Predicting in Latent Space Leads to Faithful EEG-to-Text Decoding

Anonymous ACL submission

Abstract

We present GRAPE, a JEPA-based pretraining framework for open-vocabulary EEG-to-text decoding that addresses two compounding failures in prior work: self-supervised objectives that reconstruct raw EEG amplitudes are forced to retain artefact noise, and teacher-forcing evaluation has rendered most published benchmarks unreliable. By predicting in *latent* rather than signal space, GRAPE performs lossy compression of the EEG, filtering non-semantic artefacts by construction; element-wise gated self-attention provides a learned, label-free artefact gate. Integrated into a strictly teacher-forcing-free pipeline (GRAPE-E2T), the frozen encoder recovers subject identity at **99.6%** ($29.9\times$ chance) and nine-class discourse-relation type at **56.4%** ($5.1\times$ chance) from EEG alone, with no text supervision. Under noise-input perturbation, GRAPE-E2T's BLEU-1 collapses by $> 99\%$, versus $\sim 70\%$ retention for the prior state of the art, confirming that the performance gap in prior work is largely attributable to the language prior, not EEG decoding. On LLM-judged top-3 relational accuracy, GRAPE-E2T achieves a **+16.6-point** gain over the prior state of the art. Critically, we show that BLEU *inversely* tracks EEG faithfulness: the model with higher BLEU is the one whose output is less conditioned on the EEG signal, a finding with broad implications for how EEG-to-text systems are evaluated. We introduce ETES, an EEG-aware energy-based metric that captures this distinction, and release code and checkpoints on acceptance.

Keywords: EEG-to-text decoding; brain-computer interfaces; self-supervised learning; joint-embedding predictive architecture; gated attention; posterior collapse; energy-based evaluation.

1 Introduction

Decoding language directly from electroencephalography (EEG) promises an assistive interface for patients with severe motor disorders (Birbaumer et al., 1999). Open-vocabulary EEG-to-text decoding aims to generate the text a participant was reading from the EEG recorded during reading. As a brain-computer interface (BCI) paradigm it offers far higher expressive bandwidth than fixed-vocabulary alternatives such as P300 spellers (Cecotti and Gräser, 2010) or motor-imagery interfaces (Lotte et al., 2018), at the cost of a substantially harder learning problem: parallel corpora contain only thousands of EEG–sentence pairs, naturalistic EEG is non-stationary and artefact-laden, and any given sentence admits many lexically distinct but semantically equivalent renderings.

Despite five years of work (Wang and Ji, 2022; Liu et al., 2024; Wang et al., 2024; Duan et al., 2023; Zhou et al., 2024b; Mishra et al., 2025), progress on the canonical open-vocabulary EEG-to-text benchmark ZuCo (Hollenstein et al., 2018) has been hampered by two distinct limitations that compound each other.

Limitation 1: weak self-supervision. EEG signals are non-stationary, artefact-laden, and exhibit large inter-subject variability (Lotte et al., 2018). The dominant SSL paradigms are poorly matched. Masked autoencoders (He et al., 2022; Chien et al., 2022) reconstruct raw amplitudes, allocating capacity to the structured noise that dominates EEG variance. Contrastive methods (Chen et al., 2020; Zhou et al., 2024a) require principled augmentations and large negative-sample batches, neither of which EEG provides naturally.

Limitation 2: posterior collapse and evaluation. Jo et al. (2024) showed that virtually every prior EEG-to-text model evaluated under *teacher forcing*, feeding the gold previous token at every decoding step, achieves near-identical scores when the EEG input is replaced with random noise. Reported BLEU improvements over the past five years are therefore largely attributable to the language-model

prior rather than to EEG decoding. Only a few models including the recent GLIM (Liu et al., 2025) model reports under teacher-forcing-free conditions and remains the only credible baseline. However, even under correct teacher-forcing-free evaluation GLIM retains $\sim 70\%$ of its BLEU-1 when the EEG input is replaced with Gaussian noise (§6.2), indicating substantial posterior collapse: most of its reported performance is still attributable to the language-model prior rather than to EEG conditioning. Closing the evaluation artefact is therefore a necessary but not sufficient step; the encoder must additionally be strong enough that the decoder cannot bypass it.

This work. We argue that both limitations can be addressed simultaneously through (a) a pretraining objective whose gradient is shaped by the structure of *representations* rather than raw signal values, and (b) evaluation that is sensitive to genuine EEG conditioning. We introduce GRAPE (Gated Reading-Aligned Predictive Representation for EEG), a JEPA-based pretraining recipe (LeCun, 2022; Assran et al., 2023) for reading EEG that combines (i) Semantic Subsequence Preserving masking (Foumani et al., 2024), (ii) the non-contrastive VICReg objective (Bardes et al., 2022), and (iii) *element-wise gated* self-attention. We extend GLIM's evaluation-corrected generation framework with this JEPA-pretrained encoder, using it to isolate encoder pretraining as the variable of interest: the pipeline is shared, but the learned representations entering it are fundamentally different. Throughout, we evaluate without teacher forcing and report a noise-input sanity check.

Contributions.

1. We present, to our knowledge, the first JEPA-pretrained EEG encoder integrated into an open-vocabulary EEG-to-text generation pipeline, combining SSP masking, VICReg, and element-wise gated self-attention in a unified non-contrastive framework (§3).
2. We show that this encoder learns semantically structured EEG representations *without any text supervision*: linear probing of the frozen encoder recovers subject identity and nine-class relation type far above chance (§5.2).
3. We provide direct evidence that GRAPE-E2T decodes from the EEG signal rather than from a language prior: under teacher-forcing-free evaluation it nearly collapses under a noise-input test, in contrast to the prior state of the art, and improves LLM-judged relational semantics of the generated text (§6).
4. We propose ETES, an EEG-aware energy-based score that measures, on the test set, the alignment between an EEG embedding and its generated/reference text in the model's shared latent space (§4).

Figure 1 gives a schematic of the two-stage pipeline; a fully-annotated version with tensor shapes, the internal structure of the gated-attention block, and the loss decomposition appears in Appendix B (Figure 3).

2 Related Work

Open-vocabulary EEG-to-text. Wang and Ji (2022) established the task on ZuCo using a BART-based seq-to-seq model. Subsequent work explored deeper representations (Amrani et al., 2024), multi-view pretraining with masked autoencoding (Liu et al., 2024), contrastive EEG–text masked autoencoders coupled to a frozen LM (Wang et al., 2024), discrete codebook tokenisation (Duan et al., 2023), CLIP-style contrastive alignment (Zhou et al., 2024b,a), and LLM-driven decoding from image-evoked EEG (Mishra et al., 2025). Jo et al. (2024) re-evaluated this generation of models without teacher forcing and found near-random performance, invalidating their claims.

GLIM (Liu et al., 2025) is one of the paper that addresses the evaluation flaw directly, using a Q-Merger, contrastive latent alignment, multi-text-variant training, and a frozen Flan-T5 decoder which inspired our work. Our work is closest to Liu et al. (2024); Wang et al. (2024), which also pretrain EEG encoders before downstream generation; both, however, rely on *raw-amplitude* reconstruction (MAE-style), an objective that allocates capacity to the structured noise that dominates EEG variance and uses teacher forcing. We instead predict in *latent space* (JEPA) and remove the dependence on negatives via VICReg.

From an Information Bottleneck perspective (Tishby et al., 2000), the distinction is principled. A *signal-space* (MAE) objective demands faithful reconstruction, forcing the encoder to retain structured artefact noise through a wide bottleneck. A *latent-space* (JEPA) objective performs *lossy compression* instead: only information predictable from context must be preserved; non-semantic artefacts, unpredictable by definition which are filtered out.

Our A3 ablation validates this directly: replacing signal-space reconstruction with latent prediction is the component most responsible for semantic-level abstraction (§6.7). We take GLIM's evaluation-corrected generation framework as our downstream pipeline, isolating encoder pretraining as the variable of interest; the architecture is shared but the learned representations are not.

JEPA-based pretraining for EEG. The Joint-Embedding Predictive Architecture (LeCun, 2022; Assran et al., 2023) predicts *representations* of masked targets from unmasked context in latent space, sidestepping both the amplitude-noise allocation of MAE-style reconstruction and the augmentation requirements of contrastive methods, neither of which EEG can satisfy naturally. Recent work has applied JEPA to neural (Dong et al., 2024) and EEG-specific (Guetschel et al., 2024; Hojjati et al., 2025) signals, but only for classification probes; *no JEPA-pretrained encoder has previously been used as the backbone for open-vocabulary EEG text generation*. SSP masking (Foumani et al., 2024) and VICReg (Bardes et al., 2022) have likewise not been combined within a JEPA cross-attention predictor for EEG.

