arXiv BraVL Dataset: Zero-shot EEG → Image → Text Decoding with Simple Linear Models

Benjamin Wagstaff
Department of Computer Science
University of Durham
benjamin.wagstaff@durham.co.uk

Abstract

Decoding human visual neural representations is problematic because EEG data contains strong subject-specific structure and weakly expressed semantic signals. While the original work addresses this using deep cross-modal models, I re-analyse the publicly released features using only transparent linear methods (Ridge regression, LDA, CCA, k-means) to characterise the EEG space and assess strict zero-shot decoding. After removing subject baselines and reshaping EEG representations to suppress subject-driven variance, cross-modal alignment becomes feasible.

An optimised cluster-aware zero-shot EEG→Image decoder trained solely on seen concepts and tested on unseen concepts attains $\approx 21\%$ top–1 accuracy when ranking all 1,854 candidate concepts, substantially outperforming naïve linear baselines. Separately, a CCA mapping from image to text space achieves $\approx 91\%$ top–1 accuracy, identifying EEG→Image alignment as the primary bottleneck in semantic decoding. Combined end-to-end, the linear EEG→Image→Text pipeline reaches $\approx 23\%$ top–1 and $\approx 33\%$ top–10 accuracy under strict zero-shot evaluation.

I also report negative findings, such as clustering raw EEG features and applying CCA without supervised LDA, which demonstrate that subject-specific variance dominates unprocessed EEG and must be explicitly suppressed before linear zero-shot transfer becomes feasible. These results show that while linear models are limited, manipulating the EEG representation so that concept structure becomes linearly expressible is critical for enabling non-trivial zero-shot semantic decoding.

1 Introduction and related work

Problem. BraVL [Du et al., 2023] pairs EEG responses with image and text features for 1,854 THINGS concepts [Hebart et al., 2019], enabling large-scale study of semantic brain decoding. The central challenge is *cross-concept generalisation*: decoding the semantic concept from visually evoked EEG responses when that concept was never observed during training. While large text corpora provide rich semantic representations, corresponding mappings from EEG to semantic space are sparse, noisy, and dominated by subject-specific structure.

Traditional ML vs. advanced paradigms. A conventional supervised split (e.g. random within-label partitions) primarily measures *interpolation* within a fixed label set and can substantially overestimate generalisation when subject structure or near-duplicate trials appear in both training and test data. In contrast, *zero-shot* learning evaluates transfer to disjoint classes by exploiting shared semantic representations, enabling prediction or retrieval for categories unseen during training. This paradigm has been widely studied in vision and retrieval settings, where it supports scalable generalisation beyond fixed label vocabularies [Long et al., 2017, 2018]. Zero-shot decoding therefore provides a more appropriate framework for semantic EEG decoding, where new concepts continually arise and labelled neural data is expensive to obtain.

Paradigm used here. I adopt a strict zero-shot EEG→Text setting in which all mappings are trained exclusively on *seen* concepts (and their EEG, image, and text features), while evaluation is performed only on *unseen* concepts by ranking all 1,854 labels. This protocol prevents concept leakage and directly tests cross-concept transfer, even when unseen concepts are semantically related to those observed during training.

Positioning and related work. BraVL [Du et al., 2023] trains deep multimodal models to align brain, vision, and language within a shared latent space, leveraging deep learning to learn flexible, high-capacity representations across modalities. More broadly, deep learning has been central to recent advances in representation learning by enabling hierarchical feature extraction and task-adaptive embeddings [Duan et al., 2023]. Here, I pose a complementary question: how far can transparent linear structure go when applied to the released features? Prior large-scale visual EEG studies show that EEG representations are strongly dominated by subject-specific variance and that careful preprocessing is necessary to obtain stable and generalisable decoding performance [Gifford et al., 2022]. Motivated by this, I construct a modular linear pipeline and report both positive and negative findings to clarify which representation-shaping steps are necessary for zero-shot transfer.

