391/	0.5660/	0.4319/
		0.2407	0.1037	0.4914	0.5031	0.5137	0.4870	0.4758
	ENIGMA-Enc.	0.2757/	0.1117/	0.5589/	0.5520/	0.5289/	0.5508/	0.4447/
		0.1400	0.0916	0.4822	0.5012	0.4822	0.4951	0.5746
	Dummy	0.4923	0.1456	0.5000	0.5000	0.5000	0.5000	0.6666
Grootswagers2025	LaBraM	0.2364/	0.1044/	0.5002/	0.5013/	0.4994/	0.5000/	0.4532/
		0.2403	0.1033	0.5008	0.5002	0.4993	0.5000	0.4505
	REVE	0.2409/	0.1028/	0.4989/	0.4992/	0.4982/	0.4986/	0.4554/
		0.2518	0.0988	0.4994	0.4994	0.4982	0.5005	0.4573
	LUNA	0.2451/	0.1020/	0.5001/	0.5001/	0.4999/	0.4999/	0.4569/
		0.2444	0.1025	0.4998	0.4998	0.4999	0.4997	0.4571
	ATM	0.4116/	0.1194/	0.5027/	0.4956/	0.5024/	0.5062/	0.6557/
		0.4135	0.1191	0.4981	0.4964	0.5074	0.4974	0.6571
	ENIGMA	0.0507/	0.0598/	0.5187/	0.4974/	0.4660/	0.4962/	0.7464/
		0.1029	0.0701	0.5087	0.4996	0.4911	0.4988	0.7215
	AVDE	0.1447/	0.3395/	0.5021/	0.4868/	0.4731/	0.4997/	0.5035/
		0.1657	0.3715	0.4764	0.4969	0.4784	0.4938	0.4899
	ATM-Enc.	0.2378/	0.1020/	0.5003/	0.4990/	0.4999/	0.5007/	0.4559/
		0.2440	0.1015	0.5018	0.4997	0.5008	0.4960	0.4529
	ENIGMA-Enc.	0.1365/	0.0937/	0.5025/	0.5028/	0.4995/	0.5008/	0.5484/
		0.0589	0.0794	0.5011	0.5026	0.5025	0.5059	0.6403
	Dummy	0.4673	0.1429	0.5000	0.5000	0.5000	0.5000	0.6795

Table D.3: **Raw results for video reconstruction.** For task-specific methods, `Enc.` denotes the EEG encoder evaluated in the unified pipeline, while unsuffixed methods use their original pipelines. Foundation models use the unified pipeline. Clean results are shown on the first row and noisy results on the following row.