Gated attention. Element-wise gating in self-attention is a long-standing technique (Dauphin et al., 2017). Concurrent large-scale work by Qiu et al. (2025) systematically studies 30 gating variants and identifies a head-specific sigmoid gate after scaled dot-product attention as the most effective augmentation, attributing the gain to input-dependent sparsity and elimination of the attention-sink phenomenon. Their finding, based on 15 B-parameter MoE and 1.7 B-parameter dense models trained on 3.5 T tokens offers external validation of the gating design we adopt independently. To our knowledge, this gating formulation has not previously been applied *within an EEG self-supervised encoder for reading-EEG representation*, and it is the experimental locus of our V1/V2 comparison (§6.6).

3 Method

3.1 Problem Setup

Let $\mathcal{D}=\{(X_i, Y_i)\}_{i=1}^{N}$ be a sentence-aligned dataset where $X_i \in \mathbb{R}^{T\times C}$ is a raw EEG recording with T=1280 time points and C=128 channels, and $Y_i \in V^*$ is the corresponding sentence. A decoding model factorises as $g_\varphi \circ f_\theta$, where $f_\theta : \mathbb{R}^{T\times C} \to \mathbb{R}^{N\times d}$ is an EEG encoder and g_φ is a seq-to-seq decoder. When f_θ is trained end-to-end under a language-modelling loss ℓ_{LM} alone, the gradient signal to the encoder is indirect (mediated by word-prediction error) and decoder-biased: a strong autoregressive decoder can minimise ℓ_{LM} from its language prior, suppressing pressure on f_θ. This is the posterior collapse identified by Jo et al. (2024); Liu et al. (2025). Our remedy is to front-load encoder learning into a self-supervised stage whose gradient does not flow through a language model.

3.2 Stage I: GRAPE Pretraining

Patching. We partition X along time into $N{=}T/P{=}160$ non-overlapping patches of size $P{=}8$ (62.5 ms at 128 Hz, aligned with cortical alpha/beta oscillations (Giraud and Poeppel, 2012)). Each patch is flattened and projected by Linear→LN→GELU to embedding dimension $d{=}128$ with fixed sinusoidal positional encodings.

SSP masking. Following Foumani et al. (2024) we sample two contiguous *visible* chunks covering a fraction $1{-}\rho$ of positions (ρ=0.5, block size $\lceil((1{-}\rho)N)/\beta\rceil$, β=2). Contiguity preserves the local oscillatory structure that uniform random masking destroys, and yields a non-trivial prediction problem.

Context encoder f_θ with gated self-attention. The visible token sequence Z_V is processed by a 6-block, 8-head transformer with residual MLPs. Each block uses *element-wise gated self-attention*:

$$H_{\mathrm{attn}} = \mathrm{Gate}(\sigma) \odot \mathrm{Attn}(Q, K, V), \qquad (1)$$

where $\mathrm{Gate}(\sigma)$ is a sigmoid-activated linear projection of the attention output and $\odot$ is the elementwise product. The gate provides a differentiable, per-position artefact-suppression mechanism: artefact-dominated dimensions can be down-weighted adaptively without explicit artefact-rejection preprocessing (Jas et al., 2017; Dauphin et al., 2017).

Target encoder f_ξ (EMA). f_ξ shares the architecture of f_θ but receives no gradient and is updated by exponential moving average, $\xi \leftarrow m\,\xi + (1{-}m)\,\theta$, m=0.99. Targets are taken at masked positions only: $Y = H^t_{[M]} \in \mathbb{R}^{|M|\times d}$.

Cross-attention predictor. A deliberately shallow 2-layer cross-attention predictor (Assran et al.,

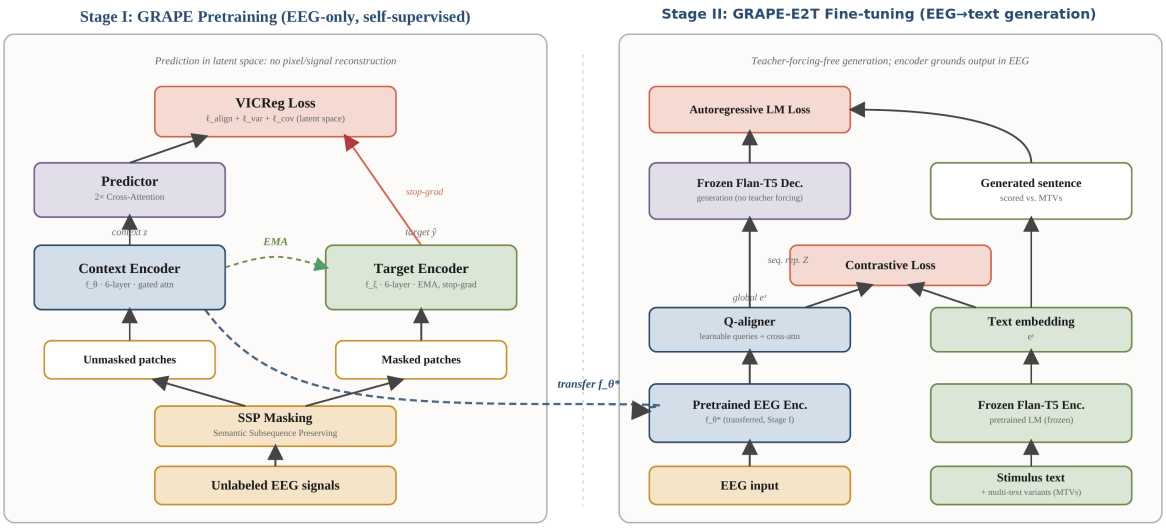

Figure 1: The two-stage GRAPE pipeline. **Stage I** (GRAPE, left): SSP-masked EEG patches feed a gated context encoder f_θ and an EMA target encoder f_ξ; a cross-attention predictor regresses latent targets under VICReg, with no signal-space reconstruction. **Stage II** (GRAPE-E2T, right): the pretrained f_{θ^*} feeds a Q-aligner bridging EEG into the frozen T5 encoder/decoder; a CMC loss and MTV language-model loss are combined. Trainable modules are shaded blue, frozen/EMA modules green, and loss terms red. Full detail in §3; annotated version with gated-attention internals in Appendix B.

2023; Baevski et al., 2022) uses learnable mask tokens with sinusoidal positional embeddings as queries and $H^c{=}f_\theta(Z_V)$ as keys/values, producing $\hat{Y} \in \mathbb{R}^{|M|\times d}$.

VICReg objective. We minimise the sum of three terms over masked positions:

$$\ell_{\text{align}} = \tfrac{1}{|M|}\textstyle\sum_n \|\hat{Y}_n - Y_n\|_2^2, \quad (2)$$

$$\ell_{\text{std}} = \tfrac{1}{d}\textstyle\sum_j \max\big(0, \gamma - \text{std}_j(\hat{Y})\big), \quad (3)$$

$$\ell_{\text{cov}} = \tfrac{1}{d}\textstyle\sum_{i\neq j}\big[\text{Cov}(\hat{Y})\big]_{ij}^2, \quad (4)$$

with $\gamma{=}1$. The total Stage I loss is $\ell_{\text{SI}} = \ell_{\text{align}} + \mu\,\ell_{\text{std}} + \nu\,\ell_{\text{cov}}$ with $\mu{=}\nu{=}1$. VICReg requires neither negative pairs (infeasible at ZuCo's sample size) nor augmentations (Bardes et al., 2022), both of which preclude contrastive alternatives in the EEG domain.

3.3 Stage II: GRAPE-E2T Fine-tuning

We build on GLIM's generation pipeline, initialising its encoder slot with θ^* from Stage I, and fine-tune end-to-end with three modules: (i) the EEG encoder f_{θ^*} producing token-level $Z \in \mathbb{R}^{L\times d_e}$ and a global embedding e^x via a Q-aligner; (ii) the Q-aligner, a cross-attention module with a learnable query that projects EEG into the frozen LM's latent space; and (iii) a frozen Flan-T5 (Chung et al., 2022) serving as both the text-side embedder and the generation decoder; freezing the LM is the principal posterior-collapse defence, leaving only f_{θ^*} and the Q-aligner free to adapt. The Q-aligner design and frozen Flan-T5 are shared with GLIM deliberately, holding the downstream pipeline fixed so that any performance difference is attributable to the encoder alone. Three things differ: *(a)* f_{θ^*} arrives warm-started from Stage I rather than randomly initialised, so the contrastive loss operates on a latent space pre-shaped to discriminate subjects, relations, and reading paradigms before any text supervision; *(b)* element-wise gated self-attention (Eq. 1) remains active throughout fine-tuning in V2 but is absent in V1 and from GLIM, making V1↔V2 a controlled ablation of the gate within an otherwise identical pipeline; and *(c)* encoder weight decay is set to zero, since Stage I VICReg regularisation renders it redundant.

