Exploring The Visual Feature Space for Multimodal Neural Decoding
— Supplementary Material —

Weihao Xia ✉ Cengiz Oztireli
University of Cambridge
https://github.com/weihaox/VINDEX

This document includes further analyses on the background knowledge, experiments, and results. We first provide more details Natural Scenes Dataset, vision encoders, and diffusion models in Sec. 1. Sec. 2 provides more implementation details in network architecture and prompt templates. We then detail the benchmark in Sec. 3 and provide additional results, analysis, and visualizations in Sec. 4. We discuss limitations and future directions, including possible choices of visual encoders and hybrid multimodal large language models, in Sec. 5.

1. Background

1.1. NSD Dataset

We use the Natural Scenes Dataset (NSD) [1] for the experiment. NSD is currently the largest released fMRI dataset, featuring detailed brain activity recordings from 8 subjects who passively viewed images sourced from the Common Objects in Context (COCO) dataset [24] for up to 40 hours in an MRI machine. Each image was displayed for three seconds and repeated three times over 30-40 scanning sessions, resulting in 22,000-30,000 fMRI response trials per participant.

We follow the data preprocessing procedure similar to prior brain visual decoding studies [34, 35, 41, 45, 46] based on NSD [1]. Specifically, we use preprocessed fMRI voxels in a 1.8-mm native volume space that corresponds to the "nsdgeneral" brain region. This region is described as the subset of voxels in the posterior cortex that are most responsive to the presented visual stimuli [1]. We train our model on the four subjects (with IDs 1, 2, 5, and 7) who completed all scanning sessions. The training set for each subject consists of 8,859 images and 24,980 fMRI trials, with each image shown up to three times. The remaining 982 images and 2,770 fMRI trials, which are common across all four participants, are used for testing. For fMRI data spanning multiple trials, we calculate the average response as in prior research [34]. Tab. 1 details characteristics of NSD and region of interests (ROIs) included in the fMRI data.

1.2. Vision Encoders

CLIP. CLIP [31], a contrastive language-image pre-training method, has gained significant attention for leveraging softmax contrastive learning on large-scale image-text datasets. As a contrastively pre-trained model, CLIP is widely used in various downstream applications, generating diverse representations for tasks such as object detection and semantic segmentation, and it demonstrates strong performance in zero-shot transfer tasks, including classification and retrieval. It is one of the most popular visual encoders in vision-language models and multimodal large language models, serving as the visual component [5, 25, 48].

DINO. DINO [29] is a self-supervised learning framework known for its ability to learn high-quality visual representations without relying on labeled data. Built on the Vision Transformer (ViT) [10] and utilizing knowledge distillation [15], DINO effectively captures semantic structures in images. It has been widely applied to various computer vision tasks, including object detection, semantic segmentation, and visual grounding. DINO's strong feature representations make it a valuable visual encoder in multimodal contexts, enhancing spatial understanding [8, 37, 42, 52].

SigLIP. SigLIP [49] is a contrastive language-image pre-training framework designed to learn high-quality visual representations using sigmoid loss. Building on the foundation of CLIP, SigLIP incorporates a memory-efficient architecture and optimization strategies that enhance performance. The pre-trained SigLIP also serves as a feature extractor in the visual component of multimodal large language models [48].

Table 1. **Details on NSD.** This table presents the training and test image-fMRI pairs for the four subjects, along with ROIs.

Training	Test	ROIs	Subject ID	Dimension
8,859	982	V1, V2, V3, hV4, VO, PHC, MT, MST, LO, IPS	`sub01`	15,724
			`sub02`	14,278
			`sub05`	13,039
			`sub07`	12,682