Contributions. I provide (1) a strict zero-shot evaluation protocol that minimises subject and concept leakage, (2) a compact linear EEG→Image→Text pipeline with strong retrieval performance, and (3) an ablation-style analysis identifying which components are required for cross-concept transfer on BraVL.

2 Dataset and paradigm

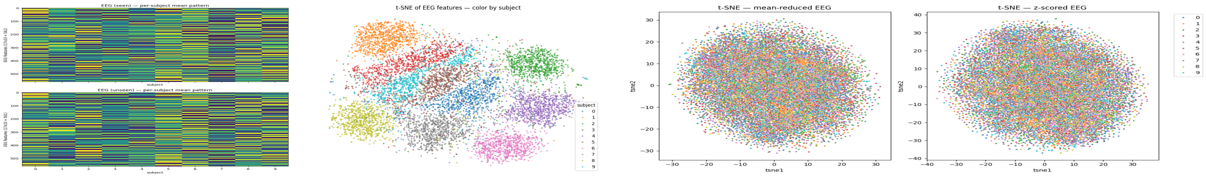
The BraVL dataset [Du et al., 2023] provides high-density EEG alongside image and text embeddings for 1,854 object concepts from the THINGS database [Hebart et al., 2019], enabling large-scale multimodal brain decoding. For each of 10 subjects the released data consists of EEG, image features, and text features. I use the following representations:

Modality	Released features	Used in this report
EEG	17 × 100 (0–800ms, 100Hz)	keep 70–400ms (33 timepoints) ⇒ 561-D; subject-mean removed; ERP-weighted
Image	1000-D PCA (CORnet-S)	first 100 (100-D)
Text	512-D CLIP-Text	as released (512-D)

The stimuli are split into *seen* (train) and *unseen* (zero-shot test) concepts so that per-subject data is:

- **Seen:** 1,654 concepts, 10 images per concept; one trial corresponds to the average of 4 EEG trials per image, giving $1{,}654 \times 10 = 16{,}540$ image–EEG pairs per subject.
- **Unseen:** 200 concepts, 1 image per concept, 80 EEG trials per image, giving 16,000 unseen trials per subject.

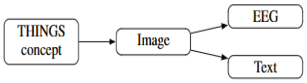
Across subjects the class counts are balanced by design (fixed images and trials per concept). In practice, unseen trials exhibit higher within-concept variance than seen trials, making cross-concept transfer harder.



(a) EEG mean per subject. (b) *t*-SNE of EEG (per subject). (c) *t*-SNE after subject-mean removal.

Figure 1: EEG inspection suggests subject identity dominates raw geometry; mean removal mitigates this.

The strongest source of variation is subject identity rather than concept identity (Figure 1b); subtracting each subject's mean reduces this separation (Figure 1c), variance division is not needed. I therefore apply per-subject mean subtraction $\tilde{x}_{s,i} = x_{s,i} - \frac{1}{N_s}\sum_j x_{s,j}$, where s indexes subjects and i trials, and I remove even time-point indices having noticed downsampling information loss, giving a 17 × 16 = 272-D EEG vector.



Paradigm: strict zero-shot EEG→text. I train only on seen concepts and evaluate on unseen concepts, ranking *all 1,854* labels. This mirrors scenarios where new categories appear but labelled EEG is costly. A random 70/30 split mainly measures interpolation within the same label set (and may partially reflect subject structure), whereas strict zero-shot measures transfer to a disjoint concept set, even when unseen concepts are semantically related to seen ones.

3 Baseline model

As a lower bound, I use a minimal linear pipeline matching the final structure but without ERP weighting, LDA, CCA, or clustering. Seen concept prototypes are used to fit Ridge maps EEG→IMG and IMG→TXT; unseen EEG trials are ranked against all 1,854 text prototypes (cosine similarity). Put simply:

$$\hat{y}_{\text{img}} = W_1 x_{\text{EEG}}, \qquad \hat{y}_{\text{text}} = W_2 \hat{y}_{\text{img}}$$

Baseline retrieval: Top–1 $= 0.0004$, *Top–10* $= 0.0037$, *mean rank* $= 839.4$ *(chance* ≈ 927.5*), MRR* $= 0.0040$.