Objective. Following Liu et al. (2025) we use multi-text-variant (MTV) training with K paraphrase variants per stimulus (Appendix G) and add a symmetric InfoNCE (Radford et al., 2021) cross-modal contrastive loss ℓ_{CMC} over EEG/text embedding pairs:

$$\ell_{\text{LM}} = -\tfrac{1}{NK}\textstyle\sum_{i,j}\log P(\hat{Y}_i^j \mid Z_i), \quad (5)$$

$$\ell_{\text{total}} = \lambda\,\ell_{\text{LM}} + (1{-}\lambda)\,\ell_{\text{CMC}}. \quad (6)$$

ℓ_{CMC} cannot be minimised by language-prior hallucination: it requires e_i^x to be closer to its matched

e_i^y than to all batch negatives, creating direct gradient pressure on f_{θ^*} to encode EEG-discriminative content.

4 Evaluation Framework

We evaluate four complementary properties.

(1) Generation, retrieval, and classification. BLEU-1, BLEU-2 (Papineni et al., 2002) scored against the multiple text variants (MTVs), and ROUGE-1 (Lin, 2004) scored against the raw stimulus text; top-k retrieval accuracy of EEG embeddings against text candidates; zero-shot semantic classification (sentiment 3-cls, relation 9-cls, corpus 2-cls) of generated text by a fixed classifier. *All decoding is autoregressive without teacher forcing.*

(2) Posterior-collapse diagnostic. Following Jo et al. (2024) we replace the EEG input with i.i.d. Gaussian noise of matched second-order statistics and recompute every metric. Genuine EEG conditioning manifests as a large gap $\Delta = m_{\text{EEG}} - m_{\text{noise}}$; posterior-collapsed models exhibit $\Delta \approx 0$.

(3) LLM-assisted semantic classification. A fixed instruction-tuned LLM is prompted to label generated sentences with a sentiment, relation type, or corpus tag. Top-1 and top-3 accuracies quantify whether generations preserve task-relevant semantics independently of lexical overlap, addressing a known weakness of n-gram metrics (Reiter, 2018).

(4) ETES: EEG–Text Energy Score. Inspired by energy-based modelling (LeCun et al., 2006), let $f_{\text{EEG}}(\cdot)$ and $f_{\text{text}}(\cdot)$ map EEG and text into the model's shared aligner space. The per-sample energy is

$$E(a, b) = -\cos(a, b), \tag{7}$$

and we define three test-set aggregates:

$$\text{ETES}_{\text{align}} = \tfrac{1}{N} \textstyle\sum_i E(z_i, \hat{t}_i), \tag{8}$$
$$\text{ETES}_{\text{ref}} = \tfrac{1}{N} \textstyle\sum_i E(z_i, t_i), \tag{9}$$
$$\text{ETES}_{\text{gap}} = \text{ETES}_{\text{align}} - \text{ETES}_{\text{ref}}, \tag{10}$$

where $\hat{t}_i$ is the generated text and t_i is the reference. $\text{ETES}_{\text{align}}$ is reference-free, ETES_{ref} functions as an upper-bound anchor, and ETES_{gap} quantifies how far generated text departs from where the reference sits in the shared space. All three are computed on the same test split as our other metrics, so ETES directly measures the property the model was optimised for.

5 Experiments

5.1 Data and configurations

ZuCo and ZuCo 2.0. We use the publicly available ZuCo (Hollenstein et al., 2018) and ZuCo 2.0 (Hollenstein et al., 2020) corpora, the standard naturalistic-reading benchmark for open-vocabulary EEG-to-text. Together they provide 128-channel EEG (500 Hz) from 30 native English speakers and over 22,000 sentence-aligned EEG–text pairs spanning two reading paradigms: normal reading (NR; passive) and task-specific reading (TSR; active reading with comprehension questions). Two properties make ZuCo well-suited to our study. First, it is deliberately *heterogeneous*, varying across reading paradigms, source corpora, recording sessions, and subjects, which is exactly the distribution shift a generalizable encoder must absorb and which our self-supervised pretraining targets. Second, its source corpora are standard NLP benchmarks, the Stanford Sentiment Treebank and a Wikipedia relation-extraction subset, so their sentiment and discourse-relation labels integrate directly with pretrained language models for both supervision and our zero-shot semantic evaluation. Eye-tracking is recorded but unused, testing a purely neural decoder. Per-group statistics, the full label set, and preprocessing are in Appendix A.

Preprocessing and configuration. EEG is downsampled from 500 Hz to 128 Hz, segmented at sentence boundaries, and zero-padded to T=1280 (10 s). We deliberately do *not* apply ICA artefact rejection or bandpass filtering, since the gated-attention mechanism (§3.2) is intended to learn this function from data (Appendix A). Splits are text-disjoint to prevent leakage; ZuCo 2.0 held-out subjects are reserved for cross-subject probing.

We compare two pretraining variants. **V1** omits gated self-attention (vanilla 2×6×6 transformer); **V2** adds element-wise gated attention in every encoder block. The V1/V2 distinction is a property of the *encoder* and carries through to the full pipeline: GRAPE-V1/V2 denote the pretrained encoders, and GRAPE-E2T-V1/V2 the corresponding fine-tuned decoders built on them. Both share patch size P=8, d=128, ρ=0.5, EMA m=0.99, AdamW (Loshchilov and Hutter, 2019) with learning rate 10^{-4}, weight decay 0.01, gradient clipping at $\|g\|_2 \leq 1$, and bf16 mixed precision. Pretraining: batch size 64 for up to 200 epochs with cosine decay. Fine-tuning: learning rate 10^{-4}, zero weight

Linear probe	Chance	V1	V2
Subject ID (30-cls)	3.3	18.8	**99.6**
Sentiment (3-cls)	33.3	34.5	33.6
Relation (9-cls)	11.1	15.6	**56.4**
Reading paradigm (2-cls)	50.0	59.8	**71.8**

Table 1: Linear-probe accuracy (%) on frozen GRAPE encoders (V1 = no gate, V2 = gated). Chance levels shown for reference.

decay, up to 200 epochs with best-validation selection. Random seed 42. Full hyperparameters and stability analysis appear in Appendix D.

5.2 Linear-probe representation quality

Before any text supervision, we freeze the GRAPE encoder and train logistic-regression classifiers with L_2 regularisation on the global embedding e^x. We probe four tasks: 30-class subject identification, 3-class sentiment, 9-class discourse- relation classification, and 2-class reading paradigm classification (NR vs. TSR). The V1↔V2 contrast functions as a controlled ablation of the gated-attention component; other Stage I components are ablated separately in §6.7.

Table 1 reports the results. Gated attention transforms the encoder qualitatively, not just quantitatively: without gating (V1), the encoder recovers only weak structure across every probe except sentiment; with gating (V2), three of the four probes leap to far-above-chance accuracy in a single architectural change. The largest reorganisation is on subject identification, where the gate flips a near-random encoder into a subject-discriminative one, suggesting the gate's adaptive suppression is the primary mechanism by which f_θ distinguishes person-level neural signatures. Relation and paradigm probes show the same direction of effect at smaller magnitudes, indicating that the gain generalises across content-level distinctions rather than being subject-specific. Sentiment remains near chance for both variants because its labels cover only the SST-NR subset ($\sim$20% of the corpus); this is a label-availability ceiling, not a representational failure.

A t-SNE projection of the V2 latent space (Figure 4, Appendix C) corroborates these probes: relation classes form visibly separated clusters under EEG-only self-supervision.

6 Results and Discussion

Our central finding is that surface BLEU disagrees most with every direct measure of EEG grounding. We organise empirical evidence around this dissociation: §6.1 establishes it on the benchmark; §6.2 shows it is a posterior-collapse artefact; the remaining subsections triangulate genuine grounding from four independent angles; §6.8 draws out the implications.

6.1 Surface fluency and EEG grounding dissociate

Table 2 surfaces a dissociation. GRAPE-E2T V2 trails GLIM on BLEU-1 yet *outperforms* V1 on every metric that probes the encoder directly; BLEU ranks them the other way. A metric that improves as the encoder gets worse is not measuring the encoder. We show below that the BLEU ordering tracks decoder fluency rather than EEG content.

6.2 The dissociation is posterior collapse, not a model quirk

The noise-input diagnostic of Jo et al. (2024) isolates the cause: a model genuinely conditions on EEG only if its outputs degrade substantially when the EEG is replaced with statistics-matched Gaussian noise. GRAPE-E2T collapses to near-zero BLEU-1 under noise ($>99.9\%$ drop for both variants; retrieval ACC@1 returns to chance), while GLIM retains roughly 70% of its clean BLEU-1, far above any reasonable collapse threshold. Most of GLIM's BLEU is therefore recoverable from its language prior alone, so its clean-EEG advantage reflects a fluent decoder rather than a faithful neural one. The conditional drop Δ, not the absolute score, separates decoding from hallucination, and on Δ the ordering reverses entirely.

