# BrainLoc: Brain Signal-Based Object Detection with Multi-modal Alignment

**Jiaqi Duan[†1], Xiaoda Yang[†2], Kaixuan Luan[§2], Hongshun Qiu[3], Weicai Yan[2], Xueyi Zhang[4], Youliang Zhang[5], Zhaoyang Li[7], Donglin Huang[2], JunYu Lu[6], Ziyue Jiang[2], Xifeng Yang[*1],**

[1]Suzhou University of Technology [2]Zhejiang University, [3]Beijing University of Technology, [4]National University of Defense Technology, [5]Tsinghua University, [6]China Merchants Research Institute of Advance Technology, [7]USTC

**Correspondence:** xfyang@cslg.edu.cn

## Abstract

Object detection is a core challenge in computer vision. Traditional methods primarily rely on intermediate modalities such as text, speech, or visual cues to interpret user intent, leading to inefficient and potentially distorted expressions of intent. Brain signals, particularly fMRI signals, emerge as a novel modality that can directly reflect user intent, eliminating ambiguities introduced during modality conversion. However, brain signal-based object detection still faces challenges in accuracy and robustness. To address these challenges, we present BrainLoc, a lightweight object detection model guided by fMRI signals. First, we employ a multi-modal alignment strategy that enhances fMRI signal feature extraction by incorporating various modalities including images and text. Second, we propose a cross-domain fusion module that promotes interaction between fMRI features and category features, improving the representation of category information in fMRI signals. Extensive experiments demonstrate that BrainLoc achieves state-of-the-art performance in brain signal-based object detection tasks, showing significant advantages in both accuracy and convenience.

## 1 Introduction

Current intelligent systems typically rely on intermediate modalities such as speech (Shi et al., 2022; Fu et al., 2024; Cheng et al., 2025), text (Radford et al., 2021; Yan et al., 2025), and images (Yang et al., 2024b; Yan et al., 2024) to understand human intent, but these modalities are merely indirect channels and abstract expressions of consciousness.

This indirectness is particularly evident in object detection tasks within computer vision. While traditional text-based detection systems (such as Grounding DINO (Liu et al., 2023)) have achieved high accuracy, they face significant limitations in practical applications: users need to describe texture and spatial information through text, a process that requires careful thinking and manual input, resulting in high cognitive costs. Moreover, the modality conversion process may lead to distortion and ambiguity of user intent.

To address these challenges, functional Magnetic Resonance Imaging (fMRI) (Belliveau et al., 1991) signals demonstrate unique advantages. fMRI is a non-invasive brain imaging technique that records neural activity patterns under visual stimulation by measuring blood oxygen level-dependent signal changes. This type of signal can directly reflect user intent, offering a promising pathway toward more natural and intuitive human-computer interaction. Compared to Electroencephalogram (EEG) (Craik et al., 2019; Zhao et al., 2024a), fMRI possesses a higher information entropy, enabling it to capture richer semantic and visual details. This advantage makes fMRI particularly well-suited for supporting tasks such as object detection.

Imagine a future where people wear smart glasses that can capture current scenes in real-time and, based on the user's brain responses, quickly and accurately locate target objects, highlighting them on the display. This technology has significant applications not only in daily life scenarios but also in military target detection, industrial automation, and medical assistance.

To this end, we combine the advantages of traditional object detection systems and brain signals to propose BrainLoc. As shown in Fig. 1, even in complex scenes containing multiple similar targets (such as multiple apples), our system can accurately identify and locate the specific target (such as a red apple) that the user has in mind. The key idea is to build a lightweight feature extractor that enhances fMRI signal comprehension through multi-modal alignment strategies, enabling it to capture fine-

[*]Corresponding author
[†]These authors contributed equally
[§]Interned at Zhejiang University



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21652–21662
November 4-9, 2025 

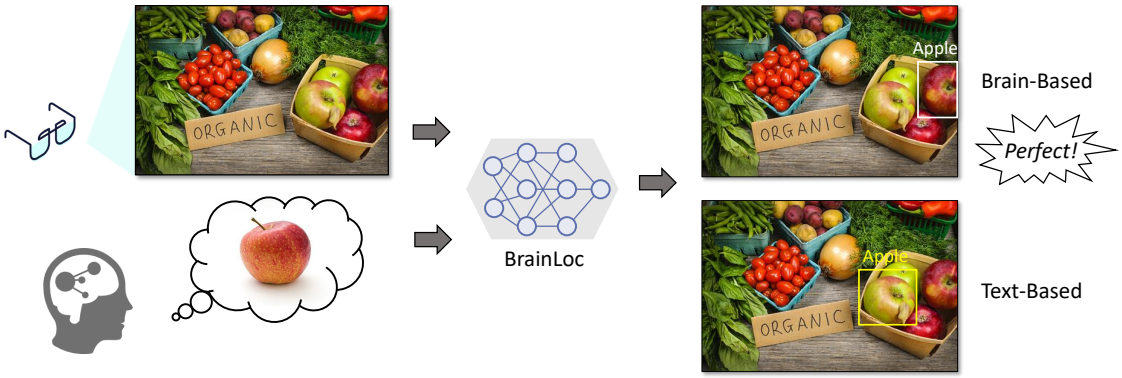


Figure 1: The main ability of our model. Our model takes fMRI signal and a visual scene as input, allowing for the detection of the desired object. Traditional models, relying on text descriptions, might identify a green apple because it stands out more. In contrast, our system incorporates brainwave information, enabling it to locate the specific red apple we want.

grained visual information from user intent. Additionally, we design a Cross-domain fusion module that organically combines visual information (such as color) from fMRI features with semantic information through attention mechanisms, thereby improving localization accuracy. Experimental results demonstrate that BrainLoc not only achieves state-of-the-art performance in localization accuracy but also shows unique advantages in handling complex scenes and fine-grained recognition tasks.