1.3. Diffusion Models

The diffusion models [16, 28, 39, 40] typically comprise a forward process and a corresponding reverse process. The forward process adds noise, while the reverse denoising process learns to remove it. The model can operate in either pixel space [16] or latent space [32]. For latent diffusion model, given clean latent tokens $\boldsymbol{z}_0$ drawn from $p(\boldsymbol{z})$, the forward diffusion process is a Markov chain that performs progressive noise addtion to the original sample:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the prior Gaussian distribution, $\beta_t \in (0, 1)$ indicates a pre-defined time-dependent variance schedule at discrete timestep t. For sampling $\boldsymbol{z}_t$ from $\boldsymbol{z}_0$ at an arbitrary timestep t [16], this can be reformulated as

$$\begin{aligned} q(\boldsymbol{z}_t|\boldsymbol{z}_0) &= \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{z}_0, (1-\bar{\alpha}_t)\mathbf{I}), \\ \boldsymbol{z}_t &= \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned} \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_t$. The reverse process in the latent diffusion model learns to denoise the added noise for latent tokens. The reverse process iteratively generates clean tokens $\boldsymbol{z}_0$ from pure noise $\boldsymbol{z}_T$ conditioned on $\mathcal{C}$, as described by

$$\boldsymbol{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\pi(\boldsymbol{z}_t; \mathcal{C}, t)\right) + \sigma_t\boldsymbol{\epsilon}, \tag{3}$$

where a π-parameterized denoiser $\boldsymbol{\epsilon}_\pi$ is trained to predict the added noise during the forward process and σ_t indicates the posterior noise variance. The objective for training the denoiser $\boldsymbol{\epsilon}_\pi$ is

$$\mathcal{L}(\pi, \boldsymbol{z}_0) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[||\boldsymbol{\epsilon}_\pi(\sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}; \mathcal{C}, t) - \boldsymbol{\epsilon}||^2\right]. \tag{4}$$

2. Implementation Details

2.1. Architecture

Our method includes a brain encoder and a multimodal large language model (MLLM). The features from visual encoders, as visual components of the MLLM, are used to train the brain encoder, allowing it to learn to predict brain features from the input brain signals. During inference, given a brain signal as input, the brain encoder predicts brain features, which replace the image features and are fed into the MLLM for instruction-following tasks. The overview is shown in Fig. 1.

Brain Encoder. For the brain encoder, we follow the UMBRAE architecture [45], which has shown to be an effective model for brain-image alignment. This encoder contains subject-specific tokenizers to handle individual subject information and a shared pre-trained trunk to capture the common information across subjects. We train the brain encoder to align the image features from the chosen vision encoder, allowing us to integrate brain features into the MLLM for brain perception tasks. Details on the trained brain encoder with different vision encoder settings and the supported MLLMs are provided in Tab. 2.

Vision Encoder. The vision encoders are used to extract features for training the brain encoder. The vision encoders, as the visual component in MLLMs, are versatile, considering the architecture-conscious LLaVA-based MLLMs [5, 25, 42, 48]. We use four vision encoders, which act as representative feature spaces to demonstrate our brain alignment idea: CLIP-224 (openai/clip-vit-large-patch14) [31], CLIP-336 (openai/clip-vit-large-patch14-336) [31], DINOv2 (facebookresearch/dinov2) [29], and SigLIP-384 (google/siglip-so400m-patch14-384) [49]. The same brain encoder trained with CLIP-224/336 can be used across SE, ME, and NF. DINOv2 is used in ME, while SigLIP is used in AF.

Table 2. **Detailed on Our pre-trained Brain Encoders.** Each pre-trained brain encoder is trained to align with vision features from a target vision encoder, using a specific image size. The predicted brain features have a shape of (batch size, tokens, dimension). The same single pre-trained brain encoder can support multiple off-the-shelf MLLMs. Notably, the trained brain encoders B-CLIP224 and B-CLIP336 are directly compatible with all eleven LLaVA models [25] using different setting available in the MODEL ZOO, without any etra modifications.