6.3 Generations preserve relational semantics better than the prior SOTA

When an LLM judge scores the *meaning* of generated sentences rather than their surface overlap, the ordering from BLEU inverts. GRAPE-E2T V1 reaches LLM-judged top-3 relation accuracy of 0.729, a +16.6-pp gain over GLIM (Table 3); all three saturate corpus classification near the raw-text ceiling. The model that loses on lexical overlap best preserves relational and topical structure, the pattern expected if BLEU rewards fluency while the judge rewards content.

Model	Generation			Retrieval		Classification		
	BLEU-1*	BLEU-2*	ROUGE-1‡	ACC@1	ACC@5	Rel@1	Rel@3	Corpus
DeWave+Contr.† (Duan et al., 2023)	0.211	0.107	0.220	–	–	–	–	–
Random baseline† (Jo et al., 2024)	–	–	0.042	0.208	0.208	0.333	0.111	0.500
GLIM (Liu et al., 2025)	**0.260**	**0.106**	**0.123**	**0.427**	**0.351**	0.325	0.571	**0.935**
+ Noise input	0.182	0.045	0.111	0.037	0.207	0.357	0.145	0.627
GRAPE-E2T V1 (ours)	0.228	0.058	0.113	0.285	0.307	0.137	0.365	0.412
+ Noise input	0.000	0.000	0.001	0.042	0.208	0.049	0.028	0.167
GRAPE-E2T V2 (ours)	0.208	0.049	0.112	0.345	0.208	0.153	0.348	0.571
+ Noise input	**0.000**	**0.000**	**0.000**	0.042	0.208	0.049	0.167	0.167

Table 2: ZuCo test-set results under **teacher-forcing-free** evaluation. Shaded noise-input rows test posterior collapse (§6.2). †: different data split with potential train–test text overlap; approximate reference only. *: BLEU scores computed against multiple text variants (MTVs). ‡: ROUGE-1 computed against raw stimulus text.

Pathway	Model	Sent @1	Rel @3	Corpus
Gen. text (LLM judge)	GLIM	0.396	0.563	0.922
	GRAPE-E2T V1	0.309	**0.729**	0.912
	GRAPE-E2T V2	0.317	0.646	0.919
EEG embed. (CLIP-like)	GLIM	**0.427**	0.571	**0.935**
	GRAPE-E2T V1	0.285	0.365	0.412
	GRAPE-E2T V2	0.324	0.348	0.571

Table 3: Semantic-content classification of generated text (top) vs. EEG embeddings (bottom). Full table in Appendix F.

Score (lower = better)	V1	V2	Δ
$\text{ETES}_{\text{align}}$	−0.451	**−0.488**	−0.037
ETES_{ref}	−0.149	−0.139	+0.010
ETES_{gap}	−0.302	**−0.348**	−0.046

Table 4: ETES on ZuCo test set. Negative ETES_{gap} indicates EEG-grounded paraphrasing: generations align more tightly to the EEG than the reference text does (see §6.4).

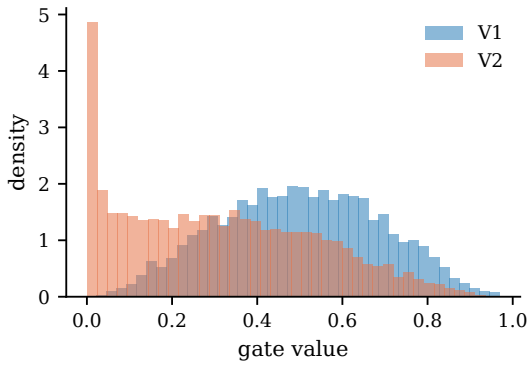

Figure 2: Gate-value distributions on the test set. V2 places more mass near 0 than V1 while preserving a broad spread.

6.4 Generations align to EEG more tightly than the reference text does

ETES measures alignment on the test set in the model's shared embedding space (Table 4). Both variants land in the "good" band; V2 improves over V1. The key detail is the *negative* gap: generations align to the EEG more tightly than the reference itself does the signature of EEG-grounded paraphrasing rather than verbatim reconstruction, and consistent with modest BLEU. Per-sample distributions are in Appendix H.

6.5 Representations transfer across unseen subjects

On the ZuCo 2.0 cross-subject paradigm-classification benchmark (Hollenstein et al., 2023), frozen GRAPE-V2 and fine-tuned GRAPE-E2T-V2 decode held-out subjects on EEG alone, competitive with concatenated hand-crafted features though below multi-modal eye-tracking baselines; full results, baselines, and interpretation are in Appendix I.

6.6 Gated attention learns artefact suppression without labels

V2 gates show selective per-position suppression rather than uniform down-weighting: higher entropy, sparsity, and std; lower mean than V1 (Figure 2; Table 11). V2's sparsity of 14.4% closely matches the 10–20% artefact prevalence typical of natural-reading EEG (Jas et al., 2017), emerging with no artefact labels. This is our most direct evidence that gated attention learns artefact suppression from self-supervision alone (behavioural match, not a comparison against ICA; see Limitations).

Linear probe	V2 full	A1 r.mask	A2 noVIC	A3 MAE	V1 nogate
Subject ID (ch. 3.3)	**99.6**	96.4	82.7	86.4	18.8
Sentiment (ch. 33.3)	33.6	31.4	31.2	33.1	**34.5**
Relation (ch. 11.1)	**56.4**	33.7	25.8	26.9	15.6
Read. par. (ch. 50.0)	**71.8**	67.8	63.9	65.7	59.8

Table 5: Stage I ablation: linear-probe accuracy (%), chance level in parentheses. Each column removes one component from V2. A1: random masking; A2: no VICReg; A3: MAE reconstruction; V1: no gate.

6.7 Each component contributes a distinct capability

Three Stage I ablations remove one component of V2 at a time (Table 5). Random masking (**A1**) primarily costs relation accuracy, identifying SSP as the main contributor to discourse structure. Removing VICReg (**A2**) hits subject identification hardest, confirming its collapse-prevention role. MAE reconstruction (**A3**) sharply reduces relation accuracy while subject ID recovers partially, identifying latent-target prediction as the critical mechanism for semantic abstraction, consistent with the Information Bottleneck argument in §2.

6.8 BLEU is an Unfaithful Proxy in EEG-to-Text

The primary discovery: V1 versus V2 exposes a metric failure. *Every* measure that probes the encoder directly—linear probes, retrieval ACC@1, LLM-judged `Rel@3`, both ETES scores—ranks V2 above V1. The only quantities ranking V1 above V2 are BLEU-1 and BLEU-2. **BLEU is an unfaithful proxy for EEG conditioning**: it ranks models in the opposite direction to every measure of genuine neural grounding. The V1/V2 comparison, differing along a single axis (gated vs. ungated encoder), was designed to expose this, and the noise-input diagnostic proves the mechanism.

The noise diagnostic proves the mechanism. Both variants collapse to near-zero BLEU-1 under noise ($> 99.9\%$), so V1's higher clean-EEG score is genuinely EEG-conditioned, yet it ranks *lower* on every semantic probe. The explanation is decoder fluency: without gating, V1's noisier embeddings cause Flan-T5 to lean on its language prior, producing fluent, lexically generic text that overlaps well with reference sentences but is less faithful to EEG relational content. V2's richer embeddings *constrain* the decoder toward content-accurate but lexically less predictable output. BLEU rewards the fluency; relation accuracy and ETES reward the content.

Practical implication. BLEU's poor correlation with task quality is well documented (Callison-Burch et al., 2006; Reiter, 2018; Mathur et al., 2020); the EEG setting sharpens it because inter-subject variability guarantees lexical divergence from any reference. A BLEU score without a noise-input control is uninterpretable as evidence of EEG decoding; higher BLEU paired with a lower conditional drop Δ is a warning sign of posterior collapse, not a gain. We propose that Δ, LLM-judged relation accuracy, and ETES accompany every BLEU number in this literature.

7 Conclusion

We presented GRAPE and GRAPE-E2T, showing that predicting in latent rather than signal space yields an EEG encoder that recovers discourse-relational structure far above chance with no text supervision, and that this faithfulness propagates into generation, producing a **+16.6 pp** gain in LLM-judged relational accuracy over the prior state of the art. The noise-input diagnostic tells the clearest story: GRAPE-E2T collapses by $>99\%$ under noise while the prior SOTA retains $\sim 70\%$, and we argue that near-total collapse is itself a state-of-the-art result on the faithfulness criterion, a model that does not collapse is not decoding from the EEG. Our secondary finding is a metric one: BLEU inversely tracks EEG faithfulness in this setting, and any score reported without a noise-input control is uninterpretable as evidence of decoding. We introduce ETES as a first step toward evaluation that is sensitive to this distinction. We hope these results shift attention in EEG-to-text research from surface fluency toward genuine neural grounding, a prerequisite for any clinically meaningful brain-computer interface.