Dataset	Baseline	Video-2-way ↑	Video-40-way ↑	Frame-2-way ↑	Frame-40-way ↑	SSIM ↑	PSNR ↑	Hue-pcc ↑
SEED-DV	LaBraM	0.7688/	0.0888/	0.6985/	0.0850/	0.2605/	8.2092/	0.7370/
		0.7386	0.0811	0.6680	0.0634	0.2336	7.7988	0.7190
	REVE	0.7689/	0.0922/	0.7113/	0.0931/	0.2597/	8.2141/	0.7420/
		0.7681	0.1110	0.5713	0.0251	0.2829	8.2804	0.6747
	LUNA	0.7658/	0.0875/	0.7136/	0.0939/	0.2579/	8.2259/	0.7342/
		0.7665	0.1100	0.5841	0.0275	0.2621	7.8563	0.6224
	EEG2Video	0.7460/	0.1256/	0.7572/	0.1109/	0.2577/	8.2868/	0.6376/
		0.7273	0.0711	0.6825	0.0632	0.2582	7.3918	0.5166
	EEGVid	0.7912/	0.1406/	0.7076/	0.0918/	0.2433/	8.6283/	0.7268/
		0.7777	0.1061	0.5974	0.0225	0.2388	7.9190	0.7113
	EEG2Video-Enc.	0.7634/	0.0967/	0.7229/	0.1016/	0.2607/	8.2767/	0.7397/
		0.7092	0.1221	0.8757	0.1873	0.2486	7.4984	0.7009
	EEGVid-Enc.	0.7738/	0.1019/	0.7118/	0.0950/	0.2534/	8.2078/	0.7309/
		0.7723	0.0847	0.7840	0.1654	0.2361	7.9970	0.6612
Smit2025	LaBraM	0.8119/	0.2300/	0.9059/	0.2700/	0.2293/	4.8959/	0.5284/
		0.7892	0.1623	0.7401	0.2427	0.2073	4.7601	0.5044
	REVE	0.8224/	0.2062/	0.9043/	0.2783/	0.2213/	4.8948/	0.5414/
		0.7931	0.1734	0.7534	0.2517	0.2232	4.8790	0.5282
	LUNA	0.8149/	0.1924/	0.9015/	0.2729/	0.2240/	4.8797/	0.5272/
		0.8073	0.1517	0.7866	0.2549	0.2020	4.7596	0.4800
	EEG2Video	0.8970/	0.3083/	0.8889/	0.2844/	0.1379/	12.5843/	0.5613/
		0.7893	0.1899	0.7760	0.2469	0.2008	5.8762	0.5200
	EEGVid	0.9616/	0.5694/	0.9769/	0.6428/	0.6990/	14.4089/	0.6270/
		0.7879	0.1970	0.7863	0.2639	0.2011	4.9838	0.6010
	EEG2Video-Enc.	0.7979/	0.1719/	0.8980/	0.2585/	0.2211/	4.9111/	0.5389/
		0.7644	0.1648	0.8025	0.2163	0.1901	4.7234	0.5229
	EEGVid-Enc.	0.8061/	0.1774/	0.8993/	0.2640/	0.2212/	4.9025/	0.5297/
		0.7595	0.1603	0.8095	0.2196	0.2266	4.7855	0.4921
EMD	LaBraM	0.7492/	0.0868/	0.7949/	0.1302/	0.1744/	5.3656/	0.5854/
		0.7194	0.0717	0.6690	0.0540	0.1745	5.2284	0.5461
	REVE	0.7474/	0.0925/	0.7824/	0.1305/	0.1755/	5.4117/	0.6076/
		0.7018	0.0693	0.6133	0.0701	0.1882	5.4532	0.5750
	LUNA	0.7518/	0.0845/	0.7802/	0.1262/	0.1724/	5.3669/	0.5882/
		0.7199	0.0694	0.6225	0.0690	0.1567	5.2828	0.5345
	EEG2Video	0.7097/	0.0511/	0.7886/	0.1125/	0.2257/	3.9767/	0.6471/
		0.7113	0.0681	0.6755	0.0538	0.1635	5.2881	0.5991
	EEGVid	0.7881/	0.1243/	0.7842/	0.1332/	0.1676/	6.4269/	0.6287/
		0.7301	0.0800	0.6531	0.0644	0.1984	5.4564	0.5849
	EEG2Video-Enc.	0.7654/	0.0939/	0.7757/	0.1157/	0.1649/	5.3435/	0.5865/
		0.7210	0.0593	0.6737	0.0781	0.1573	5.4125	0.6101
	EEGVid-Enc.	0.7650/	0.0964/	0.7936/	0.1428/	0.1757/	5.3683/	0.5905/
		0.7074	0.0890	0.6526	0.0542	0.1856	5.4695	0.5705

Table D.4: **Raw results for text reconstruction.** For task-specific methods, `Enc.` denotes the EEG encoder evaluated in the unified pipeline, while the unsuffixed method uses its original pipeline. Foundation models use the unified pipeline. Clean results are shown on the first row and noisy results on the following row; Dummy is reported once because it has no noisy result. A dash (-) denotes an unavailable result; for the Chinese datasets, the corresponding task-specific methods are unavailable because their generators do not support Chinese generation.