Brain Encoder	Vision Encoder	Size	#Token	#Dim	Supported MLLMs
B-CLIP224	CLIP-224 [31]	224	256	1024	LLaVA-1.5/1.6-7B/13B, LLaVa-MoF-7B/13B [42], M3-1.5/1.6-7B/13B [5], Shikra [6]
B-CLIP336	CLIP-336 [31]	336	576	1024	LLaVA-1.5/1.6-7B/13B, LLaVa-MoF-7B/13B [42], M3-1.5/1.6-7B/13B [5], DC [48]
B-DINO224	DINOv2 [29]	224	256	1024	LLaVa-MoF [42]
B-SIGLIP384	SigLIP-384 [49]	384	729	1152	DC [48]

MLLMs. Our method supports various MLLMs with different configurations. An MLLM consists of a vision encoder, a connector, and a base LLM. Once a pre-trained brain encoder is available, brain signals can be input to obtain predicted brain features. As shown in Fig. 1, these brain features are then fed into the connector and LLM for interaction. Tab. 2 presents MLLMs supported by each trained brain encoder. Taking B-CLIP224 as an example, for a batch of `bs` input brain signal from NSD [1], it produces brain-CLIP features of size (`bs`, 256, 1024). The connector then projects these features to (`bs`, 256, 4096) or (`bs`, 256, 5120), depending on whether the LLM is 7B or 13B, respectively. The same tokens, after projection, can be fed into the corresponding LLM or, in the NF setting, downsampled to 144, 36, 9, or 1 token for M3 [5] processing. In the AF setting, the predicted brain features have a size of (`bs`, 729, 1152). After aggregation from dense features across different layers, the size becomes (`bs`, 729, 3456) before being fed into the DC [48] connector and LLM.

Denoiser. For the denoising network, we use a small MLP following [16, 21, 40] with several residual blocks [14]. Each sequentially applies a LayerNorm (LN) [3], two linear layers with SiLU activation in between, and merges with a residual connection. The denoising MLP is conditioned on brain predictions $\mathbf{b}$. The prediction $\mathbf{b}$ is added to the time embedding of the noise schedule at timestep t, serving as the condition for the MLP in LN layers. The diffusion process follows [28]. The noise schedule has a cosine shape, with 1,000 steps at training time. The denoising network predicts the noise vector ϵ [16].

Our denoiser differs from the implementation of diffusion prior [34] in the following aspects: (1) training target: The diffusion prior is applied to CLIP embeddings, whereas ours is applied to features from different visual encoders; (2) training effect: diffusion prior predicts CLIP embeddings through additional operations, while ours is only used for training the denoiser and not during prediction. Given that both features and images have inherent redundancy, we apply a mask, which acts as an implicit form of data augmentation and regularization; (3) structure: our denoiser is more lightweight in comparison.

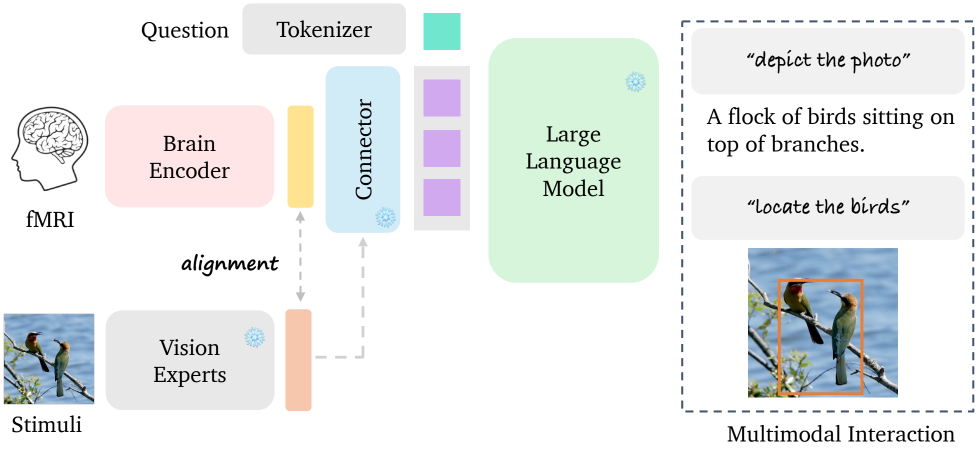


Figure 1. **Method Overview.** Once a pre-trained brain encoder is available, brain signals can be input to obtain predicted brain features. These brain features are then fed into the connector and LLM for multimodal brain interaction. Our method follows a similar overview to UMBRAE [45] but differentiates itself through the use of different vision encoders, alignment strategies, and support for additional tasks.