Limitations

Cross-subject generation not yet evaluated. Our ZuCo 2.0 cross-subject probe (§6.5) is a paradigm-classification linear probe, not open-vocabulary generation on held-out subjects. To our knowledge no prior open-vocabulary EEG-to-text study reports cross-subject generation either; in our case it was further constrained by the computational cost of full fine-tuning per held-out-subject fold. Cross-subject generation remains the most impor-

tant open evaluation for clinical relevance and is reserved for future work.

Field-level dataset scale. All work on open-vocabulary EEG-to-text decoding is bottlenecked by the scale of available naturalistic-reading EEG corpora: ZuCo contains thousands of sentence-aligned pairs, several orders of magnitude smaller than the corpora typical of language-model pre-training. Our results are therefore best read as evidence of what is achievable at this scale; scaling the JEPA recipe to larger reading-EEG corpora as they become available is a natural next step.

Single-seed reporting. All headline numbers are reported with random seed 42. The magnitude of the noise-input Δ (§6.2) and the linear-probe gaps (Table 1) are sufficiently large that seed variance is unlikely to alter the central claims, but we have not performed a multi-seed sensitivity study. Confidence intervals via bootstrap resampling of the test set could partially address this without re-training; we identify it as a near-term addition.

ETES is model-relative and measures what the model was trained to optimise. Because ETES is computed in the same aligner space the model was trained to optimise, it is susceptible to two related criticisms: it is *model-relative* (cannot directly compare two architectures with different encoders) and *circular* (measures alignment in the very space the contrastive loss shaped, meaning a model that overfits its training objective would score well without genuinely improving generation quality). We accept both limitations and stress that ETES is designed as a *within-model diagnostic and deployment-time gate*, not a cross-model benchmark. The circularity concern is mitigated in practice because (i) ETES is computed on the held-out *test* set, so overfitting to the contrastive loss would not inflate test-time alignment unless the model memorised test sentences; (ii) ETES is always reported alongside the noise-input diagnostic and LLM-judged accuracy, and the three are in agreement, a purely circular score would not co-vary with the independent LLM judge; and (iii) the negative ETES_{gap} result (generations aligning *more* tightly to EEG than the reference) is not an artefact of circularity but a structural property of MTV training, explained in Appendix H.5. In principle ETES could be computed for prior benchmarks within each model's own trained space, but doing so requires re-running every baseline's full encoder and aligner; we leave a multi-model ETES comparison to future work.

LLM self-correction bias in semantic evaluation. We use Llama 3.1-70B Instruct both to generate the multiple text variants (MTV) used as training targets and as the zero-shot judge in our semantic-consistency evaluation (§6.3, Appendix F). Because the same model family generated the targets and scores the outputs, there is a risk of *shared stylistic preference*: the judge may assign higher scores to text that conforms to Llama's generative style, independent of faithfulness to the EEG. We cannot fully eliminate this confound without changing the evaluation model, but two considerations bound its magnitude. First, the gain in LLM-judged top-3 relation accuracy over the prior state of the art is $+16.6$ pp (0.729 vs. 0.563), a margin large enough that a shared stylistic preference would need to be implausibly large to account for it entirely. Second, the same judge rates GLIM, which also uses Llama-generated MTV targets, as *lower* than our model on `Rel@3`, which argues against the bias systematically favouring any model that uses Llama training targets. We report this limitation explicitly and encourage future work to cross-validate with an independent judge (*e.g.*, GPT-4o or a specialised NLI classifier).

Gated-attention artefact-suppression claim is indirect. Our analysis of V2 gate distributions (§6.6) reports sparsity matching reported artefact prevalence in reading EEG. This is suggestive evidence that the gates learn an implicit artefact-rejection function but does not establish it directly: we do not compare against classical artefact rejection pipelines (e.g., ICA, EEGLAB) or against labelled artefact events. A rigorous test would compare against these on a corpus with artefact annotations.

Ethical Considerations

We use only the publicly released ZuCo (Hollenstein et al., 2018) and ZuCo 2.0 (Hollenstein et al., 2020) corpora; no new human-subject data were collected. EEG-to-text decoding has obvious assistive benefits but raises privacy concerns about cognitive data; deployment in clinical or consumer contexts requires informed consent, robust on-device computation, and explicit governance of the read-out content. We believe rigorous, faithfulness-aware evaluation is itself an ethical contribution: inflated scores from posterior-collapsed systems can mis-set expectations for vulnerable patient populations.

References

Hamza Amrani, Daniela Micucci, and Paolo Napoletano. 2024. Deep representation learning for open vocabulary electroencephalography-to-text decoding. *IEEE Journal of Biomedical and Health Informatics*.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of CVPR*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of ICML*.

Adrien Bardes, Jean Ponce, and Yann LeCun. 2022. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *Proceedings of ICLR*.

Niels Birbaumer, Nimr Ghanayim, Thilo Hinterberger, Iver Iversen, Boris Kotchoubey, Andrea Kübler, Juri Perelmouter, Edward Taub, and Herta Flor. 1999. A spelling device for the paralysed. *Nature*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*.

Hubert Cecotti and Axel Gräser. 2010. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):433–445.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M. Sandino, and Joseph Y. Cheng. 2022. MAEEG: Masked auto-encoder for EEG representation learning. In *NeurIPS Workshop on Learning from Time Series for Health*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, and 1 others. 2022. Scaling instruction-finetuned language models.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of ICML*.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, and 1 others. 2024. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems*.

Yiqun Duan, Charles Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. 2023. Dewave: Discrete encoding of EEG waves for EEG to text translation. In *Advances in Neural Information Processing Systems*.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. 2024. EEG2Rep: Enhancing self-supervised EEG representation through informative masked inputs. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Anne-Lise Giraud and David Poeppel. 2012. Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*.

Pierre Guetschel, Thomas Moreau, and Michael Tangermann. 2024. S-JEPA: Towards seamless cross-dataset transfer through dynamic spatial attention. *arXiv:2403.11772*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of CVPR*.

Hassan Hojjati and 1 others. 2025. EEG-VJEPA: Video joint embedding predictive architecture for EEG representations. *arXiv preprint*.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*.

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

Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A. Jäger, and Nicolas Langer. 2023. The ZuCo benchmark on cross-subject reading task classification with EEG and eye-tracking data. *Frontiers in Psychology*, 13:1028824.

Mainak Jas, Denis A. Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. 2017. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. 2024. Are EEG-to-text models working? In *Findings of EMNLP*.

Yann LeCun. 2022. A path towards autonomous machine intelligence. OpenReview preprint.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. 2006. A tutorial on energy-based learning. *Predicting Structured Data*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. 2024. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. *arXiv:2405.02165*.

Xiaotian Liu, Xinyi Shen, and Jiawei Liu. 2025. Learning interpretable representations leads to semantically faithful EEG-to-text generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR*.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. 2018. A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update. *Journal of Neural Engineering*.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of ACL*.

Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka, and Shounak Roychowdhury. 2025. Thought2Text: Text generation from EEG signal using large language models (LLMs). In *Findings of the Association for Computational Linguistics: NAACL*, pages 3747–3759.

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

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2505.06708.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 2000. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*.

Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. 2024. Enhancing EEG-to-text decoding through transferable representations from pre-trained contrastive EEG-text masked autoencoder. In *Proceedings of ACL (Long Papers)*, pages 7278–7292.

Zhenhailong Wang and Heng Ji. 2022. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jinzhao Zhou, Yiqun Duan, Fred Chang, Thomas Do, Yu-Kai Wang, and Chin-Teng Lin. 2024a. BELT-2: Bootstrapping EEG-to-language representation alignment for multi-task brain decoding. *arXiv:2409.00121*.

Jinzhao Zhou, Yiqun Duan, Yingying Zou, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. 2024b. BELT: Bootstrapping electroencephalography-to-language decoding. *arXiv:2309.12056*.

A Dataset and Data Preprocessing

A.1 Dataset composition

Table 6 breaks the combined corpus down by recording group. Each group is a distinct (corpus, paradigm) combination, and the heterogeneity across these groups, in subjects, source text, and the presence or absence of a comprehension task, is what the main text refers to as the distribution shift our encoder must absorb. Annotation coverage is uneven: discourse-relation labels accompany the task-specific Wikipedia groups, sentiment labels accompany only the SST normal-reading group, and the remaining groups are unlabelled. This is the reason the sentiment linear probe (§5.2) sits near chance, its supervision covers a single group. The nine discourse-relation classes are awarding, education, employment, founder, jobtitle, marriage, nationality, political affiliation, and visit.

Source	Para.	Corpus	Labels	Subj.
ZuCo	NR	SST	sentiment	12
ZuCo	NR	Wiki	—	12
ZuCo	TSR	Wiki	relation	12
ZuCo 2.0	NR	Wiki	—	18
ZuCo 2.0	TSR	Wiki	relation	18

Table 6: Composition of the combined corpus by recording group. NR: normal (passive) reading; TSR: task-specific reading with comprehension questions; SST: Stanford Sentiment Treebank; Wiki: Wikipedia relation-extraction subset. All groups use a 128-channel montage at 500 Hz; the dataset totals >22,000 sentence-aligned EEG–text pairs.