Our main contributions are as follows:

- We propose a lightweight object detection framework for fMRI signals. This framework directly extracts localization features from fMRI signals, avoiding image generation processes and reducing model parameters to 1/10.
- We ues a multi-modal alignment strategy. By simultaneously aligning fMRI signals with image features and dual-level text features in the CLIP feature space, we significantly enhance feature extraction effectiveness.
- We introduce an efficient cross-domain fusion module. This module integrates visual domain fMRI features and semantic domain category features through attention mechanisms, leveraging the complementary advantages of both domains to improve object detection accuracy.

## 2 Related Work

### 2.1 Brain Signal Comprehension

Recently, visual models based on brain signals (Lin et al., 2022; Scotti et al., 2024; Yang et al., 2025a,b, 2024a) have made remarkable progress. Early methods were based on image generation. Mindreader (Lin et al., 2022) projected fMRI data into the CLIP space that embeds images and captions and then used the LAFITE (Zhou et al., 2022) model adjusted with the unconditional StyleGAN2 (Karras et al., 2020) framework to perform image reconstruction. Brain-Diffuser (Ozcelik and VanRullen, 2023a) leverages pretrained diffusion models (Rombach et al., 2022) to generate high-resolution images from fMRI signals, ensuring semantic consistency. MindEye (Scotti et al., 2024), on the other hand, employs a dual-pathway model to separately handle fine-grained details and global semantics extracted from fMRI signals. By aligning multi-level features, it achieves high-quality visual reconstruction. However, these generative methods often require complex image reconstruction processes. Therefore, methods that directly encode fMRI signals have begun to attract attention. MinD-Vis (Chen et al., 2023c) uses a masked autoencoder to extract features directly from fMRI signals. Brainformer (Nguyen et al., 2025) aligns visual features and brain cognitive features through contrastive learning and has been successfully applied to object detection tasks. UMBRE (Xia et al., 2024) focuses on building a shared embedding space connecting fMRI signals with multimodal data like images and text, utilizing unsupervised learning approaches. This approach seeks to improve the robustness and generalization performance in decoding neural signals. These works have demonstrated that it is feasible to extract features directly from fMRI signals for downstream visual tasks without going through an image gener-

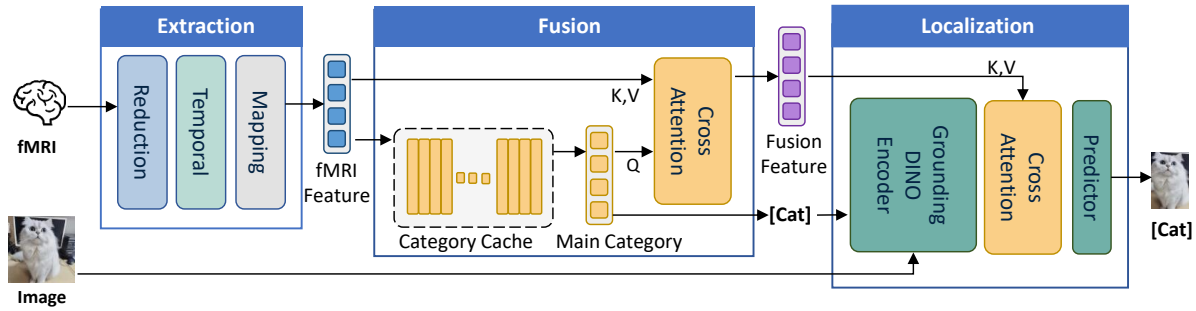


Figure 2: **An overview of the BrainLoc**. BrainLoc consists of three main components: a feature extraction module that transforms fMRI signals into embeddings, a fusion module that performs cross-domain feature integration and category prediction, and a localization module that combines the background image with fused features to generate final detection results. During inference, the model requires only fMRI signals and the background image as input.

ation step.

### 2.2 Object Detection

Carion introduced the DETR (Carion et al., 2020) detection model, which was subsequently improved by Chen (Chen et al., 2023b; Dai et al., 2021; Gao et al., 2021; Jia et al., 2023; Meng et al., 2021; Wang et al., 2022; Zhu et al., 2020; Yang et al., 2024c) through various enhancements, including the Group DETR (Chen et al., 2023b) and Deformable DETR models (Zhu et al., 2020). However, these models primarily operate within a closed set of predefined categories, making it challenging to extend their application to new categories. This limitation has prompted research into open-set object detection, which utilizes existing bounding box annotations for training and employs language generalization to achieve the detection of arbitrary categories. OV-DETR (Zareian et al., 2021) employs image and text embeddings encoded by the CLIP model as query requests to decode specific class boxes within the DETR framework (Carion et al., 2020). ViLD (Gu et al., 2021) extracts knowledge from a CLIP teacher model and transfers it to an R-CNN-like detector, allowing the learned region embeddings of the detector to incorporate text and images inferred from the teacher model. Shikra (Chen et al., 2023a) takes advantage of pretrained models (Rombach et al., 2022) to perform language-driven visual localization, accurately identifying and pinpointing specific objects or regions described in natural language. Grounding DINO (Liu et al., 2023) merges self-supervised visual features similar to DINO with the DETR detection architecture, while incorporating textual cues in an end-to-end manner. This approach achieves exceptional open-set object detection, enabling it to detect items defined by any textual description.