2.2. Prompt Template

The choice of feature spaces in vision encoders for our model, including SE, ME, AF, and NF, follows similar LLaVA-based architectures and training procedures [25]. Therefore, prompt templates for brief and detail image description [5, 25, 48] used in LLaVA [25], as illustrated in Tab. 3 and Tab. 4, can be directly used as instructions for our concise and descriptive brain captioning. It should be noted that all these prompt templates are only used for inference in our experiments and are not utilized for generating training data. For our method built upon Shika [6], we use the prompts "Describe the image `<image>` as simply as possible" for concise captioning and "Locate `<expr>` in `<image>` and provide its coordinates" for concept localization, where `<expr>` represents the target expression and `<image>` serves as a placeholder for image features. The full prompt template, including the system message, user prompt, and assistant answers, follows:

`SYSTEM MESSAGE.` USER: `<image>` `<instruction>` ASSISTANT: `<answer>`

The tags `<instruction>` and `<answer>` serve as placeholders for human instructions and assistant answers. We use variable templates for different tasks. Prompts for interactive dialogue and complex reasoning can be found in [25].

Table 3. **Prompt for Concise Brain Captioning.** The list of instructions present the same meaning with natural language variance.

- "Describe the image concisely."
- "Provide a brief description of the given image."
- "Offer a succinct explanation of the picture presented."
- "Summarize the visual content of the image."
- "Give a short and clear explanation of the subsequent image."
- "Share a concise interpretation of the image provided."
- "Present a compact description of the photo's key features."
- "Relay a brief, clear account of the picture shown."
- "Render a clear and concise summary of the photo."
- "Write a terse but informative summary of the picture."
- "Create a compact narrative representing the image presented."

Table 4. **Prompt for Descriptive Brain Captioning.** The list of instructions present the same meaning with natural language variance.

- "Describe the following image in detail"
- "Provide a detailed description of the given image"
- "Give an elaborate explanation of the image you seev
- "Share a comprehensive rundown of the presented image"
- "Offer a thorough analysis of the image"
- "Explain the various aspects of the image before you"
- "Clarify the contents of the displayed image with great detail"
- "Characterize the image using a well-detailed description"
- "Break down the elements of the image in a detailed manner"
- "Walk through the important details of the image"
- "Portray the image with a rich, descriptive narrative"
- "Narrate the contents of the image with precision"
- "Analyze the image in a comprehensive and detailed manner"
- "Illustrate the image through a descriptive explanation"
- "Examine the image closely and share its details"
- "Write an exhaustive depiction of the given image"

3. Details on MG-BrainDUB

We describe in the main paper a benchmark for evaluating detailed captions, including annotations and metrics, as well as salient question answering (QA) for complex reasoning. Here, we provide examples to elaborate the evaluation process (Sec. 3.1) and showcase constructed exemplars for detailed captioning (Sec. 3.2) and salient QA (Sec. 3.3).

3.1. Metric Calculation Explanation

We demonstrate in experiments the drawbacks of current traditional rule-based and model-based metrics, necessitating the introduction of new metrics that consider visual elements for long, detailed caption evaluation. Our metric extracts and matches each caption based on precision, recall, and F1 scores for three core visual elements: objects, attributes, and relations [9, 26].To understand the evaluation process, we break down the metric calculation into steps. Give two example captions as the candidate and reference descriptions, we parse the description into a list containing tuples of objects, attributes, and relations, following the steps below:

Candidate “A red car and a white truck are driving down a city street lined with green trees. Tall buildings in the background.”

Reference “A peaceful beach with soft white sand stretching along the coastline, where turquoise ocean waves gently roll onto the shore. Several people are sunbathing near the water while others are playing volleyball in the distance.”

Object Extraction. This step identifies and lists all entities or objects mentioned in the description. For the given candidate caption, we get a list of object labels: [‘building’, ‘city street’, ‘truck’, ‘tree’, ‘car’].

