

HyFI: Hyperbolic Feature Interpolation for Brain-Vision Alignment

Sangmin Jo, Wootae Jeong, Da-Woon Heo, Yoohwan Hwang, Heung-Il Suk*

Department of Artificial Intelligence, Korea University, Seoul, South Korea
 {sangminjo, wtjeong, daheo, yoohwan98, hisuk}@korea.ac.kr

Abstract

Recent progress in artificial intelligence has encouraged numerous attempts to understand and decode human visual system from brain signals. These prior works typically align neural activity independently with semantic and perceptual features extracted from images using pre-trained vision models. However, they fail to account for two key challenges: (1) the modality gap arising from the natural difference in the information level of representation between brain signals and images, and (2) the fact that semantic and perceptual features are highly entangled within neural activity. To address these issues, we utilize hyperbolic space, which is well-suited for considering differences in the amount of information and has the geometric property that geodesics between two points naturally bend toward the origin, where the representational capacity is lower. Leveraging these properties, we propose a novel framework, **Hyperbolic Feature Interpolation (HyFI)**, which interpolates between semantic and perceptual visual features along hyperbolic geodesics. This enables both the fusion and compression of perceptual and semantic information, effectively reflecting the limited expressiveness of brain signals and the entangled nature of these features. As a result, it facilitates better alignment between brain and visual features. We demonstrate that HyFI achieves state-of-the-art performance in zero-shot brain-to-image retrieval, outperforming prior methods with Top-1 accuracy improvements of up to +17.3% on THINGS-EEG and +9.1% on THINGS-MEG.

Code — <https://github.com/ku-milab/HyFI>

Introduction

Understanding how the human brain encodes information has long been a central topic in neuroscience, and has recently attracted growing attention in artificial intelligence. Specifically, the field of brain decoding aims to infer internal cognitive states or external sensory experiences from recorded brain activity (Naselaris et al. 2011; Oota et al. 2023). It offers insights into how the brain represents the external world and enables brain-computer interface (BCI) systems (Ko et al. 2021). In recent years, brain decoding

*Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

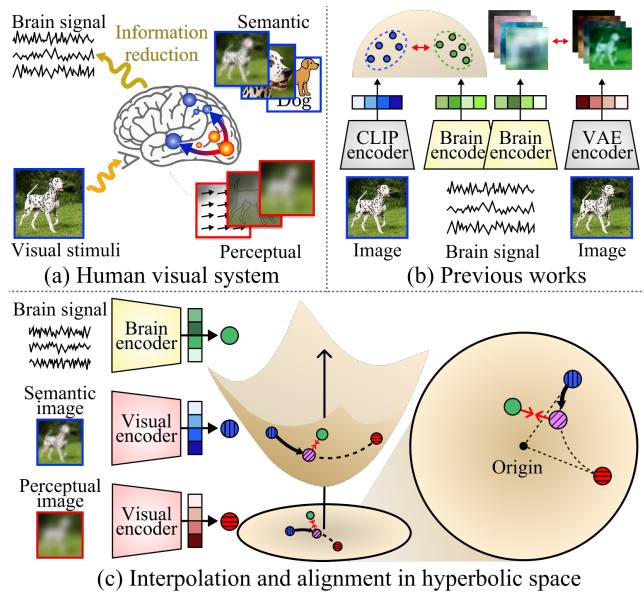


Figure 1: (a) The human visual system processes perceptual and semantic information, and some degradation occurs when neural activity is recorded. (b) Previous works aligned semantic and perceptual features through separate pathways, overlooking their entanglement in brain signals. (c) In contrast, hyperbolic interpolation merges perceptual and semantic features with lower complexity, enhancing alignment with brain signals.

techniques based on machine learning have been successfully applied across diverse cognitive domains such as vision, audition, and language (Wang and Ji 2022; Défossez et al. 2023; Scotti et al. 2024). Among these, visual brain decoding has received particular attention, given that vision is the dominant sensory modality in humans and plays a crucial role in perception and cognition (Mathis et al. 2024).

Visual brain decoding has been extensively studied using neuroimaging modalities such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG). In particular, fMRI has long been a dominant modality in this field due to its superior spatial resolution. However, its limited temporal res-

olution and bulky equipment make it less suitable for real-world applications. In contrast, EEG and MEG offer high temporal resolution, making it particularly suitable for BCI. Motivated by these advantages, we focus on brain decoding tasks based on EEG and MEG signals.

Recent brain decoding studies have increasingly adopted dual-pathway frameworks to capture both perceptual details and semantic representations from neural signals. This approach reflects the human visual system, which processes both *perceptual* and *semantic* features, as shown in Fig. 1(a). Perceptual features refer to low-level visual attributes extracted in early visual areas (e.g., V1), such as orientation, color, and edge information (Miyawaki et al. 2008). Semantic features denote high-level conceptual representations encoded in cortical regions, such as object identity and category (DiCarlo, Zoccolan, and Rust 2012). In recent approaches, semantic features are typically captured by aligning brain activity with image embeddings from pre-trained vision-language models (VLMs) like CLIP (Scotti et al. 2023). In parallel, perceptual features are often decoded by aligning representations derived from variational autoencoders (VAEs) (Shen et al. 2025). These efforts have advanced brain decoding by integrating brain-inspired models with multi-modal representations (Li, Wu, and Chen 2025).

Despite these advances, current approaches still face two key limitations. First, aligning brain signals with image embeddings remains challenging—a problem commonly referred to as the modality gap (Liang et al. 2022). This issue is known to arise from an inherent information imbalance between modalities (Schrodi et al. 2024). In brain decoding, this imbalance is particularly pronounced, as neural signals contain substantially less semantic information than image embeddings for visual tasks (Chen et al. 2024; Wu et al. 2025). The limitation is driven by human attentional bottlenecks, restricted visual working memory (Cavanagh and Alvarez 2005; Dux and Marois 2009), and the low signal-to-noise ratio and resolution of neural recordings (Srinivasan et al. 2007; Naselaris et al. 2011). Second, many existing approaches that separately align perceptual and semantic features fail to reflect the fact that neural activity encodes features in a highly entangled and interactive manner, as they are not processed independently (Pollen 1999; Naselaris et al. 2009). As a result, attempts to align these features independently may lead to suboptimal performance.

