GENERALISED ZERO-SHOT BRAIN DECODING VIA EEG–TEXT CLIP ALIGNMENT AND GENERATIVE EMBEDDING SYNTHESIS

Muhammad Ariz b. Muhammad Akmal
Department of Computer Science
Durham University
Durham, UK DH1 3LE
pgbh35@durham.ac.uk

December 16, 2025

ABSTRACT

Most existing EEG-based classifiers are trained in a fully supervised manner on fixed sets of labelled categories, limiting their ability to recognise novel concepts at test time and making them expensive to scale to rich semantic vocabularies. This report investigates a Generalised Zero-Shot Learning (GZSL) paradigm for EEG-based brain decoding, where the model must jointly classify *seen* categories with EEG training data and *unseen* categories with no EEG examples. This paradigm is complimented with a modified EEG-Text CLIP process instead of novel methods such as attribute-based or language-model GZSL. The challenge with GZSL, however are the biasness of the classifer towards *seen* classes and domain shifts. Hence, I aim to synthesise EEG training data for novel classes using a conditional Wasserstein GAN with gradient penalty (cWGAN-GP). The semantic space produced from the EEG and text encoder from CLIP will be the operating space for the cWGAN-GP, namely where the critic samples real EEG embeddings from. This approach brings forth an alternative to other methods of Generative ZSL, that improves accuracy of classifiers on brain decoding tasks.

1 Introduction

Decoding human visual perception from brain activity is an increasingly important problem in both neuroscience and artificial intelligence. Classical machine learning pipelines train discriminative models directly on labelled EEG features (e.g. logistic regression or SVM), assuming that all test-time categories are observed during training and that sufficient labelled data exist for each class. These assumptions hinder scalability to large vocabularies and prevent recognition of novel categories without costly retraining.

Advanced machine learning paradigms relax these constraints by exploiting *semantic side information*. ZSL and GZSL extend conventional supervision by conditioning on attributes or language descriptions, enabling models to infer classes with no task-specific training examples while still maintaining performance on seen categories.

2 Related Work

2.1 Zero-Shot and Generative Zero-Shot Learning

Zero-shot learning (ZSL) addresses the problem of recognising unseen classes by embedding both seen and unseen categories into a shared semantic space, for example using attribute vectors or textual descriptions. Early approaches focused on learning a compatibility function between visual features and semantic prototypes, but often suffered from bias towards seen classes in the more realistic GZSL setting. To mitigate this, Long et al. [1] proposed synthesising unseen visual features with diffusion regularisation, effectively converting ZSL into a supervised learning problem in feature space. The same authors later extended ZSL ideas to retrieval, showing that semantic attributes can support efficient instance-level search without labelled examples for every category.

3 Data and Paradigm

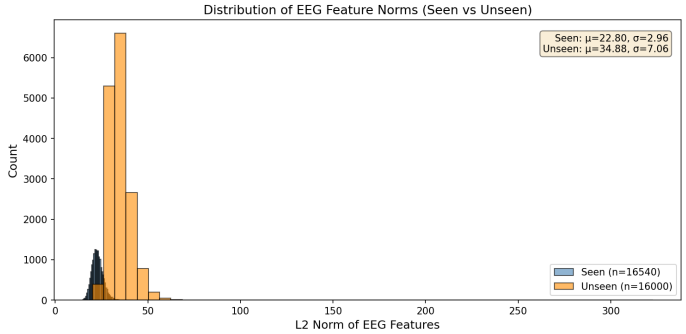

Figure 1: Empirical distribution of ℓ_2 norms of EEG feature vectors for seen and unseen trials, illustrating a distribution shift in scale and variability.

Table 1: Summary of the data splits and their roles in the hybrid GZSL paradigm.

Split	#Trials	EEG source	Classes	Used for
Train–seen	13232	b from $\mathcal{S}$	$\mathcal{S}$	Baseline LR, CLIP encoder, cWGAN-GP, GZSL classifier
Test–seen	3308	b from $\mathcal{S}$	$\mathcal{S}$	Supervised and GZSL evaluation on seen classes
Test–unseen	16000	b from $\mathcal{U}$	$\mathcal{U}$	GZSL evaluation on unseen classes

Data splitting and paradigm declaration. The experiments are based on the BraVL dataset where each trial provides an EEG feature vector $b \in \mathbb{R}^{D_b}$ recorded while a participant views an image, together with a class label $y \in \mathcal{C}$ and a precomputed text embedding $t_y \in \mathbb{R}^{D_t}$ derived from the corresponding category name or description. Table 1 summarises the splits used in the experiments. In this subset, the split corresponds to $|\mathcal{S}| = 1654$ seen classes with $n_{\text{seen}} = 16540$ trials (10 trials/class) and $|\mathcal{U}| = 200$ unseen classes with $n_{\text{unseen}} = 16000$ trials (80 trials/class), where unseen EEG is held out entirely for evaluation under the GZSL protocol (testing) to avoid leakage into the generative model or classifier training.