## 3 Method

As shown in Fig. 2, we introduce the BrainLoc, which can detect objects in people's minds with only images and fMRI signals as input. First, we employ a multi-modal alignment strategy to train a lightweight feature extractor that converts fMRI signals into features (see Sec. 3.1). Second, we design a Cross-domain fusion module that deeply integrates fMRI features with category features and predicts the category of the detected object (see Sec. 3.2). Finally, we read the background image and combine it with previously obtained information to output the localization result (see Sec. 3.3).

### 3.1 fMRI Feature Extraction

We design a lightweight feature extractor to process high-dimensional fMRI signals derived from cortical tissue. The extractor begins with a Reduction Module built upon convolutional layers, which effectively compresses feature dimensions to facilitate subsequent high-level semantic information extraction. Considering the temporal characteristics of fMRI data, we construct a Temporal Module using Residual and Transformer architectures to capture long-term dependencies within the signals. Finally, we develop a Mapping Module based on Qformer to achieve cross-modal alignment, mapping fMRI signals into a unified feature space.

This extractor maps the flattened spatial patterns of fMRI signals into the image embedding latent space of a pretrained CLIP model. To enhance the comprehensiveness and robustness of feature extraction, we adopt a multi-modal alignment strategy that simultaneously achieves align-

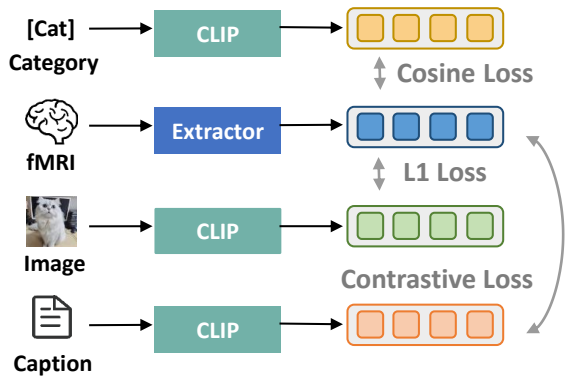


Figure 3: We introduce a variety of modality data and design a variety of loss functions to train the extractor.

ment of fMRI, image, and dual-level text modalities (both sentence-level and word-level) in the CLIP feature space. Compared to existing methods, our approach eliminates computationally intensive image generation processes, significantly reducing model complexity with parameters only 1/10 of (Scotti et al., 2024).

As shown in Fig. 3, we design a series of loss functions to guide the training of the extractor.Taking a batch as an example, we denote the obtained fMRI features as $f^i_{\text{fMRI}}$ , where $i$ represents the index of the feature in the batch.

**fMRI-Cat Loss** Let the encoded result of the text for each category be $f^c_{cat}$, where $c \in R^C$, with $C$ representing the number of categories. For each image in the dataset, the encoded result is $f^i_{img}$. We calculate the matrix $SIM \in R^{N \times C}$ through:

$$sim(i,c) = \frac{f^c_{cat} \cdot f^i_{img}}{\|f^c_{cat}\|_2 \cdot \left\|f^i_{img}\right\|_2} \in (0,1) \quad (1)$$

Taking the maximum value across columns, $\hat{c} = \arg\max(SIM) \in R^N$, then $\hat{c}_i$ is designated as the main category for the i-th image. Then the alignment loss between fMRI and the main category can be expressed as:

$$L_{fMRI-cat} = \frac{f^{\hat{c}_i}_{cat} \cdot f^i_{fMRI}}{\left\|f^{\hat{c}_i}_{cat}\right\|_2 \cdot \left\|f^i_{img}\right\|_2} \quad (2)$$

**fMRI-Img Loss** The fMRI features and image features are aligned in L1 space. Define $\tau$ as a temperature ratio, which is a hyperparameter that weights the degree of fMRI-image alignment. Then the $L_{fMRI-img}$ can be expressed as:

$$L_{fMRI-img} = \mathcal{F}\left(\|f^i_{fMRI} - f^i_{img}\|_1 \cdot \tau\right) \quad (3)$$

where $\mathcal{F}$ is a mapping function. $L_{fMRI-img}$ focuses the model on the information concerning image, restoring the features such as color.