To address these limitations, we adopt hyperbolic space for brain-vision alignment. Unlike Euclidean space, hyperbolic geometry with negative curvature offers two key advantages: (1) geodesics between two points naturally bend toward the origin, (2) representational capacity decreases near the origin due to the exponential expansion of the space with radius. Building on these insights, we introduce a novel **Hyperbolic Feature Interpolation (HyFI)** method that interpolates semantic and perceptual visual features in hyperbolic space. This allows semantic and perceptual features to be effectively integrated and compressed during interpolation, making it well suited for brain signals with limited information and entangled semantic-perceptual components. As a result, the interpolated representations become better aligned with brain activity. Our main contributions are as

follows:

- We propose a hyperbolic interpolation method that effectively integrates and compresses semantic and perceptual visual features, explicitly accounting for the limited information capacity and entangled nature of brain signals.
- Our method consistently improves performance across combinations of visual and brain encoders, demonstrating broad applicability.
- It achieves state-of-the-art (SOTA) performance on two public brain decoding benchmarks, with 68.2% Top-1 accuracy on THINGS-EEG and 35.8% on THINGS-MEG, outperforming previous methods by +17.3% and +9.1%, respectively.

Related Works

Visual Brain Decoding

Visual brain decoding has received increasing attention for its potential to uncover the mechanisms of human cognition and to enable practical BCI (Kay et al. 2008; Yang, Gee, and Shi 2024). With the rise of large-scale VLMs, recent studies have begun to leverage those representations by aligning semantic features with CLIP (Scotti et al. 2023; Song et al. 2024). A similar trend is observed in EEG-based visual decoding. However, the inherent modality gap between neural signals and pre-trained visual embeddings remains a major challenge. To mitigate this, Li et al. (2024) utilized a diffusion prior to map brain features into image space, Zhang et al. (2025) employed multi-modal fusion to increase shared information, and Wu et al. (2025) reduced visual complexity using Gaussian blur. These methods overlook the fact that semantic and perceptual representations are often entangled in neural activity. In contrast, our approach uses hyperboloid interpolation to fuse both features while reducing complexity, enabling better alignment.

Hyperbolic Representation Learning

Hyperbolic space has attracted attention in representation learning for its ability to model hierarchical data due to its negative curvature (Nickel and Kiela 2017; Chamberlain, Clough, and Deisenroth 2017). This property has led to successful applications across various modalities with inherent hierarchies, including graphs (Liu, Nickel, and Kiela 2019), text (Dhingra et al. 2018; Tifrea, Becigneul, and Ganea 2019), and images (Atigh et al. 2022). Recent works have further extended hyperbolic geometry to the multi-modal domain. For instance, Desai et al. (2023) propose hyperbolic vision-language models where image embeddings are constrained to lie within a concept cone defined by the text embeddings. Similarly, Pal et al. (2024) extend this framework to model hierarchical relations across multiple levels, including cropped-text, cropped-image, original text, and original image representations. Building on these insights, we extend hyperbolic representation learning to the domain of brain decoding. Specifically, we leverage the geodesic property of hyperbolic space to unify and compress semantic and perceptual visual information, providing a new perspective on aligning neural signals with rich visual stimuli.

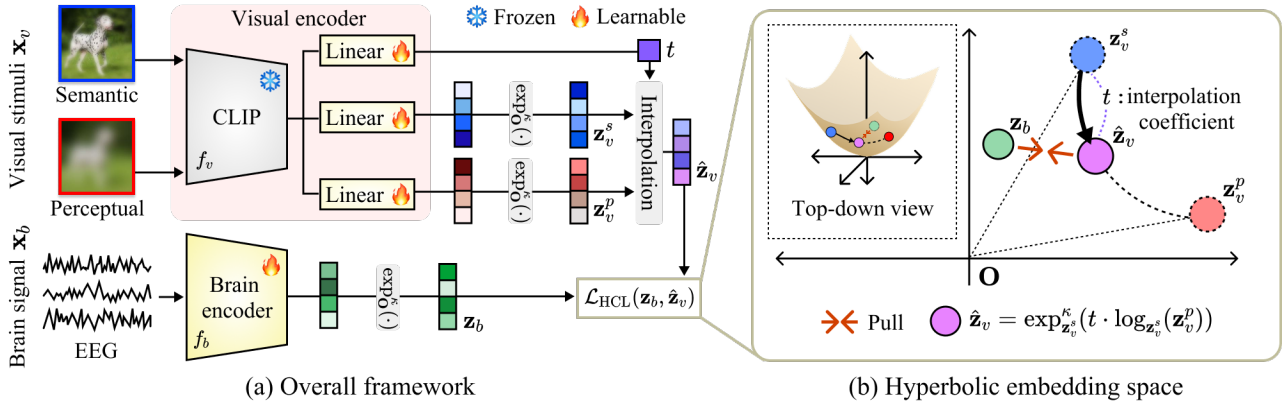


Figure 2: (a) The semantic image \mathbf{x}_v^s and perceptual image \mathbf{x}_v^p are encoded by CLIP and projected via a linear layer, and then lifted onto the hyperboloid via the exponential map. Using a learned weight t derived from the semantic image features, the two image features are interpolated on the hyperbolic manifold. Similarly, EEG inputs are encoded and projected onto the same hyperbolic space. Contrastive learning is then performed on the hyperboloid to bring paired EEG-image representations closer. (b) A schematic view of the hyperbolic embedding space. The interpolated representation $\hat{\mathbf{z}}_v$ lies along the geodesic between the semantic feature \mathbf{z}_v^s and the perceptual feature \mathbf{z}_v^p . Contrastive learning then pulls the EEG feature \mathbf{z}_b toward the target $\hat{\mathbf{z}}_v$.

Preliminaries

We formulate our approach on hyperbolic space, a Riemannian manifold with constant negative curvature, where volume grows exponentially with radius. This property makes it well suited for representing hierarchical structures and addressing modality imbalance in multi-modal learning (Le et al. 2019; Peng et al. 2021). Following prior works (Desai et al. 2023; Pal et al. 2024), we adopt the Lorentz (hyperboloid) model due to its strong empirical performance in multi-modal tasks.