Let $\mathcal{C}$ denote the set of all categories provided by the data given in the Colab code. Following the GZSL setting, we partition the label space into disjoint *seen* and *unseen* sets,

$$\mathcal{C} = \mathcal{S} \cup \mathcal{U}, \qquad \mathcal{S} \cap \mathcal{U} = \varnothing,$$

where EEG trials from $\mathcal{S}$ are available for training while classes in $\mathcal{U}$ have no EEG examples during training. For each seen-class trial we observe (b_n, y_n) with $y_n \in \mathcal{S}$, together with a text embedding t_{y_n}; for unseen classes only the text embeddings $\{t_u : u \in \mathcal{U}\}$ are available.

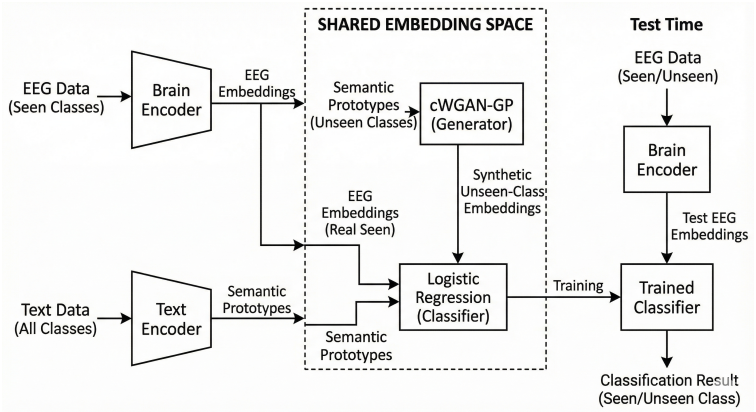

Figure 2: Schematic overview of the hybrid paradigm combining EEG–text CLIP alignment, cWGAN-GP embedding synthesis, and GZSL classification.

This project adopts a hybrid paradigm that combines ***Generalised Zero-Shot Learning***, ***Contrastive Learning***, and ***Adversarial Learning*** illustrated in Figure 2. First, a CLIP-style brain–text encoder learns a shared semantic space in which EEG embeddings $e = f_b(b)$ and text embeddings $s_y = g(t_y)$ lie in a common space $\mathbb{R}^d$ with $\|e\|_2 = \|s_y\|_2 = 1$. Second, a conditional WGAN-GP (cWGAN-GP) learns a class-conditional generator G that synthesises EEG-like embeddings for unseen classes given only their semantic prototypes s_u. Finally, a multinomial logistic regression classifier is trained in this embedding space on a mixture of real seen-class embeddings and synthetic unseen-class embeddings and is evaluated jointly on $\mathcal{S}$ and $\mathcal{U}$ under a GZSL protocol.

Paradigm justification and data exploration. This hybrid paradigm shows that we can build EEG-based decoders that recognise new, never-seen concepts purely from a text description. This novelty brings upon new opportunities such as faster adaptation of neural UX systems (e.g. assistive communication) to new commands, objects or categories alongside massively reducing annotation and preprocessing efforts. Figure 1 provides further insight into the challenges of the BraVL subset. The mismatch in scale and spread suggests a noticeable *distribution shift* in the raw EEG feature space between seen and unseen splits. A GZSL classifier trained directly on raw EEG is implicitly exposed to a covariate shift that can bias predictions towards seen classes.