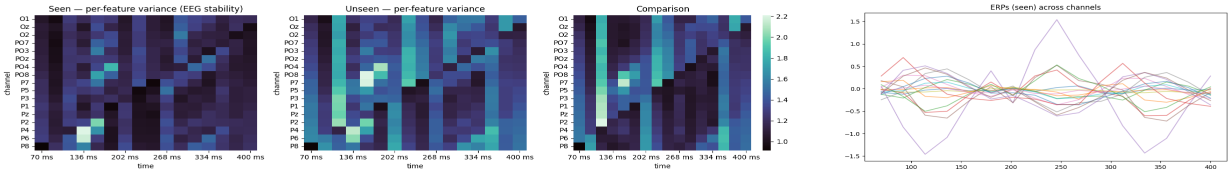
4 Specialised model

4.1 EEG preprocessing and descriptive analyses

To emphasise time–channel regions with strong evoked responses, I compute a grand-average ERP from $\tilde{X}$ and derive a normalised weight vector w:

$$E = \frac{1}{N}\sum_i \tilde{X}_i, \quad w = \frac{|\text{vec}(E)|}{\text{mean}(|\text{vec}(E)|)}, \quad x^{(\text{ERP})} = w \odot \tilde{x}.$$

Mean and variance across concepts. Figure 2 shows ERP and variance structure after subject-mean subtraction. Seen and unseen ERPs are qualitatively similar, while variance patterns reveal a dominant global temporal envelope -well approximated by a rank-1 factor across time, shared across channels. This global envelope motivates the use of ERP-derived weighting to emphasise time–channel regions with consistent evoked activity. Mean subtraction alone (without variance normalisation) yields more coherent ERP structure and improves downstream decoding. Naively removing the dominant rank-1 temporal component was found to discard semantically meaningful signal rather than isolating noise.



(a) Variance decomposition across factors (dominant rank-1 temporal component). (b) ERP-based channel–time variance.

Figure 2: ERP and variance structure after subject-mean subtraction and ERP weighting.

4.2 Linear models: LDA and CCA

4.2.1 LDA in EEG space

Given ERP-weighted EEG features $x^{(\mathrm{ERP})}$ and seen concept labels $y \in \{1, \dots, C\}$ with $C = 1{,}654$, I apply Linear Discriminant Analysis (LDA) to learn a projection $W \in \mathbb{R}^{272 \times d}$ that maximises between-class scatter relative to within-class scatter:

$$S_B w = \lambda\, S_W w.$$

I set $d = 100$ to avoid overfitting and ensure numerically stable within-class covariance estimation, and fit on seen trials.

4.2.2 CCA between EEG and image prototypes

From LDA-projected EEG prototypes $P^{\mathrm{EEG}}_{\mathrm{seen}} \in \mathbb{R}^{1654 \times 100}$ and image prototypes $P^{\mathrm{IMG}}_{\mathrm{seen}} \in \mathbb{R}^{1654 \times 1000}$, Canonical Correlation Analysis (CCA) finds projections : $u = A^{\top} p^{\mathrm{EEG}}, \quad v = B^{\top} p^{\mathrm{IMG}}$, maximising $\mathrm{corr}(u_k, v_k)$ per canonical dimension under orthogonality constraints. This leads to the standard CCA eigenvalue problem:

$$\Sigma_{EE}^{-1} \Sigma_{EI} \Sigma_{II}^{-1} \Sigma_{IE} a = \rho^2 a.$$

CCA is fit on *standardised* seen concept prototypes (feature-wise z-scoring via `StandardScaler`) and applied to all prototypes using the same scalers; I keep up to 50 canonical dimensions.

Training protocol. All mappings are fit on the full seen set; unseen concepts are reserved exclusively for zero-shot evaluation, so no additional validation split is used. I set LDA to $d = 100$ and CCA to 50 components. Cluster-aware retrieval parameters are selected from a sweep over $k \in [5, 200]$ and $M \in \{1, 2, 3, 4\}$ using zero-shot top–1 and mean-rank performance.