Definition The Lorentz model \mathbb{L}^n of n -dimensional hyperbolic space with constant negative curvature is realized as the ‘‘upper sheet’’ of a two-sheeted hyperboloid in $(n + 1)$ -dimensional Minkowski space (Cannon et al. 1997). Concretely, a point $\mathbf{p} \in \mathbb{R}^{n+1}$ is represented as $\mathbf{p} = (p_0, \tilde{\mathbf{p}})$, where $p_0 > 0$ denotes the time component and $\tilde{\mathbf{p}} \in \mathbb{R}^n$ are spatial coordinates. The Lorentz manifold is defined as:

$$\mathbb{L}^n = \left\{ \mathbf{p} \in \mathbb{R}^{n+1} : \langle \mathbf{p}, \mathbf{p} \rangle_{\mathbb{L}} = -\frac{1}{\kappa}, p_0 > 0 \right\}, \quad (1)$$

where $-\kappa \in \mathbb{R}$ is the curvature of the space. The Lorentzian inner product for two vectors $\mathbf{p}, \mathbf{q} \in \mathbb{L}^n$ is defined as:

$$\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}} = -p_0 q_0 + \langle \tilde{\mathbf{p}}, \tilde{\mathbf{q}} \rangle_{\mathbb{E}}, \quad (2)$$

where $\langle \cdot, \cdot \rangle_{\mathbb{E}}$ denotes the standard Euclidean dot product.

Geodesics A geodesic is the shortest curve connecting two points on the manifold. In the Lorentz model, the geodesic distance between two points $\mathbf{p}, \mathbf{q} \in \mathbb{L}^n$ is defined as:

$$d_{\mathbb{L}}(\mathbf{p}, \mathbf{q}) = \sqrt{1/\kappa} \cdot \cosh^{-1}(-\kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}}). \quad (3)$$

Exponential and Logarithmic Map The exponential map defines a smooth mapping from the tangent space onto the Lorentz manifold. For a point $\mathbf{p} \in \mathbb{L}^n$, the tangent space is defined as:

$$T_{\mathbf{p}}\mathbb{L}^n = \{ \mathbf{v} \in \mathbb{R}^{n+1} \mid \langle \mathbf{p}, \mathbf{v} \rangle_{\mathbb{L}} = 0 \}. \quad (4)$$

Given a tangent vector $\mathbf{v} \in T_{\mathbf{p}}\mathbb{L}^n$, the exponential map traces the geodesic from \mathbf{p} in the direction of \mathbf{v} and is parameterized as $\gamma(t) = \exp_{\mathbf{p}}^{\kappa}(t\mathbf{v})$, where $t \in [0, 1]$. It is explicitly defined as:

$$\exp_{\mathbf{p}}^{\kappa}(t\mathbf{v}) = \cosh(t\sqrt{\kappa}\|\mathbf{v}\|_{\mathbb{L}}) \mathbf{p} + \frac{\sinh(t\sqrt{\kappa}\|\mathbf{v}\|_{\mathbb{L}})}{\sqrt{\kappa}\|\mathbf{v}\|_{\mathbb{L}}} \mathbf{v}, \quad (5)$$

where $\|\mathbf{v}\|_{\mathbb{L}} = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle_{\mathbb{L}}}$. Conversely, a point $\mathbf{q} \in \mathbb{L}^n$ on the hyperboloid can be projected onto the tangent space via the logarithmic map $\log_{\mathbf{p}}^{\kappa}(\cdot) : \mathbb{L}^n \rightarrow T_{\mathbf{p}}\mathbb{L}^n$, as follows:

$$\log_{\mathbf{p}}^{\kappa}(\mathbf{q}) = \frac{\cosh^{-1}(-\kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}})}{\sqrt{(\kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}})^2 - 1}} (\mathbf{q} + \kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}} \mathbf{p}). \quad (6)$$

In practice, the point \mathbf{p} is commonly set to the time origin $\mathbf{O} = (\sqrt{1/\kappa}, 0, \dots, 0)^{\top} \in \mathbb{L}^n$. Under this setting, a vector $\mathbf{v} = [0, \mathbf{v}_{\text{enc}}] \in \mathbb{R}^{n+1}$ lies in the tangent space at \mathbf{O} and can be mapped onto the hyperboloid via exponential map $\exp_{\mathbf{O}}^{\kappa}(\cdot)$, where \mathbf{v}_{enc} denotes the encoder output.

Method

Problem Formulation

Given the brain signals space \mathcal{X}_b and the visual stimuli space \mathcal{X}_v , the goal of visual brain decoding is to map brain signals $\mathbf{x}_b \in \mathcal{X}_b$ into a shared space \mathcal{H} aligned with $\mathbf{x}_v \in \mathcal{X}_v$. To this end, we learn a brain encoder $f_b : \mathcal{X}_b \rightarrow \mathcal{H}$ and a visual encoder $Wf_v : \mathcal{X}_v \rightarrow \mathcal{H}$, where f_v is a frozen backbone of pre-trained VLM and W is a linear layer. Specifically, we use n -dimensional Lorentz space \mathbb{L}^n as a semantically aligned space \mathcal{H} . Embeddings in \mathbb{L}^n are obtained via the exponential map $\exp_{\mathbf{O}}^{\kappa}(\cdot)$ at the time origin \mathbf{O} . The visual embedding is defined as $\mathbf{z}_v = \exp_{\mathbf{O}}^{\kappa}(\alpha_v \cdot Wf_v(\mathbf{x}_v))$ and the brain embedding $\mathbf{z}_b = \exp_{\mathbf{O}}^{\kappa}(\alpha_b \cdot f_b(\mathbf{x}_b))$, where α_v and α_b are learnable projection scalars that reduce the norm of the embeddings to keep them near the origin \mathbf{O} . An overview of our overall framework is illustrated in Fig 2(a).

Hyperbolic Brain-Vision Contrastive Learning

To align neural embeddings with corresponding visual embedding, we utilize contrastive learning in hyperbolic space. Unlike Euclidean space, hyperbolic space provides a natural embedding space for aligning modalities with asymmetric information capacity and has shown empirical success in multi-modal representation learning (Desai et al. 2023; Pal et al. 2024; Mandica et al. 2025).

Given a batch of EEG-image pairs $\{(\mathbf{z}_{b,i}, \mathbf{z}_{v,i})\}_{i=1}^B$, hyperbolic contrastive learning is formulated as:

$$\mathcal{L}(\mathbf{z}_v, \mathbf{z}_b) = - \sum_{i \in B} \log \frac{\exp(d_{\mathbb{L}}(\mathbf{z}_{v,i}, \mathbf{z}_{b,i})/\tau)}{\sum_{k=1, k \neq i}^B \exp(d_{\mathbb{L}}(\mathbf{z}_{v,i}, \mathbf{z}_{b,k})/\tau)}, \quad (7)$$

where B denote the batch, $d_{\mathbb{L}}$ denote the negative Lorentz distance, τ is temperature parameter.