Attribute Mapping. This step identifies attributes associated with each object, which describe their properties or characteristics. The attribute mapping for the caption is a dictionary mapping object labels to their attributes as follows: {‘car’: {‘red’}, ‘tree’: {‘green’}, ‘truck’: {‘white’}, ‘building’: {‘tall’}}. Each object is paired with its respective attributes, providing essential information for evaluating the model’s ability to recognize both the objects and their attributes.

Relation Extraction. This discerns the relationships between different objects in the scene, which describe their spatial or functional connections. In the example caption, the relationships are: {(‘truck’, ‘drive down’, ‘city street’), (‘car’, ‘drive down’, ‘city street’), (‘building’, ‘in’, ‘background’)}. This information is essential for evaluating the model’s ability to reason and represent spatial relationships accurately in the caption.

The structured results for the candidate and reference captions are as follows. This hierarchical data structure aids in evaluating the model’s ability to recognize objects, their attributes, and the relationships between them in the scene.

Candidate objects: ‘building’, ‘city street’, ‘truck’, ‘tree’, ‘car’
attributes: ‘car’: ‘red’, ‘tree’: ‘green’, ‘truck’: ‘white’, ‘building’: ‘tall’,
relations: (‘truck’, ‘drive down’, ‘city street’), (‘car’, ‘drive down’, ‘city street’), (‘building’, ‘in’, ‘background’)

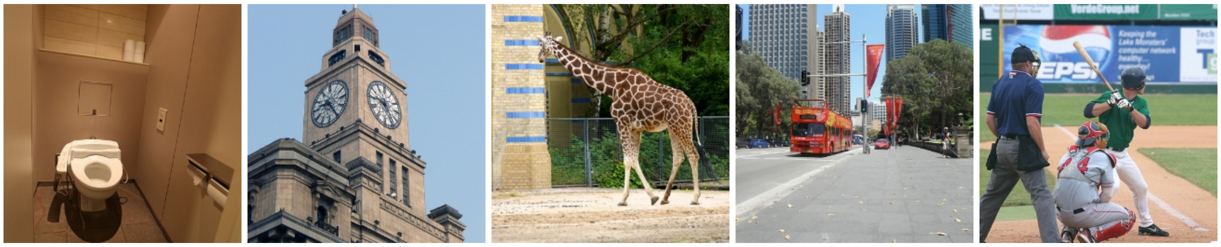
Reference objects: ‘sky’, ‘edifice’, ‘car’, ‘street’, ‘truck’, ‘city street’, ‘tree’
attributes: ‘car’: ‘red’, ‘city street’: ‘busy’, ‘truck’: ‘white’, ‘tree’: ‘green’, ‘edifice’: ‘modern’, ‘sky’: ‘blue’, ‘clear’
relations: (‘tree’, ‘surround’, ‘street’), (‘edifice’, ‘stand under’, ‘sky’), (‘car’, ‘run in front of’, ‘city street’)

Once the elements are extracted from both the ground truth and candidate captions, we can compute the scores for objects, attributes, and relationships using the following metrics: (a) **Precision**, which measures the accuracy of the model on all mentioned samples in the candidate; (b) **Recall**, which measures the accuracy on all actual samples in the reference; (c) **F1 score**, which combines precision and recall by representing their harmonic mean. Similar to the long detailed caption evaluation [9, 26], the scores for objects, attributes, and relationships are calculated as follows:

$$\text{Precision} = \frac{N(\text{Matched})}{N(\text{Candidate})}, \quad \text{Recall} = \frac{N(\text{Matched})}{N(\text{Reference})}, \quad \text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5}$$

where $N(\text{Matched})$ is the number of correctly matched items, $N(\text{Candidate})$ the total items in the candidate, and $N(\text{Reference})$ the total in the reference.