4.3 Cluster-aware retrieval in shared space

In the EEG–image CCA space, I cluster seen EEG prototypes using k-means with $k = 170$ clusters and restrict retrieval for each test point to its $M = 3$ nearest clusters:

1. Find the 3 nearest EEG clusters to an unseen EEG point.
2. Restrict candidate image prototypes to those whose EEG prototypes fall in these clusters.
3. Rank candidates by cosine similarity within the restricted set.

This narrows the effective candidate set and exploits local EEG–image structure.

4.3.1 Image→Text mapping

Given image prototypes $P^{\mathrm{IMG}}_{\mathrm{seen}}$ and text prototypes $P^{\mathrm{TXT}}_{\mathrm{seen}}$ (CLIP-Text), I first fit Ridge regression $\hat{t} = W_{\mathrm{ridge}} p^{\mathrm{IMG}}$, trained on seen concepts and evaluated against all 1,854 text prototypes using cosine similarity. This yields $\approx 44.6\%$ top–1 and $\approx 86\%$ top–10. I then fit CCA between image and text prototypes (seen only) and use the shared space for retrieval, using this mapping as the second stage in the full EEG→Image→Text pipeline.

- **Image→Text CCA (seen→all):** top–1 ≈ 0.91, top–10 ≈ 0.99, mean rank ≈ 1.4, MRR ≈ 0.95.

5 Result analysis

I evaluate performance under the baseline ridge pipeline and my specialised variant incorporating ERP-weighted EEG, LDA projection, CCA alignment, and cluster-aware retrieval. In a strict zero-shot setting, I report retrieval metrics: top-k accuracy, mean rank, and mean reciprocal rank (MRR), which are well suited to 1,854-way ranking. Precision/recall/F1 are not directly applicable because the model outputs rankings rather than class probabilities. As all components are linear or closed-form, optimisation is characterised via pipeline architecture and structural sweeps (e.g. cluster granularity and candidate narrowing) rather than gradient-based learning curves.

Table 1: Summary of main model variants (zero-shot on unseen concepts, retrieval over all 1,854 labels; metrics averaged over unseen concepts using multiple averaged EEG queries per concept).

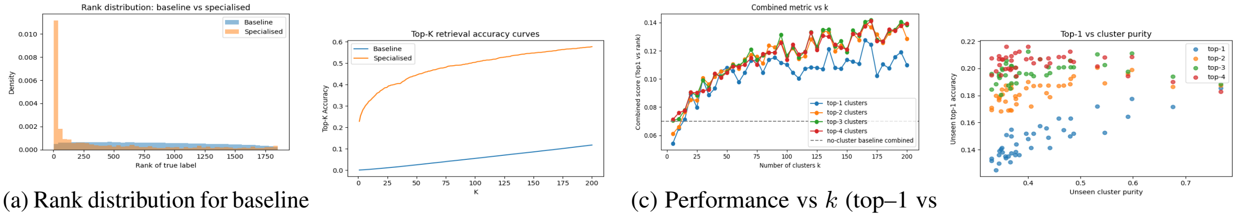
Model	Top-1	Top-10	Mean rank	MRR
Baseline EEG→IMG→TXT (Ridge stack)	0.0004	0.0037	839.4	0.0040
EEG→IMG, ERP+LDA+CCA (no clustering)	0.1805	0.4470	79.1	0.2715
EEG→IMG, ERP+LDA+CCA, k=170, top-3 clusters	0.2060	0.4545	22.7	0.2857
IMG→TXT (CCA, seen→all)	0.9148	0.9946	1.39	0.947
Full EEG→IMG (cluster-aware)→TXT (CCA)	0.2280	0.3250	390.2	0.2631

Ablation summary. Ablation reveals a strict dependency chain: ERP weighting stabilises evoked structure, supervised LDA injects class-level separability, CCA enforces cross-modal geometry, and clustering reduces effective search complexity. Removing LDA produces the largest drop in top–1 and mean-rank performance.