**fMRI-Cap Loss** Contrastive learning is an effective representation learning method that learns representations across multi-modal data by maximizing the cosine similarity of positive sample pairs and minimizing the similarity of negative sample pairs. Previous research suggests that combining contrastive learning with neural data can yield significant benefits (Défossez et al.; Schneider et al., 2023). CLIP is an example of a multi-modal contrastive model that maps images and text captions to a shared embedding space. BrainLoc is designed to incorporate fMRI as an additional modality into the embedding space of a pretrained CLIP model, while keeping the CLIP image space fixed, similar to the approach used in locked-image text tuning (LiT). We utilize the CLIP loss as our contrastive objective. Let the embedding representation of the caption after processing by CLIP be denoted as $f_{cap}$. Then, $L_{fMRI-cap}$ can be expressed as:

$$S = \frac{f_{fMRI} \cdot f^T_{cap}}{\tau} \quad (4)$$

$$r_i = \frac{e^{S_{ii}}}{\sum_j e^{S_{ij}}}; c_i = \frac{e^{S_{ii}}}{\sum_j e^{S_{ji}}} \quad (5)$$

$$L_{fMRI-cap} = -\frac{1}{N} \cdot \sum_{i=1}^{N} [\lambda \cdot log(r_i) + (1-\lambda) \cdot log(c_i)] \quad (6)$$

where $\tau$ is a temperature hyperparameter, and $\lambda$ controls the degree of contrastive learning in two directions.

Then the loss of the extract can be expressed as:

$$L_{total} = \lambda_1 \cdot L_{fMRI-cat} + \lambda_2 \cdot L_{fMRI-img} + \lambda_3 \cdot L_{fMRI-cap} \quad (7)$$

Through the experiment, we set $\lambda_1$ to 0.5, and $\lambda_2$ and $\lambda_3$ to 0.25. This configuration effectively integrates the information from different modalities while avoiding the excessive dominance of a specific loss term.

### 3.2 Cross-domain Fusion

After projecting fMRI signals into the CLIP space, we observe that the visual domain (fMRI features) and semantic domain (category features) share the same latent space, with each domain exhibiting

distinct advantages: semantic domain features possess higher semantic purity, while visual domain fMRI features contain richer perceptual information. For instance, fMRI features not only encode visual details of targets (such as the color of an apple) but also preserve spatial layout and perceptual information. However, due to limitations in current feature extraction techniques, the semantic information may exhibit some ambiguity. Therefore, we propose a cross-domain feature fusion strategy to achieve mutual enhancement of visual perception and semantic understanding.

Specifically, in the semantic domain, we first use CLIP to encode text from common categories into feature vectors $f_{cat}^{c} \in R^{N \times D}$ and store them in the cache, where N represents the total number of categories and D represents the feature dimension. Subsequently, we compute the similarity between visual domain fMRI features and semantic domain category features:

$$f_{cat}^{\hat{c}} = \arg\max_{c \in C}(f_{fMRI}^{i} \cdot f_{cat}^{c}) \tag{8}$$

Through this cross-domain matching, we obtain precise category information of target objects (such as "cat") and use it along with the original image as input to the localization module. To fully utilize the complementary information from both domains, we employ a cross-attention mechanism for feature fusion, using the semantic domain category vector $f_{cat}^{i}$ as key and value, and the visual domain fMRI feature $f_{fMRI}^{i}$ as query, ensuring that the generated feature vector $f_{fusion}^{i}$ effectively integrates the advantageous information from both domains. The $f_{fusion}^{i}$ is then fed into the localization module for subsequent object localization tasks.

### 3.3 Localization Module

Our localization module takes the image and category signals generated from fMRI as input, which are initially processed through Grounding DINO. Specifically, we use the hidden layer outputs from Grounding DINO as queries, and the previously obtained $f_{fusion}^{i}$ as keys and values, integrating features through a cross-attention mechanism. Subsequently, we reconnect Grounding DINO's prediction head to generate target boxes. We denote the candidate boxes of an image as $\{Q_i\}_{i=1}^{n}$, where $n$ is the number of candidate boxes and $Q_i$ is composed of $(x, y, width, height)$. The ground truth boxes for the image are denoted as $\{G_i\}_{i=1}^{m}$, where $m$ is the number of ground truth boxes. We optimize the model's multi-object detection capability directly to improve its performance. First, we maintain a cost matrix $L_{metric}$. To ensure accurate matching, we apply the Hungarian algorithm (Kuhn, 2004) to calculate the total loss based on $L_{metric}$. The cost matrix $L_{metric}$ consists of two components: the classification loss $L_{class}$, which measures the difference between the predicted and true categories, and the IoU loss $L_{IoU}$, which evaluates the overlap between the predicted and ground truth boxes to improve the model's localization accuracy. By combining these two loss components, the model can simultaneously optimize both the target's category prediction and location matching.

**Classification Loss** In object localization tasks, the first priority is to ensure that the model correctly classifies the objects in the candidate boxes. To achieve this, we use the cross-entropy loss function to measure the difference between the predicted and true categories. The classification loss between the i-th candidate box and the j-th ground truth box can be expressed as:

$$L_{class}^{ij} = -\sum_{c=1}^{C} y_{j,c} \log(\hat{p}_{i,c}) \tag{9}$$

where $c$ represents the class, and $y_{j,c}$ indicates whether the j-th ground truth box belongs to class $c$, with $y_{j,c} = 1$ if it does. Similarly, $\hat{p}_{i,c}$ denotes the probability that the i-th candidate box is predicted to belong to class $c$.