4 Model Development

Baseline. The baseline is a multinomial logistic regression (LR) model trained on raw EEG features from seen classes only. Given (b_i, y_i) with $y_i \in \mathcal{S}$, LR minimises the regularised empirical risk

$$\min_{W,\beta} \frac{1}{N}\sum_{i=1}^{N}\Big(-\log p(y_i \mid b_i; W, \beta)\Big) + \frac{\lambda}{2}\|W\|_F^2, \qquad p(y \mid b; W, \beta) = \frac{\exp\big(w_y^\top b + \beta_y\big)}{\sum_{c\in\mathcal{S}}\exp(w_c^\top b + \beta_c)}. \tag{1}$$

A naïve GZSL evaluation can be obtained by applying this seen-only classifier to the joint test set; however, predictions are restricted to $\mathcal{S}$ and therefore provide near-zero recognition on $\mathcal{U}$, motivating the proposed hybrid improvement.

Improvements. Inspired by CLIP-style contrastive learning and its adaptation to EEG–language alignment [2], we learn two projection functions into a shared embedding space $\mathbb{R}^d$:

$$f_b : \mathbb{R}^{D_b} \to \mathbb{R}^d, \quad g : \mathbb{R}^{D_t} \to \mathbb{R}^d, \qquad e_i = \frac{f_b(b_i)}{\|f_b(b_i)\|_2}, \quad s_i = \frac{g(t_{y_i})}{\|g(t_{y_i})\|_2}. \tag{2}$$

The L2 normalisation places embeddings on the unit sphere, making cosine similarity equivalent to the dot product $e_i^\top s_j$. For a mini-batch of size N, define logits

$$\ell_{ij} = \frac{1}{\tau} e_i^\top s_j, \tag{3}$$

where $\tau > 0$ is a temperature. We optimise a symmetric InfoNCE objective:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\Big(\mathcal{L}_{b\to t} + \mathcal{L}_{t\to b}\Big), \quad \mathcal{L}_{b\to t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\ell_{ii})}{\sum_{j=1}^{N}\exp(\ell_{ij})}, \quad \mathcal{L}_{t\to b} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\ell_{ii})}{\sum_{j=1}^{N}\exp(\ell_{ji})}. \tag{4}$$

Hyperparameters: $d = 64$, $\tau = 0.07$, epochs = 20, $B = 256$, $\eta = 10^{-3}$

Theoretical expectation: if contrastive alignment succeeds, then matched pairs should have larger cosine similarity than mismatched pairs, i.e. $\mathbb{E}[e^\top s_y] > \mathbb{E}[e^\top s_{y'}]$ for $y' \neq y$. This induces a geometry where text-derived prototypes can act as semantic anchors for recognition beyond the seen label set.

After training, we compute a semantic prototype for each class $c \in \mathcal{C}$ by aggregating projected text embeddings:

$$s_c = \frac{\sum_{i:\, y_i = c} g(t_{y_i})}{\Big\|\sum_{i:\, y_i = c} g(t_{y_i})\Big\|_2} \in \mathbb{R}^d. \tag{5}$$

Because text embeddings are available for both $c \in \mathcal{S}$ and $c \in \mathcal{U}$, prototypes $\{s_u : u \in \mathcal{U}\}$ provide semantic side information for unseen classes without requiring EEG training examples.

Generative ZSL synthesises feature-space samples for unseen classes to reduce the GZSL bias toward seen classes [1]. In our setting, synthesis occurs in the learned CLIP embedding space. Let $e \in \mathbb{R}^d$ denote a real EEG embedding and $s_c \in \mathbb{R}^d$ its class prototype. We train a conditional generator and critic:

$$\tilde{e} = G(z, s_c), \quad z \sim \mathcal{N}(0, I), \qquad D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}. \tag{6}$$

Following WGAN-GP[3, 4], the critic is trained to assign higher scores to real pairs (e, s_c) than to synthetic pairs $(\tilde{e}, s_c)$ while enforcing a 1-Lipschitz constraint via a gradient penalty. Writing $\hat{e} = \epsilon e + (1-\epsilon)\tilde{e}$ with $\epsilon \sim \text{Uniform}(0,1)$, the objectives are:

$$\mathcal{L}_D = -\mathbb{E}_{(e,c)\sim p_{\text{data}}}\big[D(e, s_c)\big] + \mathbb{E}_{(c,z)}\big[D(G(z, s_c), s_c)\big] + \lambda_{\text{gp}}\,\mathbb{E}_{(c,\hat{e})}\Big(\|\nabla_{\hat{e}} D(\hat{e}, s_c)\|_2 - 1\Big)^2, \tag{7}$$

$$\mathcal{L}_G = -\mathbb{E}_{(c,z)}\big[D(G(z, s_c), s_c)\big]. \tag{8}$$

Hyperparameters: $z_{\text{dim}} = 100$, $\eta = 10^{-4}$, $(\beta_1, \beta_2) = (0.0, 0.9)$, $\lambda_{\text{gp}} = 10$, $n_{\text{critic}} = 5$, $n_{\text{steps}} = 10{,}000$, $B = 256$.