Hyperbolic Feature Interpolation

With hyperbolic space established, we aim to learn visual representations aligned with brain signals by capturing their inherent properties. In particular, we fuse and compress semantic and perceptual visual features. Our approach is motivated by two key observations: (1) semantic and perceptual visual features are often entangled in neural activity, (2) and brain signals inherently contain less information than natural images. To address the first property, we describe how semantic and perceptual features are extracted from an image, then present a hyperbolic interpolation method. Finally, we show that this interpolation naturally leads to information compression, thereby addressing the second observation.

Extracting Semantic and Perceptual Features We first apply image-level augmentations to obtain the semantic and perceptual visual inputs. The semantic image \mathbf{x}_v^s is generated via fovea blur, simulating peripheral vision to preserve semantics and enhance alignment with brain signals (Wu et al. 2025). The perceptual image \mathbf{x}_v^p is obtained by applying Gaussian blurring to suppress high-frequency components and retain coarse structure. This augmentation amplifies perceptual attributes in the CLIP embeddings. These augmentations and their effects are presented in Fig. 3.

These inputs are then projected into hyperbolic space as semantic visual features $\mathbf{z}_v^s = \exp_{\mathbb{O}}^{\kappa}(\alpha_v \cdot W_s f_v(\mathbf{x}_v^s))$ and perceptual visual features $\mathbf{z}_v^p = \exp_{\mathbb{O}}^{\kappa}(\alpha_v \cdot W_p f_v(\mathbf{x}_v^p))$, respectively. Here, $W_s, W_p \in \mathbb{R}^{d \times d}$ are learnable matrices and d denotes the CLIP embedding dimension.

Hyperbolic Interpolation To perform interpolation of these features, we approximate the geodesic in the Lorentz model using the exponential map. The perceptual feature \mathbf{z}_v^p is projected onto the tangent space at \mathbf{z}_v^s using the logarithmic map, i.e., $\log_{\mathbf{z}_v^s}^{\kappa}(\mathbf{z}_v^p) \in T_{\mathbf{z}_v^s} \mathbb{L}^n$. This tangent vector is scaled and mapped back to the hyperbolic space via the exponential map. The resulting interpolated visual representation $\hat{\mathbf{z}}_v$ follows the geodesic from \mathbf{z}_v^s to \mathbf{z}_v^p :

$$\hat{\mathbf{z}}_v = \gamma_{\mathbf{z}_v^s \rightarrow \mathbf{z}_v^p}(t) = \exp_{\mathbf{z}_v^s}^{\kappa} \left(t \cdot \log_{\mathbf{z}_v^s}^{\kappa}(\mathbf{z}_v^p) \right), \quad (8)$$

where $t \in [0, 1]$ is the interpolation coefficient.

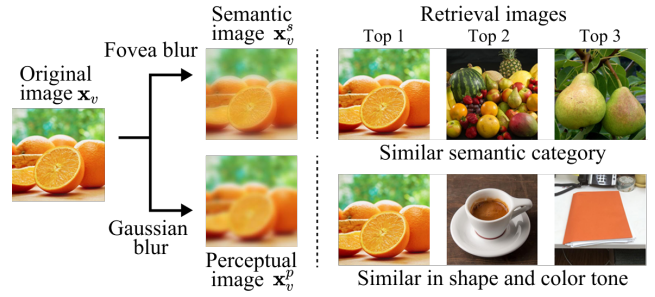


Figure 3: Examples of image augmentations and retrieval results. The semantic image \mathbf{x}_v^s and perceptual image \mathbf{x}_v^p are generated via fovea blur and Gaussian blur, respectively. Retrieval results using CLIP embedding show that semantic queries return category-relevant matches (e.g., fruits), while perceptual queries retrieve images with similar low-level attributes such as color and shape.

We compute the interpolation coefficient t dynamically to reflect image-specific variation in the relative importance of semantic and perceptual features:

$$t = \sigma(W_t f_v(\mathbf{x}_v^s)), \quad (9)$$

where $W_t \in \mathbb{R}^{1 \times d}$ is a learnable matrix and σ denotes the sigmoid function.

Compression Effects The proposed interpolation mechanism concurrently compresses and fuses semantic and perceptual features. To analyze how this compression arises, we revisit the geodesic formulation in the Lorentz model. We reformulate the geodesic in Eq. (8) as follows:

$$\gamma_{\mathbf{p} \rightarrow \mathbf{q}}(t) = \frac{\sinh((1-t)\beta)}{\sinh(\beta)} \mathbf{p} + \frac{\sinh(t\beta)}{\sinh(\beta)} \mathbf{q}, \quad (10)$$

where $\beta = \sqrt{\kappa} \cdot d_{\mathbb{L}}(\mathbf{p}, \mathbf{q})$. A detailed derivation is provided in the appendix.

Unlike linear interpolation $(1-t)\mathbf{p} + t\mathbf{q}$ in Euclidean space, the hyperbolic interpolation weights $\frac{\sinh((1-t)\beta)}{\sinh(\beta)}$ and $\frac{\sinh(t\beta)}{\sinh(\beta)}$ are strictly smaller than $(1-t)$ and t , respectively. This causes the interpolated points to lie closer to the origin.

To understand how this contraction relates to the geometry of hyperbolic space, we revisit the hyperboloid constraint in Eq. (1), which gives:

$$p_0 = \sqrt{1/\kappa + \|\tilde{\mathbf{p}}\|^2}, \quad (11)$$

A smaller p_0 constrains $\|\tilde{\mathbf{p}}\|$ more tightly, thus reducing the expressive capacity of embedding. This is consistent with prior work (Ganea, Becigneul, and Hofmann 2018; Khulkov et al. 2020), suggesting that points near the origin represent more abstract concepts.