Dataset	Baseline	Lexical Overlap		N-way / Recall					Diversity		Dist.
		BLEU-1 ↑	BLEU-2 ↑	2-WAY ↑	4-WAY ↑	10-WAY ↑	24-WAY ↑	C.RECALL ↑	DIST-1 ↑	DIST-2 ↑	H.ENT ↑	FD ↓
ZuCo 1	LaBraM	0.2290/	0.0731/	0.5464/	0.2903/	0.1169/	0.0484/	0.0083/	0.1245/	0.3841/	4.5644/	0.5956/
		0.0177	0.0015	0.4897	0.2424	0.0957	0.0405	0.0002	0.0343	0.0883	6.1236	1.1403
	REVE	0.0000/	0.0000/	0.5004/	0.2494/	0.1000/	0.0423/	0.0000/	0.0003/	0.0005/	0.6087/	1.8884/
		0.0000	0.0000	0.4994	0.2498	0.1001	0.0417	0.0000	0.0000	0.0000	0.0590	1.9649
	LUNA	0.2563/	0.0933/	0.5988/	0.3451/	0.1551/	0.0694/	0.0115/	0.1183/	0.3547/	4.2208/	0.6199/
		0.1552	0.0264	0.4994	0.2504	0.1016	0.0433	0.0048	0.0565	0.2090	6.6477	0.8080
	EEG2Text	0.1497/	0.0163/	0.5020/	0.2508/	0.0991/	0.0400/	0.0043/	0.0070/	0.0167/	0.0104/	1.4696/
		0.1582	0.0260	0.4986	0.2472	0.0969	0.0407	0.0040	0.0145	0.0325	0.1475	1.4335
	GLIM	0.1546/	0.0356/	0.5555/	0.3059/	0.1306/	0.0549/	0.0073/	0.1055/	0.2834/	6.2725/	0.6936/
		0.1322	0.0191	0.5006	0.2515	0.1030	0.0459	0.0085	0.0428	0.0995	1.5203	1.1832
	R-EEG2Text	0.1275/	0.0237/	0.6669/	0.4096/	0.1913/	0.0887/	0.0224/	0.0608/	0.1830/	7.1179/	0.7258/
		0.1423	0.0366	0.5001	0.2490	0.0993	0.0414	0.0148	0.0017	0.0026	0.0449	1.7017
	EEG2Text-Enc.	-	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-
	GLIM-Enc.	0.2956/	0.1411/	0.7108/	0.4578/	0.2183/	0.0973/	0.0175/	0.0923/	0.2566/	5.2343/	0.6240/
		0.1811	0.0451	0.4987	0.2471	0.1000	0.0426	0.0058	0.0567	0.1707	4.8222	0.7643
	Dummy	0.1025	0.0221	0.5000	0.2500	0.1000	0.0417	0.0133	0.0007	0.0008	-0.0000	1.6433
ZuCo 2	LaBraM	0.2089/	0.0480/	0.5088/	0.2606/	0.1039/	0.0410/	0.0106/	0.1506/	0.4509/	5.8343/	0.6589/
		0.0013	0.0001	0.5019	0.2507	0.1036	0.0467	0.0000	0.0170	0.0281	5.9991	1.5315
	REVE	0.0027/	0.0001/	0.4909/	0.2449/	0.1008/	0.0466/	0.0000/	0.0336/	0.0566/	6.2135/	1.5600/
		0.0099	0.0010	0.5073	0.2457	0.0919	0.0396	0.0005	0.0083	0.0273	6.3647	1.3586
	LUNA	0.2183/	0.0551/	0.5090/	0.2654/	0.1137/	0.0491/	0.0089/	0.1283/	0.3846/	5.3692/	0.6620/
		0.1437	0.0323	0.5174	0.2679	0.1076	0.0445	0.0056	0.0407	0.1345	5.4493	1.0073
	EEG2Text	0.1562/	0.0342/	0.4977/	0.2513/	0.0993/	0.0405/	0.0104/	0.0163/	0.0435/	1.8123/	1.3262/
		0.1292	0.0216	0.5093	0.2568	0.0986	0.0365	0.0103	0.0274	0.0754	2.8657	1.1982
	GLIM	0.2388/	0.0896/	0.5690/	0.3413/	0.1711/	0.0829/	0.0187/	0.0378/	0.1060/	2.3277/	0.8037/
		0.2280	0.0721	0.4889	0.2372	0.0956	0.0418	0.0030	0.0532	0.1481	3.4712	0.9144
	R-EEG2Text	0.1260/	0.0284/	0.5666/	0.3293/	0.1652/	0.0869/	0.0231/	0.0662/	0.1941/	6.7161/	0.8052/
		0.1520	0.0408	0.5000	0.2500	0.1000	0.0417	0.0175	0.0011	0.0013	-0.0000	1.6698
	EEG2Text-Enc.	-	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-
	GLIM-Enc.	0.2427/	0.1051/	0.7038/	0.4524/	0.229/	0.1097/	0.0172/	0.1010/	0.2661/	5.978/	0.6257/
		0.1223	0.0288	0.5033	0.2552	0.1062	0.0439	0.0027	0.0511	0.1676	5.037	0.9662
	Dummy	0.2010	0.0699	0.5000	0.2500	0.1000	0.0417	0.0015	0.0013	0.0013	-0.0000	1.5402
chisco	LaBraM	0.0684/	0.0066/	0.4971/	0.2476/	0.0979/	0.0407/	0.0000/	0.0053/	0.0083/	0.4534/	0.7643/
		0.1365	0.0061	0.5008	0.2510	0.1010	0.0426	0.0000	0.0061	0.0278	0.0000	0.7747
	REVE	0.0132/	0.0000/	0.5000/	0.2498/	0.1000/	0.0417/	0.0000/	0.0055/	0.3333/	0.9864/	1.4161/
		0.0000	0.0000	0.5000	0.2500	0.0999	0.0417	0.0000	0.0000	0.0000	0.0000	1.7144
	LUNA	0.1271/	0.0107/	0.5000/	0.2458/	0.0988/	0.0399/	0.0000/	0.7077/	0.9525/	8.8148/	0.1727/
		0.0000	0.0000	0.5000	0.2500	0.1000	0.0417	0.0000	0.0000	0.0000	0.0000	1.7289
	EEG2Text	-	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-
	GLIM	-	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-
	R-EEG2Text	-	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-
	EEG2Text-Enc.	-	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-
	GLIM-Enc.	0.1760/	0.0144/	0.4994/	0.2525/	0.1045/	0.0451/	0.0000/	0.0118/	0.5000/	-0.0000/	0.6470/
		0.1372	0.0201	0.5064	0.2560	0.1044	0.0429	0.0000	0.0250	0.0678	2.0778	0.5605
	Dummy	0.1067	0.0122	0.5000	0.2500	0.1000	0.0417	0.0000	0.0008	0.0000	0.0000	0.5941
tmnred	LaBraM	0.1515/	0.0244/	0.5002/	0.2502/	0.1009/	0.0430/	0.0000/	0.0203/	1.0000/	0.0000/	0.7269/
		0.0074	0.0000	0.5038	0.2587	0.1098	0.0492	0.0000	0.6105	0.7179	3.1494	1.0797
	REVE	0.0000/	0.0000/	0.4989/	0.2496/	0.0989/	0.0406/	0.0000/	0.0021/	0.0000/	0.0000/	1.5760/
		0.0007	0.0000	0.4923	0.2453	0.0971	0.0379	0.0000	0.1666	0.1482	4.8897	1.3369
	LUNA	0.1579/	0.0321/	0.5018/	0.2595/	0.1062/	0.0449/	0.0000/	0.6795/	0.7532/	5.1996/	0.4001/
		0.0000	0.0000	0.4978	0.2484	0.1048	0.0477	0.0000	0.0958	0.1439	6.3948	1.4327