Theoretical expectation: conditioning on s_c encourages G to model the class-conditional embedding distribution in the CLIP space, enabling synthesis for $u \in \mathcal{U}$ by sampling $\tilde{e} \sim p_G(e \mid s_u)$, and thereby providing training mass for unseen labels without accessing unseen EEG trials.

Final model. After training f_b and g on seen trials and fitting the cWGAN-GP on the resulting embeddings, we construct a GZSL training set in $\mathbb{R}^d$:

$$\mathcal{D}_{\text{GZSL}} = \{(e_i, y_i) : y_i \in \mathcal{S}\} \;\cup\; \{(\tilde{e}_{u,j}, u) : u \in \mathcal{U},\; j = 1, \ldots, m\}, \tag{9}$$

where m is the number of synthetic samples per unseen class. We then train multinomial LR on $\mathcal{D}_{\text{GZSL}}$ using the same objective form as in (1), but over the expanded label set $\mathcal{C}$ and using embeddings as inputs. This yields a simple linear classifier whose capacity is deliberately limited, so that any improvements can be attributed to representation learning (CLIP) and distribution-aware augmentation (cWGAN-GP), rather than to an overly expressive downstream model.

5 Result Analysis

Table 2: Ablation study under GZSL evaluation on real seen-test EEG and real unseen EEG. Methods: (A) baseline LR on raw EEG; (B) CLIP nearest-prototype retrieval; (C) LR trained on CLIP EEG embeddings for seen classes only; (D) full model trained on real seen embeddings + cWGAN-GP synthetic unseen embeddings.

Method	Acc_S	Acc_U	H	Macro-F1$_S$	Macro-F1$_U$
(A) Raw EEG + LR	0.0181	0.0000	0.0000	0.0162	0.0000
(B) CLIP Prototype	0.0169	0.0227	0.0194	0.0123	0.0194
(C) CLIP + LR (seen)	0.0305	0.0000	0.0000	0.0202	0.0000
(D) CLIP + cWGAN-GP + LR	0.0003	0.0216	0.0006	0.0011	0.0170

Table 3: Bias (routing) table for the full model (D). Rows are the true group; columns indicate whether the predicted label lies in the seen label set $\mathcal{S}$ or unseen label set $\mathcal{U}$. Percentages are row-normalised.

	Pred: Seen		**Pred: Unseen**	
True: Seen	12	(0.4%)	3296	(99.6%)
True: Unseen	11	(0.1%)	15989	(99.9%)

Table 2 summarises the GZSL performance across four methods. We report seen accuracy Acc_S, unseen accuracy Acc_U, and the harmonic mean, H together with macro-F1 on seen and unseen subsets.

Call-back. The baseline (A) and the seen-only classifier (C) achieve non-trivial Acc_S but essentially zero Acc_U, confirming that purely supervised learning on EEG labels does not transfer to unseen categories (due to distribution shift illustrated in Data and Paradigms 3). In contrast, the CLIP prototype method (B) achieves the best harmonic mean, indicating that EEG–text alignment alone already provides a viable semantic transfer mechanism. The full model (D) achieves non-zero Acc_U by training a classifier over $\mathcal{S} \cup \mathcal{U}$ using cWGAN-GP synthetic unseen embeddings, consistent with the intended role of generative ZSL. However, (D) exhibits a strong trade-off with a marked reduction in Acc_S, motivating an explicit bias and failure-case analysis.

Figure 3 provides a direct reliability check for the Theoretical Expectations (stated in Model Development 4): for real seen EEG, similarities to the matched prototype are higher on average than to mismatched prototypes; for synthetic unseen embeddings, the positive distribution is more strongly shifted, indicating that cWGAN-GP generation is class-conditional in prototype geometry. These diagnostics support why prototype retrieval (B) can perform competitively, and why synthesis in (D) is able to expose the classifier to unseen labels during training.