Final Objective Function Finally, we train our model by aligning the interpolated visual representations with the brain embeddings in hyperbolic space. The final hyperbolic contrastive loss is defined as:

$$\mathcal{L}_{\text{HCL}} = \mathcal{L}(\hat{\mathbf{z}}_v, \mathbf{z}_b) + \mathcal{L}(\mathbf{z}_b, \hat{\mathbf{z}}_v). \quad (12)$$

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Average	
	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Intra-subject: train and test on one subject																						
BraVL	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE	13.2	39.5	13.5	40.3	14.5	42.7	20.6	52.7	10.1	31.5	16.5	44.0	17.0	42.1	22.9	56.1	15.4	41.6	17.4	45.8	16.1	43.6
ATM-S	25.6	60.4	22.0	54.5	25.0	62.4	31.4	60.9	12.9	43.0	21.3	51.1	30.5	61.5	38.8	72.0	30.4	51.5	29.1	63.5	28.5	60.4
Cog-cap	31.4	79.7	31.4	77.8	38.2	85.7	40.4	85.8	24.4	66.3	34.8	78.8	34.7	81.0	48.1	88.6	37.4	79.4	35.6	79.3	35.6	80.2
UBP	<u>41.2</u>	<u>70.5</u>	<u>51.2</u>	<u>80.9</u>	<u>51.2</u>	<u>82.0</u>	<u>51.1</u>	<u>76.9</u>	<u>42.2</u>	<u>72.8</u>	<u>57.5</u>	<u>83.5</u>	<u>49.0</u>	<u>79.9</u>	<u>58.6</u>	<u>85.8</u>	<u>45.1</u>	<u>76.2</u>	<u>61.5</u>	<u>88.2</u>	<u>50.9</u>	<u>79.7</u>
HyFI	60.6	85.3	65.9	94.0	69.5	93.9	66.5	89.8	55.0	86.0	74.4	95.0	68.4	91.3	78.9	96.9	66.0	90.6	77.0	96.4	68.2	91.9
Inter-subject: leave one subject out for test																						
BraVL	2.3	8.0	1.5	6.3	1.9	6.7	2.1	8.1	2.2	7.6	1.6	6.4	2.3	8.5	1.8	7.0	1.4	5.9	1.7	6.7	1.5	5.6
NICE	7.6	22.8	5.9	20.5	6.0	22.3	6.3	20.7	4.4	18.3	5.6	22.2	5.6	19.7	6.3	22.0	5.7	17.6	8.4	28.3	6.2	21.4
NICE-G	5.9	21.4	6.4	22.7	5.5	20.1	6.1	21.0	4.7	19.5	6.2	22.5	5.9	19.1	7.3	25.3	6.2	18.3	6.2	26.3	5.9	21.6
ATM-S	10.5	26.8	7.1	24.8	11.9	33.8	<u>14.7</u>	<u>39.4</u>	7.0	23.9	11.1	35.8	16.1	43.5	15.0	40.3	4.9	22.7	<u>20.5</u>	<u>46.5</u>	11.8	<u>33.7</u>
UBP	<u>11.5</u>	<u>29.7</u>	<u>15.5</u>	<u>40.0</u>	<u>9.8</u>	<u>27.0</u>	13.0	32.3	<u>8.8</u>	33.8	<u>11.7</u>	31.0	10.2	23.8	12.2	<u>32.2</u>	15.5	40.5	16.0	43.5	12.4	33.4
HyFI	16.2	35.8	20.0	47.7	7.5	26.7	18.8	41.3	9.7	<u>27.5</u>	15.5	<u>33.8</u>	<u>11.0</u>	<u>34.5</u>	<u>13.2</u>	30.3	<u>13.3</u>	<u>38.8</u>	25.3	55.7	15.1	37.2

Table 1: Top-1 (T-1) and top-5 (T-5) accuracy (%) results in 200-way zero-shot brain-to-image retrieval on THINGS-EEG, reported for intra- and inter-subject settings. **Bold** and underline indicate the best and second-best results, respectively.

Method	Subject 1		Subject 2		Subject 3		Subject 4		Average	
	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5	T-1	T-5
Intra-subject: train and test on one subject										
NICE	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.3	42.3
UBP	<u>15.0</u>	<u>38.0</u>	<u>46.0</u>	<u>80.5</u>	<u>27.3</u>	<u>59.0</u>	<u>18.5</u>	<u>43.5</u>	<u>26.7</u>	<u>55.2</u>
HyFI	17.6	40.1	63.8	91.1	38.0	76.9	23.7	50.2	35.8	64.6
Inter-subject: leave one subject out for test										
UBP	2.0	5.7	1.5	17.2	2.7	10.5	2.5	8.0	2.2	10.4
HyFI	2.7	6.0	4.3	17.0	3.5	12.5	<u>1.0</u>	<u>7.6</u>	3.2	11.5

Table 2: Top-1 and top-5 accuracy (%) results in 200-way zero-shot brain-to-image retrieval on THINGS-MEG.

Experiments and Results

Datasets

THINGS-EEG We used the THINGS-EEG dataset (Gifford et al. 2022), a large-scale EEG benchmark collected under the Rapid Serial Visual Presentation (RSVP) paradigm, which provides image-EEG paired data from 10 subjects viewing rapid sequences of images. The training set consists of 1,654 object concepts, each represented by 10 distinct images, with each image repeated 4 times per subject. The test set includes 200 concepts with one image per concept and each image repeated 80 times per subject. Following prior preprocessing protocols (Song et al. 2024; Wu et al. 2025), we averaged trials, downsampled EEG signals to 250 Hz, and selected 17 channels over occipital and parietal regions.

THINGS-MEG We additionally utilized the THINGS-MEG dataset (Hebart et al. 2023), which contains recordings from four participants using 271 MEG channels. The training set includes 1,854 object concepts, each presented once

with 12 different images, while the test set consists of 200 concepts, each associated with a single image repeated 12 times. To ensure fair comparison, we follow the same preprocessing pipeline as described in (Song et al. 2024; Wu et al. 2025); more details are provided in the appendix.

Implementation Details

We used AdamW (Loshchilov and Hutter 2017) with learning rate 3×10^{-4} , weight decay 1×10^{-4} , and batch size 1024. We trained the model for 50 epochs. The curvature parameter κ was set to 1 at initialization and optimized during training. To control feature norms, we initialize the scaling factors α_v and α_b to $\sqrt{1/d}$, following (Desai et al. 2023). We empirically observed that fixing $\alpha_v = 1$ led to better training stability for ResNet-based and hyperbolic vision encoder, due to the already exhibit low norm of their features. All experiments were conducted on a GPU, GTX 1080 Ti (12GB). We applied fovea blur and Gaussian blur. Augmentation details are in the appendix.

Vision Encoders We evaluated a range of visual backbones, including RN50, RN101, ViT-B/16, ViT-B/32, ViT-L/14, and ViT-H/14. In addition, we also used two recent hyperbolic vision-language models, MERU (Desai et al. 2023) and HyCoCLIP (Pal et al. 2024), which are pretrained to capture hierarchical relationships between images and text. We adopted RN50 as the default vision encoder.