Baseline performance. The baseline EEG→Image→Text yields extremely low top–1 accuracy and a near-uniform rank distribution (Figure 3a), indicating that naive linear mapping fails to recover discriminative cross-modal structure. This establishes that performance gains are not attributable to dataset bias or evaluation artefacts.

Specialised model performance. Incorporating ERP weighting, LDA, CCA alignment, and cluster-aware retrieval produces a substantial shift in the first mapping: top–1 exceeds 20%, and mean rank improves from ~ 840 to ~ 23. Figure 3b shows improvements across the top-k spectrum, suggesting a globally better-aligned embedding rather than a local correction.

Cluster-aware retrieval and purity. Increasing the number of clusters reduces mean rank and improves top–1 up to an optimal region near 120–170 clusters (Figure 3c). Performance saturates and then degrades beyond this region, indicating an intrinsic trade-off rather than monotonic tuning. Cluster purity correlates positively with unseen-class accuracy (Figure 3d), supporting that performance gains follow meaningful structure rather than arbitrary partitioning.



(a) Rank distribution for baseline vs specialised.

(b) Top-k accuracy curves.

(c) Performance vs k (top–1 vs rank).

(d) Top–1 vs unseen purity.

6 Discussion

This report revisits BraVL [Du et al., 2023] using a compact set of linear tools (Ridge, LDA, CCA, k-means). Three themes emerge:

- **Subject and ERP structure dominate raw EEG.** Prior to preprocessing, variance is driven by subject identity and a global temporal envelope, leaving limited concept-specific geometry.
- **Transformations are essential for zero-shot.** Raw EEG yields low semantic structure; supervised LDA on seen classes produces a space more amenable to cross-modal alignment.
- **CCA provides a usable shared geometry once representations are conditioned.** CCA between LDA–EEG and image prototypes yields a coherent latent space where retrieval improves; cluster-aware narrowing further reduces the effective search difficulty.
- **Image→Text is easy; EEG→Image is the bottleneck.** CCA maps image features to text with ≈91% top–1 accuracy, indicating that the primary difficulty lies in mapping EEG responses to visual image representations.

Limitations and future work. Working from precomputed features limits exploration of alternative preprocessing or ROIs. All mappings are linear; modest non-linearities applied after representation conditioning (e.g., kernel CCA or shallow MLPs) may improve alignment without sacrificing interpretability. Error stratification by semantic similarity [Hebart et al., 2019] could clarify where zero-shot decoding succeeds or fails.

Compared to a mean-reduced Ridge baseline that behaves almost randomly, my specialised pipeline recovers substantial semantic structure from EEG alone. These results show that, once subject structure is controlled and representations are properly conditioned, linear methods suffice to support non-trivial zero-shot semantic decoding from EEG.

References

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023. doi:10.1109/TPAMI.2023.3263181.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):1–24, 10 2019. doi:10.1371/journal.pone.0223792. URL `https://doi.org/10.1371/journal.pone.0223792`.

Yang Long, Li Liu, Fumin Shen, Ling Shao, and Xuelong Li. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2498–2512, 2017. doi:10.1109/TPAMI.2017.2762295.

Yang Long, Li Liu, Yuming Shen, and Ling Shao. Towards affordable semantic searching: Zero-shot retrieval via dominant attributes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi:10.1609/aaai.v32i1.12280. URL `https://ojs.aaai.org/index.php/AAAI/article/view/12280`.

Haoran Duan, Yang Long, Shidong Wang, Haofeng Zhang, Chris G. Willcocks, and Ling Shao. Dynamic unary convolution in transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12747–12759, 2023. doi:10.1109/TPAMI.2022.3233482.

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. ISSN 1053-8119. doi:https://doi.org/10.1016/j.neuroimage.2022.119754. URL `https://www.sciencedirect.com/science/article/pii/S1053811922008758`.