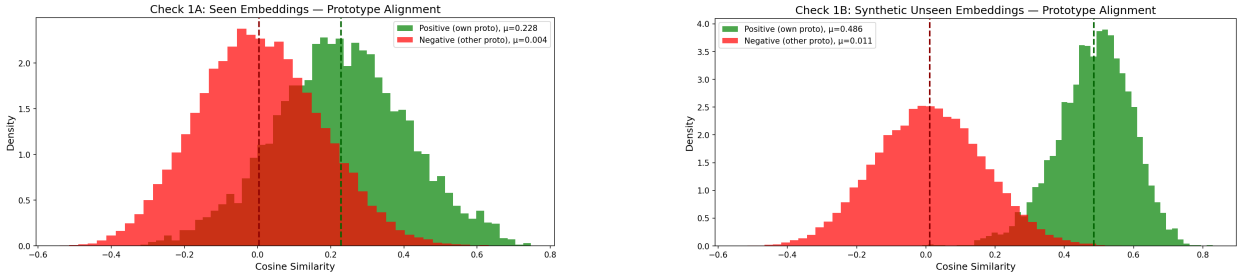

Figure 3: Prototype-alignment histograms. Left: real seen EEG embeddings vs their matched seen text prototype (positive) compared to random mismatched prototypes (negative). Right: the same check for synthetic unseen embeddings conditioned on unseen prototypes.

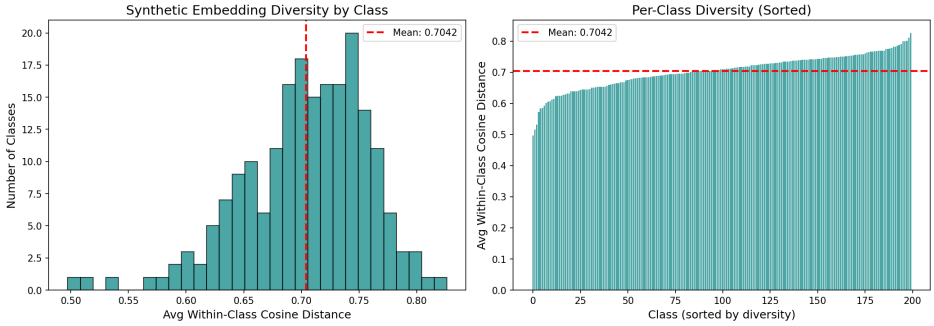

Figure 4: per-class diversity statistics for synthetic unseen embeddings which shows a quantitative check against mode collapse.

Failure Case and Causal Explanation. Although (D) improves Acc_U, it severely degrades Acc_S. We attribute this to a calibration (routing) shift in which the classifier assigns most test points, including true seen examples to the unseen label set. Table 3 shows that model (D) routes the vast majority of true seen trials to unseen labels (99.6%), which explains the collapse in Acc_S. Importantly, the model is not exhibiting the *classic* GZSL failure mode of predicting only seen labels; rather, it exhibits an *inverse* bias towards unseen labels. A plausible cause, is that synthetic unseen embeddings are highly prototype-aligned and comparatively "easy" in the learned geometry, whereas real EEG embeddings remain noisy and weakly aligned. Moreover, the generator does not exhibit severe mode collapse: synthetic embeddings maintain appreciable within-class diversity across unseen categories (Figure 4), suggesting that the main failure mode arises from classifier calibration rather than degenerate synthesis.

Key research finding. Across ablations, the strongest balanced GZSL behaviour is achieved by CLIP prototype retrieval (B), suggesting that in this extremely high-class, low-trial regime, similarity-based semantic decoding can outperform a learned classifier. The full hybrid model (D) demonstrates the intended capability of generative GZSL; enabling a classifier to predict unseen labels, but highlights a central limitation: the learned decision rule can become miscalibrated between seen and unseen label sets. This trade-off, together with the alignment and synthesis diagnostics, provides a principled basis for interpreting success and failure cases in EEG-based GZSL brain decoding.

References

[1] 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, 2018. doi:10.1109/TPAMI.2017.2762295.

[2] Tidiane Camaret Ndir, Robin T. Schirrmeister, and Tonio Ball. Eeg-clip: learning eeg representations from natural language descriptions. *Frontiers in Robotics and AI*, 12:1625731, 2025. doi:10.3389/frobt.2025.1625731.

[3] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[4] George Halal. cWGAN-GP: A conditional wasserstein generative adversarial network with gradient penalty (github repository). `https://github.com/georgehalal/cWGAN-GP`, 2025. Accessed: 2025-12-15.