Brain Encoders We applied our method using a variety of EEG encoders to demonstrate its generality across different architectures. Specifically, we adopted models that are either widely used in the literature or recently proposed in neural encoding studies, including ShallowNet (Schirmer et al. 2017), EEGNet (Lawhern et al. 2018), and TSCoV (Li et al. 2024), as well as EEGProject (Wu et al. 2025). We used EEGProject as default EEG encoder.

Interpolation	Hyperbolic	THINGS-EEG		THINGS-MEG	
		T-1	T-5	T-1	T-5
–	–	49.4	81.0	23.1	47.8
–	✓	54.3	82.5	28.8	55.1
✓	–	59.7	86.8	25.2	51.8
✓	✓	68.2	91.9	35.8	64.6

Table 3: Ablation study: effect of interpolation and hyperbolic space. The row with only interpolation checked refer to interpolation in CLIP space.

Results

We compare our method with several recent neural decoding approaches for brain-to-image retrieval, including BraVL (Du et al. 2023), NICE (Song et al. 2024), ATM (Li et al. 2024), CogCap (Zhang et al. 2025), and UBP (Wu et al. 2025). Detailed baseline descriptions are in the appendix.

To evaluate decoding performance, we perform 200-way zero-shot retrieval on THINGS-EEG and THINGS-MEG, following prior work (Du et al. 2023; Wu et al. 2025). This evaluates alignment quality and generalization to novel concepts. The results in Tables 1 and 2. We averaged over 5 runs, and all improvements are statistically significant ($p < 0.01$). On the THINGS-EEG dataset, our method achieves a top-1 accuracy of 68.2% and top-5 accuracy of 91.9%, outperforming the previous SOTA (UBP) by +17.3% and +12.2%, respectively. On the THINGS-MEG dataset, HyFI achieves a top-1 accuracy of 35.8% and top-5 accuracy of 64.6%, improving upon UBP by +9.1% and +9.4%, respectively.

Furthermore, we conduct a qualitative comparison of retrieval results with UBP, as shown in Fig. 4. For fair comparison, we used subject 4 with near-average performance. As illustrated in the figure, our method preserves both semantic and perceptual coherence, whereas previous approaches fail to maintain this consistency. This indicates that our method effectively fuses perceptual and semantic features, thereby maintaining coherence in both aspects. More examples of retrieved image are provided in the appendix.

Ablation Study

Effect of Hyperbolic Geometry and Feature Interpolation To evaluate the contributions of hyperbolic space and feature interpolation, we conduct an ablation study and results in Table 3. First, we observe that aligning brain and image features is more effective in hyperbolic space than in CLIP space (Euclidean space). This supports the suitability of hyperbolic space for modeling the information imbalance between brain and visual modalities. Next, we compare interpolating features in the original CLIP space and in hyperbolic space. Even in the CLIP space, fusing low-level visual features improves performance, highlighting the benefit of integrating perceptual information. However, performing interpolation in hyperbolic space leads to the best results. This finding indicates that hyperbolic interpolation effectively reduces redundancy and representational complexity, while facilitating feature fusion.

Method	Architecture	Model	THINGS-EEG		THINGS-MEG	
			w/o HyFI	Ours	w/o HyFI	Ours
CLIP	CNN	RN50	49.4	68.2	23.1	35.8
		RN101	44.4	62.1	21.0	34.8
	ViT	ViT-B/16	36.0	41.6	18.6	23.3
		ViT-B/32	42.5	46.9	20.0	25.1
		ViT-L/14	30.1	34.1	13.0	21.0
	ViT-H/14	42.5	43.8	18.7	27.9	
MERU	ViT	ViT-S/16	43.7	52.4	21.6	31.1
		ViT-B/16	31.7	48.8	15.7	24.6
		ViT-L/16	23.1	30.9	11.0	19.6
HyCoCLIP	ViT	ViT-S/16	44.3	51.6	21.2	27.9
		ViT-B/16	35.6	45.6	18.1	27.0

Table 4: Top-1 image retrieval accuracy (%) across visual encoders. The “w/o HyFI” represents alignment in CLIP space; MERU and HyCoCLIP use hyperbolic alignment.

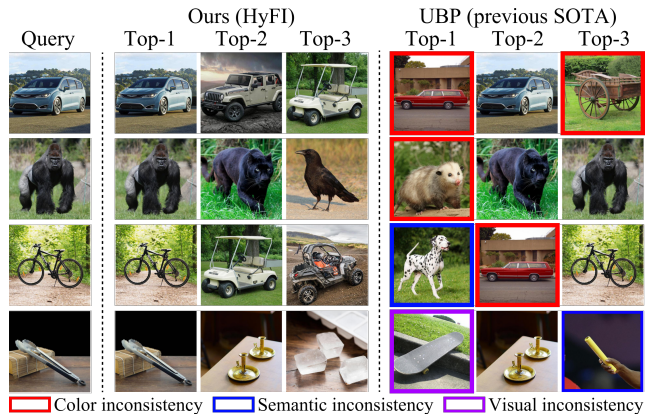


Figure 4: Qualitative comparison of image retrieval results. Our method retrieves semantically and perceptually coherent images, while the previous method often suffers from color or semantic inconsistencies.

Effect of Vision and Brain Encoders

We investigate the impact of different vision and brain encoders on brain-to-image retrieval performance, the results summarized in Table 4 and 5. For this analysis, we use EEGProject as the brain encoder and compare various pre-trained vision encoders. We observe that CNN-based backbones yield stronger alignment with EEG signals compared to transformer-based models. Interestingly, lightweight architectures outperform deeper models. These results suggest that compact visual representations may be more compatible with neural signals. Importantly, our method consistently outperforms the baseline across all vision architectures. We also evaluate the effect of different brain encoders while fixing the vision encoder to CLIP-RN50. We observe that our method consistently improves performance across all brain encoder architectures. This highlights the general applicability and robustness of our approach, regardless of the specific choice of brain encoder.

Brain Encoder	THINGS-EEG		THINGS-MEG	
	w/o HyFI	Ours	w/o HyFI	Ours
ShallowNet	36.6	50.5	15.1	23.2
EEGNet	36.2	51.6	19.5	26.4
TSCov	41.1	57.1	21.9	30.5
EEGProject	49.4	68.2	23.1	35.8

Table 5: Top-1 image retrieval accuracy (%) using different brain encoders and CLIP-RN50.