A.2 Reading paradigms

The corpora contain two reading paradigms. **Normal Reading (NR)** presents SST and Wikipedia sentences for silent reading at the participant's own pace, closest to ecologically valid reading.

Task-Specific Reading (TSR) presents the same Wikipedia sentences but cues participants to identify a specific entity-relation, inducing top-down attentional control. The NR vs. TSR contrast supports the reading-paradigm probe (§5.2) and the cross-subject probe (Appendix I).

A.3 Preprocessing pipeline

The pipeline operates on the raw ZuCo MAT files and produces tensor-format mini-batches: (1) load the EEG matrix $X \in \mathbb{R}^{T_{\text{raw}} \times C_{\text{raw}}}$ and align to sentence onsets; (2) downsample $500 \rightarrow 128$ Hz with anti-alias filtering; (3) drop channel index 105 (NaN-only across all sessions); (4) zero-pad missing channels to a fixed $C{=}128$ montage; (5) zero-pad the time axis to $T{=}1280$ (a 10 s epoch at 128 Hz, exceeding the longest sentence at our downsampled rate); and (6) persist as HDF5 with sentence-disjoint train/val/test splits to prevent leakage; ZuCo 2.0 held-out subjects are reserved for cross-subject probing.

We deliberately do *not* apply ICA-based artefact rejection, bandpass filtering, or per-subject normalisation. The first two risk removing genuine cortical-language signal in the alpha/beta bands of interest (Giraud and Poeppel, 2012); the third would defeat cross-subject transfer evaluation. The element-wise gate (§3.2) is intended to learn the artefact-suppression function that classical pipelines hard-code.

B Detailed Architecture

Figure 3 provides a complete diagram of the two-stage architecture, with Stage I pretraining (top), the internal gated-self-attention block (middle), and Stage II fine-tuning (bottom). The per-component walk-through is given in the caption; colour coding follows the legend described below.

Throughout the figure, colour coding follows the legend at the top: blue boxes are trainable, green boxes are frozen or EMA-updated, yellow boxes are data/I/O, purple boxes are masking/auxiliary tokens, and red boxes are loss terms. Solid arrows represent forward computation; the dashed red arrow is the EMA update path (no gradient).

C Latent-Space Visualisation

Figure 4 shows t-SNE projections of the frozen GRAPE-V2 patch embeddings on the test set, coloured by four factors: (a) subject identity (30 classes), (b) relation type (9 classes), (c) sentiment (3 classes), and (d) reading paradigm (NR vs. TSR, 2 classes). Strong cluster structure emerges for subject identity and reading paradigm, with partial organisation for relation type, all from EEG-only self-supervision; sentiment shows little separation, consistent with the near-chance linear-probe result (Table 1). We did not tune t-SNE hyperparameters extensively; perplexity is set to 30 and learning rate to 200 following common practice.

D Hyperparameters

Table 7 lists all hyperparameters with their sources or rationales.

Hyperparameter	Value	Source
EEG length T	1280	10 s @ 128 Hz
EEG channels C	128	full ZuCo montage
Patch size P	8	62.5 ms; α/β band
Num. patches N	160	T/P
Embedding d	128	data scale
Encoder blocks	6	I-JEPA
Attention heads	8	standard, $d{=}128$
Predictor depth	2 (x-attn)	(Assran et al., 2023)
EMA momentum m	0.99	I-JEPA
Masking ratio ρ	0.50	(Foumani et al., 2024)
Pretrain epochs	100–200	cosine decay
Pretrain batch size	64	VICReg-stable
Pretrain LR / WD	10^{-4} / 0.01	AdamW
Fine-tune epochs (max)	200	early-stop
Fine-tune LR / WD	10^{-4} / 0	pre-regularised
Mixed precision	bf16	numerical stab.
Random seed	42	main; multi-seed planned

Table 7: Full hyperparameter list across both stages.

E Training Stability (V1 vs. V2)

V1 (no gate) exhibited three large gradient spikes within the first ten epochs and steadily-rising total loss after epoch 28, terminating at 87 epochs. V2 (gated) trained stably for 158 epochs with a single late spike near epoch 140, suggesting that gated attention smooths the optimisation landscape during early SSL training when the EMA target is itself non-stationary. Per-loss component decomposition (alignment / variance / covariance) confirms that the variance and covariance terms remain bounded in both versions; the difference is concentrated in the alignment term. Loss curves are reproducible from the released code.

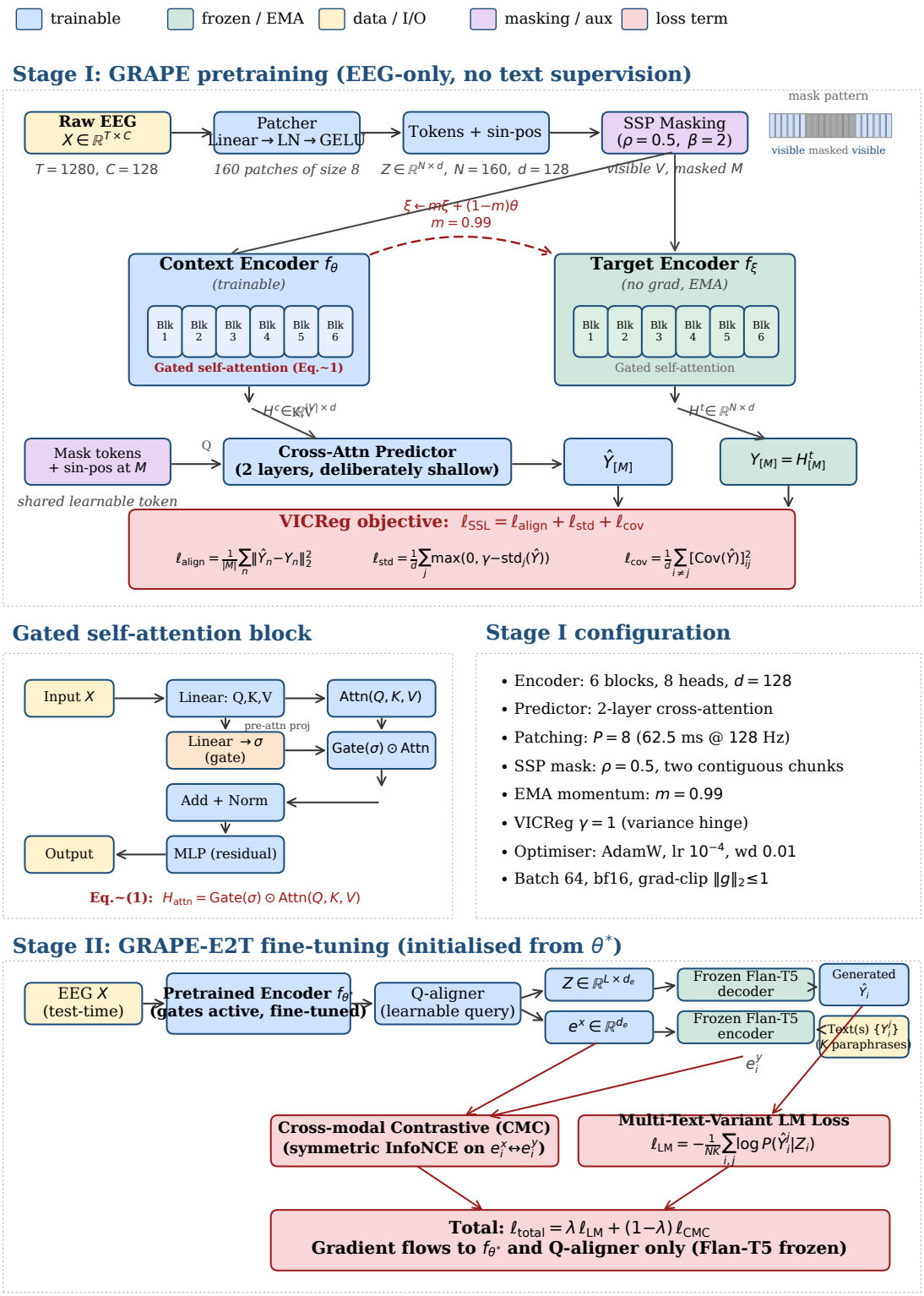

Figure 3: Detailed two-stage architecture of GRAPE / GRAPE-E2T. **Top:** Stage I pretraining: SSP-masked EEG patches feed a context encoder f_θ with element-wise gated self-attention; an EMA-updated target encoder f_ξ produces prediction targets at masked positions; a deliberately shallow cross-attention predictor outputs $\hat{Y}$; the VICReg objective combines alignment, variance-hinge, and covariance terms. **Middle-left:** internal structure of one gated self-attention block (Eq. 1). **Middle-right:** Stage I configuration summary. **Bottom:** Stage II fine-tuning: the pretrained encoder f_{θ^*} (gates active) feeds a Q-aligner that produces a token sequence Z and a global embedding e^x; a frozen Flan-T5 serves as both the text encoder (yielding e^y) and the decoder; the total loss is a convex combination of the multi-text-variant LM loss and the symmetric InfoNCE cross-modal contrastive (CMC) loss; gradient flows only to f_{θ^*} and the Q-aligner.