**IoU Loss** In addition to ensuring that the predicted candidate boxes match the ground truth boxes in terms of classification, it is also important to accurately localize the objects. For this, we introduce the IoU loss. The formula for calculating the IoU between the i-th candidate box $Q_i$ and the j-th ground truth box $G_j$ in target region $Area$ is:

$$L_{IoU}^{ij} = 1 - \frac{Area(Q_i \cap G_j)}{Area(Q_i \cup G_j)} \tag{10}$$

The total loss matrix is expressed as:

$$L_{metric}^{ij} = \alpha \cdot L_{class}^{ij} + \beta \cdot L_{IoU}^{ij} \tag{11}$$

We set $\alpha$ to 5 and $\beta$ to 2 following the DETR.

By applying the Hungarian Algorithm in the Appendix, the minimum total loss can be achieved. This approach ensures a globally optimal solution rather than a local optimum.

## 4 Experiment

### 4.1 Dataset

This study primarily utilizes two datasets: the Natural Scenes Dataset (NSD (Allen et al., 2022)) and the Generic Object Decoding (GOD (Horikawa and Kamitani, 2017)) dataset. The NSD contains fMRI data collected from 8 subjects while viewing 10,000 natural scene images. The GOD dataset comprises fMRI data from 5 subjects viewing images of 200 object categories.

Through studies of human visual cognitive mechanisms, we observe that people tend to mentally focus on the target object rather than the entire scene during object search. Based on this observation, we analyzed the characteristics of two datasets: The NSD dataset, derived from COCO (Lin et al., 2014), better represents real-world scenarios but contains multiple object categories within each image, resulting in higher noise levels in fMRI signals. In contrast, the GOD dataset (based on ImageNet (Deng et al., 2009)) contains single-object images, producing fMRI signals that better align with targeted object search patterns.

Therefore, we designed a two-stage training strategy to leverage the advantages of both datasets: In the first stage, we train the feature extractor using the NSD dataset.In the second stage, based on freezing the parameters of the feature extractor, the fMRI signals from the GOD dataset were paired with the images from the NSD dataset that contained the corresponding objects. According to the default division of the GOD dataset, the training and testing data were obtained, which were used for subsequent training and evaluation of the fusion and localization modules.

### 4.2 Experimental Setup

**Implement Details.** The architecture of our fMRI feature extractor consists of one ConvBlock (Alaeddine and Jihene, 2021), three ResidualBlocks (Goceri, 2019), three TransformerBlocks (Min et al., 2022)(each containing four layers), one Qformer (Zhang et al., 2024), and several MLP layers, totaling 130M trainable parameters. The model was trained for 1,000 epochs on four A800 GPUs. For learning rate scheduling, we employ a LambdaLR (Paszke et al., 2019) with a warm-up period of 100 iterations, where the learning rate starts at zero and increases linearly to the maximum value of 1e-4 during warm-up, then decreases from 1e-4 to a minimum of 1e-7 using cosine annealing over 1,000 iterations. For feature extraction from other modalities, we utilize the ViT/b-32 architecture from the CLIP model.

**Baseline.** Considering the limited number of end-to-end object detection models based on brain signals, we refer to UMBRAE (Xia et al., 2024) and adapt several advanced brain-signal-based image reconstruction methods (Ozcelik and VanRullen, 2023a; Scotti et al., 2024; Xia et al., 2024), as object detection baselines. Specifically, the adaptation procedure consists of two steps: first, reconstructing images from brain signals using these models; second, feeding the reconstructed images into the Shikra model, accompanied by the instruction `"Please interpret this image and provide coordinates [x1,y1,x2,y2] foreach object you mention"` to extract object bounding boxes. In addition, we also conduct comparisons with purely text-driven localization models (Liu et al., 2023; Chen et al., 2023a).

**Metrics.** The evaluation of baseline model performance primarily relies on two metrics: accuracy(acc@m) and Intersection over Union (IoU). Acc@m quantifies the percentage of correctly localized predictions where the IoU between predicted and ground-truth bounding boxes exceeds a predefined threshold m; consistently, we select acc@0.5 throughout our experiments, as it serves as an effective indicator of localization reliability. Concurrently, IoU directly measures the degree of overlap between predicted and ground-truth bounding boxes. To facilitate an in-depth analysis, we adopt a categorization scheme inspired by UMBRAE (Xia et al., 2024), dividing the 80 classes of the COCO dataset into two main groups—"**Salient**" and "**Inconspicuous**"—based on their prominence in natural scenes. The "Salient" category is further subdivided into "**Salient Creatures**" (e.g., humans, animals) and "**Salient Objects**" (e.g., cars, beds, tables), while the "Inconspicuous" category includes items such as backpacks, knives, and toothbrushes. This extensive evaluation approach seeks to verify the robustness and real-world applicability of BrainLoc across diverse object detection scenarios.