Analysis

Feature Visualization To investigate the effect of feature interpolation in hyperbolic space, we visualize the distribution of distances from the root, as shown in Fig. 5. Following (Desai et al. 2023), we define the root as the mean of all embeddings in CLIP space and the time origin \mathbf{O} in hyperbolic space. As shown in Fig. 5, (a) in the CLIP space, the interpolated image embeddings lie between the semantic and perceptual features, whereas (b) in the hyperbolic space, the interpolated embeddings are located closer to the time origin \mathbf{O} . This behavior reflects the nature of hyperbolic interpolation, which bends toward the origin and results in a tighter bound on the spatial components, as described in Eq. (11). It suggests that effectively compressing and integrating semantic and perceptual features leads to better alignment with brain signals. Notably, the EEG embeddings lie farther from the origin, as their high variability requires regions with fewer constraints (i.e., away from the time origin).

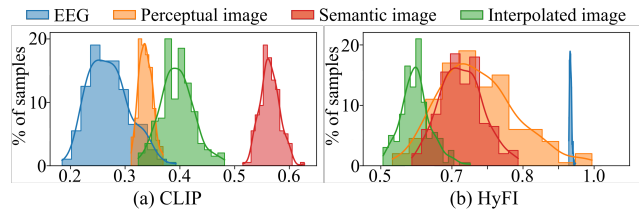


Figure 5: Distributions of embedding distances from the root in (a) CLIP space and (b) hyperbolic space. Interpolated image embeddings lie closer to the root in hyperbolic space, unlike in CLIP space.

Analysis of interpolation coefficient To understand how the model adaptively integrates semantic and perceptual information, we analyze the learned interpolation coefficient $t \in [0, 1]$. The distribution of coefficients is shown in Fig. 6. We observe that t is predominantly distributed below 0.5, indicating that the model tends to focus more on semantic features during interpolation. To further interpret this behavior, we also examine images corresponding to both low and high t values in the test set. Images with lower t values typically contain objects that are iconic examples of their higher-level categories (e.g., banana–fruit, cheetah–mammal, and van–car). In contrast, images with higher t values tend to exhibit salient low-level visual attributes such as orientation and color.

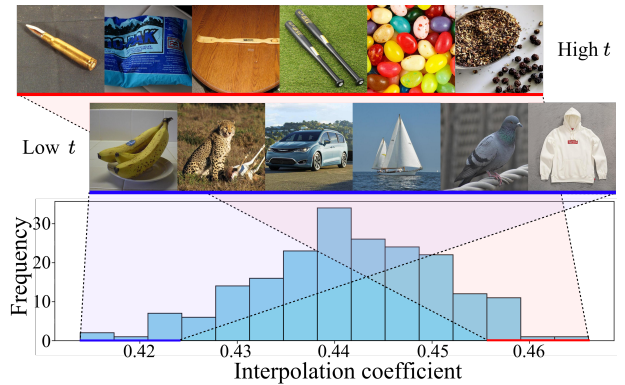


Figure 6: Distribution of the interpolation coefficient t and example images with low and high t values.

Analyzing the effect of image augmentation To investigate the effect of image augmentations on a pre-trained VLM, we conduct image-to-image retrieval using perceptual and semantic images as queries in CLIP-RN50. The results are shown in Fig. 7. When using the semantic images as queries, the retrieved results tend to be semantically consistent. In contrast, perceptual images bias the retrieval toward low-level visual attributes such as color and object orientation. Additional results with other VLMs are in the appendix.

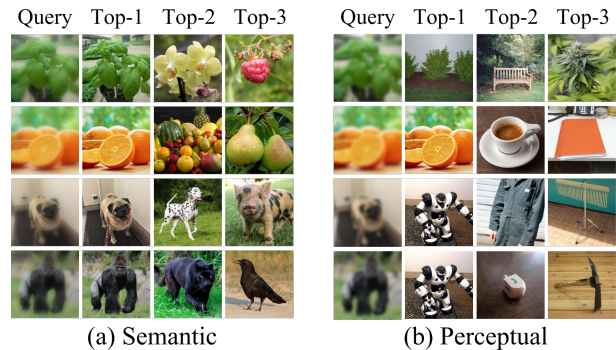


Figure 7: Comparison of Top-3 retrieval results using (a) semantic and (b) perceptual images. Semantic queries tend to retrieve conceptually similar images (e.g., plants, fruits, animals). In contrast, perceptual queries retrieve images with shared low-level visual feature such as color and orientation.

Conclusion

We propose Hyperbolic Feature Interpolation (HyFI), a framework that interpolates semantic and perceptual visual features in hyperbolic space for improved alignment with brain signals. Leveraging the geodesic curvature toward the origin, HyFI enables effective fusion and compression of visual representations. This leads to better alignment with neural activity by reflecting both limited brain information and feature entanglement. Experiments on THINGS-EEG and THINGS-MEG show that HyFI achieves SOTA performance on zero-shot brain-to-image retrieval task.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) under two projects: (No. RS-2019-II190079, Artificial Intelligence Graduate School Program at Korea University) and (No. RS-2024-00457882, National AI Research Lab Project).