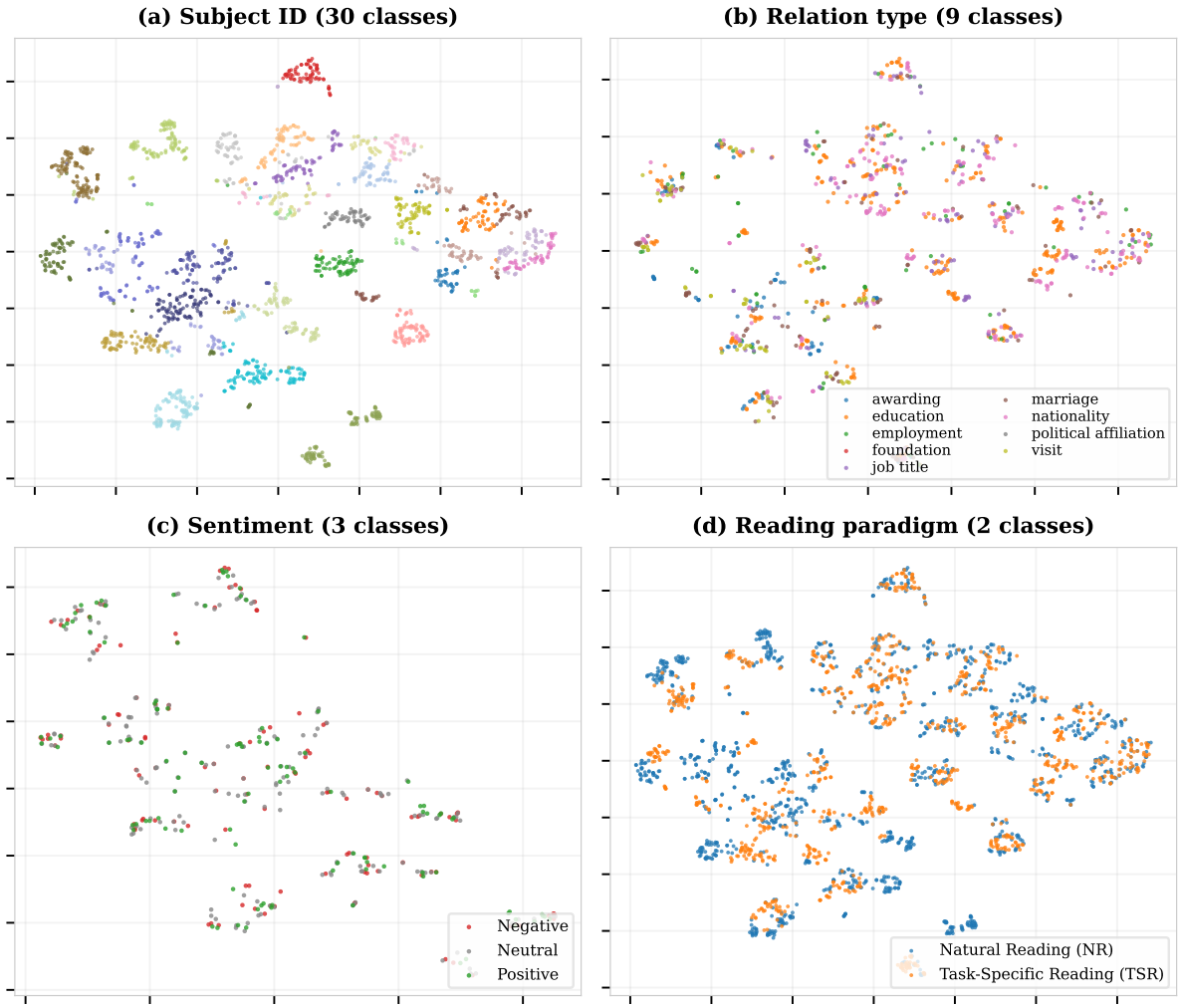

Figure 4: t-SNE of frozen GRAPE-V2 patch embeddings, coloured by (a) subject, (b) relation type, (c) sentiment, and (d) reading paradigm. See text for interpretation.

F Full Semantic-Consistency Table

Table 8 reports the complete five-pathway semantic-consistency analysis described in §6.3. The LLM judge is **Llama 3.1-70B Instruct**, the same model used to generate the multiple text variants (Appendix G), queried zero-shot at temperature 0 with a fixed classification prompt for each semantic axis (sentiment, relation, corpus). Using one model for both variant generation and judging keeps the semantic label space consistent between training targets and evaluation; we report results as a directional, complementary signal rather than a calibrated absolute score.

G Multiple Text Variants

Following Liu et al. (2025), we train against multiple text variants (MTVs) rather than the single raw stimulus sentence. For each stimulus, K paraphrases are generated that preserve core semantics while varying surface form, which reframes decoding as semantic summarisation rather than verbatim reconstruction and reduces overfitting to lexical patterns under limited data. We use the same construction as GLIM: variants are produced by **Llama 3.1-70B Instruct** prompted to rewrite each sentence with diverse syntax and lexicon while preserving meaning. The same model is reused as the zero-shot semantic judge (Appendix F), so generation targets and evaluation share a consistent semantic space. BLEU in the main results is scored against this MTV set (multi-reference), whereas ROUGE-1 is scored against the single raw stimulus (single-reference); the asterisk and double-dagger marks in Table 2 flag this distinction.

Pathway	Model	Sent @1	Rel @1	Rel @3	Corpus
EEG embedding (CLIP-like)	GLIM	**0.427**	**0.325**	**0.571**	**0.935**
	GRAPE-E2T V1	0.285	0.137	0.365	0.412
	GRAPE-E2T V2	0.324	0.153	0.348	0.571
Generated text (LLM judge)	GLIM	0.396	0.072	0.563	0.922
	GRAPE-E2T V1	0.309	**0.078**	**0.729**	0.912
	GRAPE-E2T V2	0.317	0.080	0.646	0.919
Generated text embedding (CLIP-like)	GLIM	0.396	0.207	0.433	**0.874**
	GRAPE-E2T V1	0.288	0.302	**0.483**	0.391
	GRAPE-E2T V2	**0.391**	**0.364**	0.483	0.615
Raw text (LLM judge; ceiling)	GLIM	0.734	0.097	0.755	0.861
	GRAPE-E2T V1	0.734	0.075	0.755	0.864
	GRAPE-E2T V2	0.734	0.077	**0.770**	0.864
Raw text embedding (CLIP-like)	GLIM	**0.456**	0.253	**0.497**	**0.919**
	GRAPE-E2T V1	0.420	**0.379**	0.379	0.507
	GRAPE-E2T V2	0.410	0.296	0.483	0.541

Table 8: Full five-pathway semantic-consistency evaluation. The "raw text" rows function as approximate upper-bounds; differences on these rows across models are dataset-intrinsic.

H The ETES Metric: Detailed Treatment

This appendix gives a substantially more detailed treatment of the EEG–Text Energy Score (ETES) than space permitted in §4 and §6.4.

H.1 Motivation and positioning

Standard text-generation metrics (BLEU, ROUGE, METEOR) measure *lexical overlap* between a generated sentence and a reference. They are insensitive to whether the generated sentence was actually conditioned on the input modality (*e.g.* EEG), a property that posterior-collapse experiments (§6.2) make starkly visible. Embedding-based metrics such as BERTScore measure *semantic similarity between two text strings*, but they still operate text-to-text and cannot distinguish a model that is reading the brain from one that is hallucinating from a language prior.

We propose ETES to fill this gap. ETES is an *EEG-aware* alignment score: measured on the test set, in the model's shared latent space, it quantifies how well a generated (or reference) text aligns with the EEG signal that the model was given. Because the metric is computed in the same space the model was *optimised to align*, it directly probes the quantity the model was trained to maximise, in a manner closely analogous to the energy formulation of LeCun et al. (2006): low energy corresponds to a compatible (well-aligned) input–output pair, high energy to an incompatible one.