### 4.3 Experimental Results

Tab. 1 summarizes the detection results of each model across categories including "All," "Salient," and "Inconspicuous." Among fMRI-based methods, BrainLoc consistently outperforms other baselines. We attribute this superiority primarily to BrainLoc's ability to avoid the inherent informa-

Table 1: **Comparison result**. Text-based models refer to location systems that rely on textual input, offering high accuracy but also incurring significant interaction costs. Brain-based models, on the other hand, achieve localization through brain signals. UMBRAE-S refers to the model trained with a single subject only. Shikra-w/method provides visual grounding results using images produced by reconstruction model based on brain signals.

| Method | All | | Salient | | Salient Creatures | | Salient Objects | | Inconspicuous | |
|---|---|---|---|---|---|---|---|---|---|---|
| | acc@0.5 | IoU | acc@0.5 | IoU | acc@0.5 | IoU | acc@0.5 | IoU | acc@0.5 | IoU |
| Text-based | | | | | | | | | | |
| Grounding DINO (Liu et al., 2023) | 80.16 | 48.66 | 80.44 | 47.19 | 81.06 | 44.08 | 77.07 | 44.50 | 78.63 | 42.47 |
| Shikra (Chen et al., 2023a) | 51.96 | 47.22 | 62.92 | 56.44 | 66.71 | 59.34 | 58.79 | 53.27 | 38.29 | 35.71 |
| fMRI-based | | | | | | | | | | |
| Shikra-w/ Brain-Diffuser (Ozcelik and VanRullen, 2023b) | 17.49 | 19.34 | 27.18 | 27.46 | 38.71 | 34.63 | 14.62 | 19.66 | 5.39 | 9.20 |
| Shikra-w/ MindEye (Scotti et al., 2024) | 15.34 | 18.65 | 23.83 | 26.96 | 29.29 | 31.64 | 17.88 | 21.86 | 4.74 | 8.28 |
| Shikra-w/ DREAM (Xia et al.) | 16.21 | 18.65 | 26.51 | 27.35 | 34.43 | 33.85 | 17.88 | 20.28 | 3.35 | 7.78 |
| Shikra-w/ UMBRAE | 16.83 | 18.69 | 27.10 | 27.55 | 34.14 | 33.65 | 19.44 | 20.92 | 4.00 | 7.64 |
| UMBRAE-S | 13.72 | 17.56 | 21.52 | 25.14 | 26.00 | 29.06 | 16.64 | 20.88 | 4.00 | 8.08 |
| UMBRAE (Xia et al., 2024) | 18.93 | 21.28 | 30.23 | 30.18 | 39.57 | 36.64 | 20.06 | 23.14 | 4.83 | 10.18 |
| **BrainLoc** | **64.13** | **67.08** | **67.18** | **67.79** | **70.11** | **68.65** | **61.94** | **62.69** | **61.79** | **63.92** |


Figure 4: The visualization of the BrainLoc. The numbers on the bounding box represent the confidence level of the target detection within the range of (0, 1).

tion loss encountered in traditional fMRI-to-image reconstruction approaches and its cross-domain fusion module that enables the integration of multi-modal features. Compared with text-based methods, BrainLoc achieves higher IoU scores than Grounding DINO, though it slightly trails behind in terms of acc@0.5. This discrepancy likely arises due to the distribution characteristics of Grounding DINO's (Liu et al., 2023) predictions around the IoU threshold of 0.5, where a higher acc@0.5 may mask its lower average localization accuracy. Additionally, text-only localization approaches face inherent challenges in fine-grained discrimination tasks, such as distinguishing between differently colored apples, thus potentially constraining their IoU scores.

The visualization results in Fig. 4 further validate the effectiveness of BrainLoc, highlighting its ability to accurately identify and localize the positions of objects.

### 4.4 Ablation Study

To evaluate the individual contributions of each component in the BrainLoc model, we conducted a series of ablation studies, with results summarized in the table. These experiments primarily focused on variations involving the following three core modules:

**Feature Extraction Module:** We explored the effects of three variations: removing multi-modal alignment, eliminating main category optimization, and replacing contrastive learning with a cosine loss. Results consistently indicate performance deterioration across these modifications. Particularly, removal of main category optimization significantly reduced performance in the "Salient" category, highlighting the essential contribution of these components and strategies.

**Fusion Module:** We evaluated this module's role by removing it entirely, thus forcing the localization module to rely solely on category-level

Table 2: **Ablation study**. w/o Fusion Module indicates that only the Category is provided; w/o Multi-modal Alignment refers to using a refined structure of (Scotti et al., 2024) to extract features; w/o Main Category refers to the absence of main category optimization; and w/ Shikra means using Shikra as the localization module.

| Method | All | | Salient | | Salient Creatures | | Salient Objects | | Inconspicuous | |
|---|---|---|---|---|---|---|---|---|---|---|
| | acc@0.5 | IoU | acc@0.5 | IoU | acc@0.5 | IoU | acc@0.5 | IoU | acc@0.5 | IoU |
| w/o Multimodal Alignment | 63.52 | 65.89 | 66.59 | 66.38 | 69.45 | 67.69 | 61.37 | 61.77 | 62.11 | 64.21 |
| w/o Main Category | 60.97 | 63.25 | 63.92 | 63.72 | 66.67 | 64.98 | 58.91 | 59.29 | 59.62 | 61.64 |
| w/o Contrastive Learning | 62.24 | 64.57 | 65.25 | 65.05 | 68.06 | 66.33 | 60.14 | 60.53 | 60.86 | 62.92 |
| w/o Fusion Module | 60.92 | 63.72 | 63.82 | 64.4 | 66.60 | 65.21 | 58.84 | 59.55 | 58.70 | 60.72 |
| w/ Shikra | 44.16 | 40.13 | 53.48 | 47.97 | 56.70 | 50.43 | 49.97 | 45.27 | 32.54 | 30.35 |
| BrainLoc | 64.13 | 67.08 | 67.18 | 67.79 | 70.11 | 68.65 | 61.94 | 62.69 | 61.79 | 63.92 |

information for predictions. The considerable drop in performance, as shown in the "w/o Fusion Module" row of the table, confirms the module's pivotal role in effectively integrating multi-source information for accurate object localization.