References

- Atigh, M. G.; Schoep, J.; Acar, E.; Van Noord, N.; and Mettes, P. 2022. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4453–4462.
- Cannon, J. W.; Floyd, W. J.; Kenyon, R.; Parry, W. R.; et al. 1997. Hyperbolic geometry. *Flavors of geometry*, 31(59-115): 2.
- Cavanagh, P.; and Alvarez, G. A. 2005. Tracking multiple targets with multifocal attention. *Trends in cognitive sciences*, 9(7): 349–354.
- Chamberlain, B. P.; Clough, J.; and Deisenroth, M. P. 2017. Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*.
- Chen, H.; He, L.; Liu, Y.; and Yang, L. 2024. Visual Neural Decoding via Improved Visual-EEG Semantic Consistency. *arXiv preprint arXiv:2408.06788*.
- Défossez, A.; Caucheteux, C.; Rapin, J.; Kabeli, O.; and King, J.-R. 2023. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10): 1097–1107.
- Desai, K.; Nickel, M.; Rajpurohit, T.; Johnson, J.; and Vedantam, S. R. 2023. Hyperbolic image-text representations. In *International Conference on Machine Learning*, 7694–7731. PMLR.
- Dhingra, B.; Shallue, C. J.; Norouzi, M.; Dai, A. M.; and Dahl, G. E. 2018. Embedding text in hyperbolic spaces. *arXiv preprint arXiv:1806.04313*.
- DiCarlo, J. J.; Zoccolan, D.; and Rust, N. C. 2012. How does the brain solve visual object recognition? *Neuron*, 73(3): 415–434.
- Du, C.; Fu, K.; Li, J.; and He, H. 2023. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10760–10777.
- Dux, P. E.; and Marois, R. 2009. The attentional blink: A review of data and theory. *Attention, Perception, & Psychophysics*, 71(8): 1683–1700.
- Ganea, O.; Becigneul, G.; and Hofmann, T. 2018. Hyperbolic Entailment Cones for Learning Hierarchical Embeddings. In *International Conference on Machine Learning*, 1646–1655.
- Gifford, A. T.; Dwivedi, K.; Roig, G.; and Cichy, R. M. 2022. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264: 119754.
- Hebart, M. N.; Contier, O.; Teichmann, L.; Rockter, A. H.; Zheng, C. Y.; Kidder, A.; Corriveau, A.; Vaziri-Pashkam, M.; and Baker, C. I. 2023. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12: e82580.
- Kay, K. N.; Naselaris, T.; Prenger, R. J.; and Gallant, J. L. 2008. Identifying natural images from human brain activity. *Nature*, 452(7185): 352–355.
- Khrulkov, V.; Mirvakhobova, L.; Ustinova, E.; Oseledets, I.; and Lempitsky, V. 2020. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6418–6428.
- Ko, W.; Jeon, E.; Jeong, S.; and Suk, H.-I. 2021. Multi-scale neural network for EEG representation learning in BCI. *IEEE Computational Intelligence Magazine*, 16(2): 31–45.
- Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of neural engineering*, 15(5): 056013.
- Le, M.; Roller, S.; Papaxanthos, L.; Kiela, D.; and Nickel, M. 2019. Inferring concept hierarchies from text corpora via hyperbolic embeddings. *arXiv preprint arXiv:1902.00913*.
- Li, D.; Wei, C.; Li, S.; Zou, J.; Qin, H.; and Liu, Q. 2024. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*.
- Li, H.; Wu, H.; and Chen, B. 2025. NeuralDiffuser: Neuroscience-inspired Diffusion Guidance for fMRI Visual Reconstruction. *IEEE Transactions on Image Processing*.
- Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625.
- Liu, Q.; Nickel, M.; and Kiela, D. 2019. Hyperbolic graph neural networks. *Advances in neural information processing systems*, 32.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Mandica, P.; Franco, L.; Kallidromitis, K.; Petryk, S.; and Galasso, F. 2025. Hyperbolic learning with multimodal large language models. In *European Conference on Computer Vision*, 382–398. Springer.
- Mathis, M. W.; Rotondo, A. P.; Chang, E. F.; Tolia, A. S.; and Mathis, A. 2024. Decoding the brain: From neural representations to mechanistic models. *Cell*, 187(21): 5814–5832.
- Miyawaki, Y.; Uchida, H.; Yamashita, O.; Sato, M.-a.; Morito, Y.; Tanabe, H. C.; Sadato, N.; and Kamitani, Y. 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5): 915–929.
- Naselaris, T.; Kay, K. N.; Nishimoto, S.; and Gallant, J. L. 2011. Encoding and decoding in fMRI. *NeuroImage*, 56(2): 400–410.

- Naselaris, T.; Prenger, R. J.; Kay, K. N.; Oliver, M.; and Gallant, J. L. 2009. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6): 902–915.
- Nickel, M.; and Kiela, D. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30.
- Oota, S. R.; Chen, Z.; Gupta, M.; Bapi, R. S.; Jobard, G.; Alexandre, F.; and Hinaut, X. 2023. Deep neural networks and brain alignment: Brain encoding and decoding (survey). *arXiv preprint arXiv:2307.10246*.
- Pal, A.; van Spengler, M.; di Melendugno, G. M. D.; Flaborea, A.; Galasso, F.; and Mettes, P. 2024. Compositional Entailment Learning for Hyperbolic Vision-Language Models. *arXiv preprint arXiv:2410.06912*.
- Peng, W.; Varanka, T.; Mostafa, A.; Shi, H.; and Zhao, G. 2021. Hyperbolic deep neural networks: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 44(12): 10023–10044.
- Pollen, D. A. 1999. On the neural correlates of visual perception. *Cerebral cortex*, 9(1): 4–19.
- Schirmer, R. T.; Springenberg, J. T.; Fiederer, L. D. J.; Glasstetter, M.; Eggenberger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; and Ball, T. 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, 38(11): 5391–5420.
- Schrodi, S.; Hoffmann, D. T.; Argus, M.; Fischer, V.; and Brox, T. 2024. Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. *arXiv preprint arXiv:2404.07983*.
- Scotti, P.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Dempster, A.; Verlinde, N.; Yundler, E.; Weisberg, D.; Norman, K.; et al. 2023. Reconstructing the mind’s eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36: 24705–24728.
- Scotti, P. S.; Tripathy, M.; Villanueva, C. K. T.; Kneeland, R.; Chen, T.; Narang, A.; Santhirasegaran, C.; Xu, J.; Naselaris, T.; Norman, K. A.; et al. 2024. MindEye2: shared-subject models enable fMRI-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, 44038–44059.
- Shen, G.; Zhao, D.; He, X.; Feng, L.; Dong, Y.; Wang, J.; Zhang, Q.; and Zeng, Y. 2025. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. *Advances in Neural Information Processing Systems*, 37: 98083–98110.
- Song, Y.; Liu, B.; Li, X.; Shi, N.; Wang, Y.; and Gao, X. 2024. Decoding Natural Images from EEG for Object Recognition. In *The Twelfth International Conference on Learning Representations*.
- Srinivasan, R.; Winter, W. R.; Ding, J.; and Nunez, P. L. 2007. EEG and MEG coherence: measures of functional connectivity at distinct spatial scales of neocortical dynamics. *Journal of neuroscience methods*, 166(1): 41–52.
- Tifrea, A.; Becigneul, G.; and Ganea, O.-E. 2019. Poincaré Glove: Hyperbolic Word Embeddings. In *International Conference on Learning Representations*.
- Wang, Z.; and Ji, H. 2022. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5350–5358.
- Wu, H.; Li, Q.; Zhang, C.; He, Z.; and Ying, X. 2025. Bridging the Vision-Brain Gap with an Uncertainty-Aware Blur Prior. *arXiv preprint arXiv:2503.04207*.
- Yang, H.; Gee, J.; and Shi, J. 2024. Brain decodes deep nets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23030–23040.
- Zhang, K.; He, L.; Jiang, X.; Lu, W.; Wang, D.; and Gao, X. 2025. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14486–14493.