H.2 Formal definition

Let $f_{\text{EEG}} : \mathbb{R}^{T\times C} \to \mathbb{R}^{d_e}$ be the EEG embedding map produced by the GRAPE encoder followed by the Q-aligner (§3.3), and let $f_{\text{text}} : V^* \to \mathbb{R}^{d_e}$ be the text embedding map produced by the frozen Flan-T5 encoder with a [CLS]-style pooling. Both maps output unit-normalised vectors. For an EEG X_i paired with a text b, the pairwise energy is

$$E(X_i, b) \;=\; -\cos\big(f_{\text{EEG}}(X_i),\, f_{\text{text}}(b)\big). \quad (11)$$

Energy ranges over $[-1, +1]$; lower is better. We aggregate this energy over a test set of N EEG–text–generation triples $\{(X_i, t_i, \hat{t}_i)\}_{i=1}^{N}$ (where t_i is the reference text and $\hat{t}_i$ the model's generated text) into three scores:

$$\text{ETES}_{\text{align}} = \tfrac{1}{N}\sum_i E(X_i, \hat{t}_i), \quad (12)$$

$$\text{ETES}_{\text{ref}} = \tfrac{1}{N}\sum_i E(X_i, t_i), \quad (13)$$

$$\text{ETES}_{\text{gap}} = \text{ETES}_{\text{align}} - \text{ETES}_{\text{ref}}. \quad (14)$$

Variable-length text. Both t_i and $\hat{t}_i$ may have arbitrary length; f_{text} handles this via the Flan-T5 encoder's sequence-to-vector pooling. We use the standard [CLS]-style first-token pooling; mean-pooling and last-token pooling give qualitatively similar results in pilot experiments but were not used for the reported numbers.

H.3 The three scores: what each diagnoses

$\text{ETES}_{\text{align}}$ is the headline number. It is reference-free in the sense that it does not require the gold

text t_i to compute: only X_i and $\hat{t}_i$. This makes it the natural metric for a deployment-time evaluation in which only the EEG and the generated output are available (*e.g.* BCI inference).

ETES_{ref} functions as a reference-text upper-bound anchor: it is what alignment energy is achievable with the *exact reference sentence as the text*. A sensible expectation is that $\text{ETES}_{\text{align}} \geq \text{ETES}_{\text{ref}}$, *i.e.* the gold text aligns at least as well as any generated paraphrase. We discuss the empirical violation of this expectation in §H.5.

ETES_{gap} measures how far the model's generations deviate (in the shared space) from where the reference sits. Positive gaps indicate generations that are *less* aligned than the reference (the expected case); negative gaps indicate generations that are *more* aligned than the reference (an informative phenomenon we discuss below).

H.4 Interpretation bands

We propose the qualitative bands in Table 9. These thresholds are *heuristic* and chosen so that the band boundaries fall at qualitatively meaningful cosine-similarity levels ($-0.8 \approx$ tight alignment, $-0.5 \approx$ clear alignment well above chance, $-0.2 \approx$ weak but distinguishable alignment, $> 0 \approx$ misalignment). We recommend reporting raw ETES values rather than band labels in publications, but the bands are useful as a sanity check.

$\text{ETES}_{\text{align}}$	Interpretation
< -0.8	Excellent alignment
< -0.5	Good alignment
< -0.2	Fair alignment
> 0	Poor alignment (dissimilar embeddings)

Table 9: Qualitative bands for the alignment energy $\text{ETES}_{\text{align}}$. These are heuristic; we recommend reporting raw values in papers.

H.5 Why the gap can be negative

Our experiments report a negative ETES_{gap} for both V1 and V2 (Table 4). At first reading this is counter-intuitive: the model's generated paraphrase aligns *more closely* to the EEG than the reference sentence *itself*. There is no contradiction here:

1. The reference t_i is a single surface realisation of the neural content. Many paraphrases of it could be equally valid renderings, and some may be *more* natural in the embedding space the model has shaped during alignment training.
2. Multi-text-variant (MTV) training (Liu et al., 2025) explicitly encourages the model to associate a single EEG with multiple paraphrases. The model's converged generations live somewhere in the convex hull of those paraphrases, which can be closer to the EEG embedding than any single reference.
3. The shared space is shaped by the contrastive cross-modal loss; the reference text was never seen by the contrastive objective in isolation, only as one of K variants.

We read negative ETES_{gap} as evidence of *EEG-grounded paraphrasing*: the model is generating in the shared latent space, not transcribing in surface space. This is consistent with the dissociation between low BLEU and high LLM-judged `Rel@3` (§6.3).

H.6 Reference-free use

$\text{ETES}_{\text{align}}$ does not require t_i. This is operationally important for two scenarios.

Inference-time evaluation. In any real BCI deployment, the gold sentence is unknown; the only signals available are the EEG and whatever text the model produced. $\text{ETES}_{\text{align}}$ provides a per-utterance quality score in exactly this setting. By contrast, BLEU, ROUGE, and ETES_{ref} all require t_i and cannot be computed.

Self-monitoring. A confident system could use $\text{ETES}_{\text{align}}$ as a runtime gate: if the energy on a particular utterance exceeds a learnt threshold, the system can refuse to output rather than hallucinate. We do not evaluate this use case here but flag it as a natural next step.

H.7 Limitations

ETES has three concrete limitations that users should be aware of.

Model-relative. ETES is measured in *this model's* aligner space. It cannot be used to directly compare two architectures with different encoders, because each architecture has its own latent geometry. To compare across models one needs either a shared external embedder (in which case ETES collapses into a BERTScore-style text-vs-text metric), or to compute ETES separately within each model and compare *rankings* rather than raw values.

Heuristic bands. The interpretation bands in Table 9 are not derived from any calibration study;

they are chosen by inspection. A proper calibration would require a curated test set with graded alignment judgements (high, medium, low), which does not currently exist for EEG-to-text.

Possible BLEU correlation. Because the shared embedding space is itself shaped by the cross-modal contrastive loss, it is possible that ETES correlates with BLEU on some slice of the data even when the central diagnostic distinguishes them on other slices. We report ETES *alongside* BLEU rather than in place of it, and recommend that future work do the same.

I Cross-Subject Probing Protocol and Full Results

Protocol. The cross-subject probe uses the ZuCo 2.0 NR-vs-TSR reading paradigm from Hollenstein et al. (2020, 2023). Inputs are 128-channel EEG-only (no eye-tracking, no text features). Held-out subjects are not seen during GRAPE pretraining or GRAPE-E2T fine-tuning. We freeze the encoder, extract the global embedding e^x, and train a logistic-regression probe with L_2 regularisation (λ=10^{-3}). Reported numbers are mean accuracy over held-out subjects.

Full results. Table 10 reports all baselines from the Hollenstein et al. (2023) benchmark alongside our frozen and fine-tuned variants.

Method	Modality	Acc. (%)
Random baseline	–	50.0
BERT (text-dependent)	Text only	53.0
Eye-tracking features	Eye-tracking	65.0
Eye-tracking + EEG	Multi-modal	69.0
Concatenated EEG features	EEG only	55.0
Concatenated EEG + PCA	EEG only	**66.0**
GRAPE-V2 (frozen, ours)	EEG only	52.3
GRAPE-E2T-V2 (fine-tuned, ours)	EEG only	55.6

Table 10: Cross-subject reading paradigm classification on ZuCo 2.0 held-out subjects (Hollenstein et al., 2023).

Honest interpretation. Our learned EEG representations are *competitive with* concatenated hand-crafted electrode features (55.6 vs. 55.0) while requiring no manual feature engineering, but *do not exceed* feature-engineered baselines that apply PCA dimensionality reduction (66.0%). Multi-modal eye-tracking baselines (65–69%) remain higher but require precise gaze hardware not available in a purely neural BCI deployment. The fine-tuning gain (+3.3 pp[1] over the frozen probe) suggests that text-side alignment functions as a cross-subject regulariser by mapping disparate brain signatures onto a shared linguistic manifold. Closing the gap to PCA-engineered EEG features while maintaining the end-to-end learning paradigm is a clear direction for future work.

J Gate-Level Diagnostics

Table 11 reports the gate-level summary statistics discussed in §6.6, complementing the distribution plot in Figure 2.

Gate metric	V1	V2	Δ
Entropy	0.988	0.997	+0.009
Mean	0.472	0.444	−0.028
Sparsity	0.136	0.144	+0.007
Std	0.258	0.267	+0.009

Table 11: Gate-level diagnostics on the ZuCo test set (V1 = no gate, V2 = gated).

K Reproducibility

Code, preprocessed dataset manifests, training scripts, and checkpoints will be released under a permissive licence on acceptance. All runs use a fixed random seed (seed 42); the magnitude of the noise-input Δ (§6.2) and linear-probe results (Table 1) are sufficiently large that seed variance is unlikely to materially affect the central claims of the paper.

L Computational Budget

Pretraining of the GRAPE encoder took approximately 4 GPU-hours on a single RTX 6000 Pro (NVIDIA, via RunPod cloud infrastructure). Fine-tuning of GRAPE-E2T was performed on a single RTX 5040 GPU, with GRAPE-V1 requiring 33 hours and GRAPE-V2 requiring 18 hours. The faster convergence of V2 reflects the improved optimisation landscape afforded by element-wise gated self-attention (Appendix E). All experiments used mixed-precision training (bf16) to reduce memory overhead and accelerate computation.

[1] "pp" denotes percentage points, the absolute difference between two percentages on the same scale.