**Localization Module:** To investigate how the choice of the foundational localization model affects BrainLoc's overall performance and its compatibility with other models, we replaced the original localization module with Shikra ("w/ Shikra"). This replacement led to a decline in performance, indicating that BrainLoc's overall efficacy is partially dependent on the capabilities of its underlying localization model. Nonetheless, it underscores the modular flexibility of the BrainLoc architecture, suggesting potential adaptability for future model improvements and iterative upgrades.

In general, these experiments confirm the rationality of BrainLoc's current design and the necessity of each module. Given that the effectiveness of the related fundamental techniques (Défossez et al., 2023) has already been extensively validated in prior studies [(Xia et al., 2024; Lin et al., 2022; Scotti et al., 2024)], we refrain from redundant discussions here.

### 4.5 Discussion

**Why choose this combination of loss functions?** If both $L_{fMRI-img}$ and $L_{fMRI-cap}$ employ contrastive learning, the model collapses, and training fails. When both $L_{fMRI-img}$ and $L_{fMRI-cap}$ employ $L_1$ loss, the performance is inferior to using contrastive learning for $L_{fMRI-cap}$. Because the CLIP model employs contrastive learning to align text and image modalities, we also adopt contrastive learning to map fMRI signals into the CLIP space. We choose cosine similarity for $L_{fMRI-cat}$ because the downstream retrieval tasks are also based on cosine similarity.

**In Fig. 1, both 'green apple' and 'red apple' are labeled simply as 'apple.' So, why can our model distinguish between them during localization while text-based models cannot?** We first align brain signals with image and other modalities, thus assuming they capture the visual information, including color details. The brain signals and textual signals are then fused through an attention mechanism and fed into the localization network. This process embeds color information into the model, enabling the model to locate the items exactly as the user intended, rather than just the categories. By contrast, text-based models trained on datasets like COCO have coarser label granularity, distinguishing only 'apple' as a category without further specifying details like color.

## 5 Conclusion

In this paper, we presented BrainLoc, a novel lightweight brain-based object detection model that leverages fMRI signals to identify and locate target objects in complex scenes. Experimental results demonstrate that BrainLoc achieves SOTA performance in brain-based localization tasks, combining the precision of traditional systems with the convenience of brain-based approaches. This significant advancement highlights the potential of brain-computer interface (BCI) technologies (Zhao et al., 2024b) in various applications, including assistive technologies, military target detection, and industrial automation. Given that fMRI data typically contain less signal noise compared to EEG, our research primarily focused on fMRI. In future work, we plan to shift our efforts toward EEG to enhance real-time applications. overall, BrainLoc

represents a significant step forward in brain-based object detection, offering high accuracy and portability, and is expected to inspire further research and development in the field of BCI.

## Limitations

Although our work has shown promising results in brain signal-based object detection, several limitations remain. The performance of our model heavily depends on the quality and quantity of brain signal data. While fMRI signals can accurately reflect brain activity, their high collection costs make it difficult to obtain large-scale, high-quality fMRI datasets. Meanwhile, significant individual differences in brain activity patterns may affect the model's generalization ability across different subjects. Additionally, the current datasets (NSD and GOD) are collected in controlled laboratory settings, which may not fully capture the complexity and variability of real-world scenarios.

## References


Hmidi Alaeddine and Malek Jihene. 2021. A convblock for convolutional neural networks. In *Deep Learning Applications in Medical Imaging*, pages 100–113. IGI Global.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. 2022. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126.

Jack W Belliveau, David N Kennedy, Robert C McKinstry, Bradley R Buchbinder, Robert M Weisskoff, Mark S Cohen, JM Vevea, Thomas J Brady, and Bruce R Rosen. 1991. Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, 254(5032):716–719.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023a. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. 2023b. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6633–6642.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. 2023c. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720.

Xize Cheng, Dongjie Fu, Xiaoda Yang, Minghui Fang, Ruofan Hu, Jingyu Lu, Bai Jionghao, Zehan Wang, Shengpeng Ji, Rongjie Huang, Linjun Li, Yu Chen, Tao Jin, and Zhou Zhao. 2025. Omnichat: Enhancing spoken dialogue systems with scalable synthetic data for diverse scenarios. *Preprint*, arXiv:2501.01384.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. 2019. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001.

Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. 2021. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech from non-invasive brain recordings, august 2022. *URL http://arxiv. org/abs/2208.12266*.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. 2023. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dongjie Fu, Xize Cheng, Xiaoda Yang, Wang Hanting, Zhou Zhao, and Tao Jin. 2024. Boosting speech recognition robustness to modality-distortion with contrast-augmented prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 3838–3847, New York, NY, USA. Association for Computing Machinery.

Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. 2021. Fast convergence of detr with spatially modulated co-attention. 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 3601–3610.

Evgin Goceri. 2019. Analysis of deep networks with residual blocks and different activation functions: classification of skin diseases. In *2019 Ninth international conference on image processing theory, tools and applications (IPTA)*, pages 1–6. IEEE.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Tomoyasu Horikawa and Yukiyasu Kamitani. 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037.

Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. 2023. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19702–19712.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

Harold W Kuhn. 2004. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1):7–21.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. 2022. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. 2021. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660.

Erxue Min, Runfa Chen, Yatao Bian, Tingyang Xu, Kangfei Zhao, Wenbing Huang, Peilin Zhao, Junzhou Huang, Sophia Ananiadou, and Yu Rong. 2022. Transformer for graphs: An overview from architecture perspective. *arXiv preprint arXiv:2202.08455*.

Xuan-Bac Nguyen, Xin Li, Pawan Sinha, Samee U Khan, and Khoa Luu. 2025. Brainformer: Mimic human visual brain functions to machine vision models via fmri. *Neurocomputing*, 620:129213.

Furkan Ozcelik and Rufin VanRullen. 2023a. Natural scene reconstruction from fmri signals using generative latent diffusion. *Preprint*, arXiv:2303.05334.

Furkan Ozcelik and Rufin VanRullen. 2023b. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. 2023. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. 2024. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.

Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. 2022. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2567–2575.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system—supplementary material—.

Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. 2024. Umbrae: Unified multimodal decoding of brain signals. *arXiv preprint arXiv:2404.07202*.

Weicai Yan, Wang Lin, Zirun Guo, Ye Wang, Fangming Feng, Xiaoda Yang, Zehan Wang, and Tao Jin. 2025. Diff-prompt: Diffusion-driven prompt generator with mask supervision. In *The Thirteenth International Conference on Learning Representations*.

Weicai Yan, Ye Wang, Wang Lin, Zirun Guo, Zhou Zhao, and Tao Jin. 2024. Low-rank prompt interaction for continual vision-language retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8257–8266.

Xiaoda Yang, Xize Cheng, Jiaqi Duan, Hongshun Qiu, Minjie Hong, Minghui Fang, Shengpeng Ji, Jialong Zuo, Zhiqing Hong, Zhimeng Zhang, et al. 2024a. Audiovsr: Enhancing video speech recognition with audio data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15352–15361.

Xiaoda Yang, Xize Cheng, Minghui Fang, Hongshun Qiu, Yuhang Ma, JunYu Lu, Jiaqi Duan, Sihang Cai, Zehan Wang, Ruofan Hu, et al. 2025a. Multimodal conditional retrieval with high controllability. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 3577–3585.

Xiaoda Yang, Xize Cheng, Dongjie Fu, Minghui Fang, Jialung Zuo, Shengpeng Ji, Tao Jin, and Zhou Zhao. 2024b. Synctalklip: Highly synchronized lip-readable speaker generation with multi-task learning. In *ACM Multimedia 2024*.

Xiaoda Yang, Xize Cheng, Dongjie Fu, Minghui Fang, Jialung Zuo, Shengpeng Ji, Zhou Zhao, and Jin Tao. 2024c. Synctalklip: Highly synchronized lip-readable speaker generation with multi-task learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8149–8158.

Xiaoda Yang, Jiayang Xu, Kaixuan Luan, Xinyu Zhan, Hongshun Qiu, Shijun Shi, Hao Li, Shuai Yang, Li Zhang, Checheng Yu, et al. 2025b. Omnicam: Unified multimodal video generation via camera control. *arXiv preprint arXiv:2504.02312*.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. 2024. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Peng Zhao, Ruicong Wang, Zijie Lin, Zexu Pan, Haizhou Li, and Xueyi Zhang. 2024a. Ensemble deep learning models for eeg-based auditory attention decoding. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 339–343. IEEE.

Peng Zhao, Ruicong Wang, Xueyi Zhang, Mingrui Lao, and Siqi Cai. 2024b. Binary-temporal convolutional neural network for multi-class auditory spatial attention detection. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.

Y Zhou, R Zhang, C Chen, C Li, C Tensmeyer, T Yu, J Gu, J Xu, and T Sun. 2022. Lafite: Towards language-free training for text-to-image generation. arxiv. *arXiv preprint arXiv:2111.13792*.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

# A Appendix

## A.1 Algorithm for Optimal Loss Minimization

We introduce Alg. 1, which minimizes the total loss matrix by balancing classification and IoU losses to ensure a globally optimal solution for object detection.

**Algorithm 1** Hungarian Algorithm for Maximum Matching Problem

**Input:** IoU matrix $M \in R^{m \times n}$
**Output:** Optimal matching $(\text{row\_ind}, \text{col\_ind})$
Initialize $M[i][j] = 0, \quad \forall i \in [0, m-1],\ j \in [0, n-1]$
Compute IoU values: $M[i][j] = \text{IoU}(p_i, g_j), \quad \forall i \in [0, m-1],\ j \in [0, n-1]$
**Row reduction:** $M'[i,j] \leftarrow M[i,j] - \min(M[i,:])$
**Column reduction:** $M''[i,j] \leftarrow M'[i,j] - \min(M'[:,j])$
Mark zero elements and check if all zeros can be covered by lines
**if** it is possible to cover all zeros with $m$ lines **then**
    Matching is complete
**else**
    Find the smallest uncovered element $\delta$
    Adjust uncovered elements: $M'''[i,j] \leftarrow M''[i,j] \pm \delta$
**end if**
Repeat marking zero elements and adjusting the matrix until all zeros are covered by $m$ lines
Return the matched row and column indices $(\text{row\_ind}, \text{col\_ind})$