Continued on next page

Table D.4: **Raw results for text reconstruction.** For task-specific methods, `Enc.` denotes the EEG encoder evaluated in the unified pipeline, while the unsuffixed method uses its original pipeline. Foundation models use the unified pipeline. Clean results are shown on the first row and noisy results on the following row; Dummy is reported once because it has no noisy result. A dash (-) denotes an unavailable result; for the Chinese datasets, the corresponding task-specific methods are unavailable because their generators do not support Chinese generation. (Continued)

Dataset	Baseline	Lexical Overlap		N-way / Recall					Diversity			Dist.
		BLEU-1 ↑	BLEU-2 ↑	2-WAY ↑	4-WAY ↑	10-WAY ↑	24-WAY ↑	C.RECALL ↑	DIST-1 ↑	DIST-2 ↑	H.ENT ↑	FD ↓
	EEG2Text	-	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-
	GLIM	-	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-
	R-EEG2Text	-	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-
	EEG2Text-Enc.	-	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-
	GLIM-Enc.	0.2040/	0.0462/	0.4957/	0.2408/	0.0948/	0.0407/	0.0000/	0.0506/	1.0000/	1.0000/	0.6888/
		0.0197	0.0011	0.4969	0.2468	0.1074	0.0487	0.0000	0.8769	0.9995	8.7211	0.8878
	Dummy	0.0985	0.0344	0.5000	0.2500	0.1000	0.0417	0.0000	0.0020	0.0000	0.0000	0.7078