Given that objects, attributes, and relationships extracted from ground truth and candidate captions are often not the same, we process the data through three matching steps: (a) exact matching: this step checks for precise word matches between the ground truth and candidate captions, (b) synonym matching: this step matches words based on their similar meanings, (c) semantic matching: for any remaining unmatched elements, the cosine similarity of their word embeddings is computed to determine their relevance. Using the same example, ‘city street’, ‘truck’, ‘tree’, and ‘car’ pass exact matching, while ‘building’ passes synonym matching as it shares a similar meaning with ‘edifice’. This gives 5 correct matches and 2 missing matches. Therefore, the precision is 100%, which is calculated as 5/5, recall is 71.43% (5/7) and F1 is 83.33%.



(a) toilet (b) tower (c) giraffe (d) street (e) player

Figure 2. **Example NSD Images.** Below, we present detailed captions generated by MLLMs using images as input, alongside captions from our method using different feature spaces with brain signals as input.

3.2. Detailed Caption Examples

We provide COCO [24] captions and detailed captions for example images (as shown in Fig. 2) from the NSD test set. Results from LLaVA-v1.5 7B (LLaVA) [25] serve as pseudo ground truth for detailed captioning evaluation, while others, including LLaVA-MoF (MoF) [42], DenseConnector-v1.5 7B (DC) [48], and Matryoshka-MM-v1.5 7B (M3) [5] are for reference.

MLLM	Caption
COCO	(a) (14 words) A very small bathroom stall with a toilet and several rolls of toilet paper.
	(b) (10 words) A very tall clock tower with two clocks on it.
	(c) (10 words) A large giraffe walking across a dirt lot near trees.
	(d) (12 words) A red and yellow double decker bus on street next to trees.
	(e) (10 words) A baseball player holding a bat next to home plate.
LLaVA	(a) (44 words) The image shows a small bathroom with a white toilet sitting on a tiled floor. The toilet is positioned in the corner of the room, and there is a toilet paper roll on the wall nearby. The bathroom appears to be clean and well-maintained.
	(b) (59 words) The image features a tall clock tower with two clocks on its side. The clocks are positioned at different heights, one higher than the other. The tower is made of stone and has a prominent presence in the scene. The clocks are visible from different angles, making it easy to see the time from various points around the tower.
	(c) (60 words) The image features a giraffe walking in a dirt area, possibly a zoo enclosure. The giraffe is the main focus of the scene, with its long neck and legs prominently visible. The enclosure is surrounded by a fence, and there is a building nearby. The giraffe appears to be walking towards the building, possibly seeking shelter or exploring its surroundings.
	(d) (114 words) The image features a red double-decker bus driving down a city street, surrounded by tall buildings. The bus is positioned in the middle of the scene, and there are several cars on the road as well. Some cars are behind the bus, while others are in front of it. There are multiple traffic lights along the street, with one near the left side of the bus, another further down the road, and a third one on the right side. A few pedestrians can be seen walking along the sidewalk, with one person closer to the bus and two others further away. Additionally, there is a handbag visible near the right side of the scene.
	(e) (114 words) The image captures a baseball game in progress, with a batter holding a baseball bat and preparing to swing. The batter is wearing a green shirt and is positioned near the center of the scene. The catcher, wearing a red shirt, is crouching behind the batter, ready to catch the ball. There are several other people in the scene, including a person standing to the left of the batter and another person on the right side. A baseball glove can be seen in the hands of one of the players, and a sports ball is visible in the air. The scene is set on a baseball field, with a bench located in the background.
MoF	(a) (67 words) The image features a small, clean bathroom with a white toilet situated in the corner. The toilet is equipped with a control panel and a remote, which can be used to operate the toilet. There are also several rolls of toilet paper placed on the wall above the toilet, ensuring that users have easy access to them. The bathroom appears to be well-maintained and ready for use.
	(b) (58 words) The image features a tall clock tower with two clocks on its sides. The clocks are positioned at different heights, with one being higher up and the other lower down. The tower is made of stone and has a brown color. The clocks are visible from various angles, making it easy to tell the time from different directions.
	(c) (60 words) The image features a giraffe standing in a dirt area, possibly a zoo enclosure, with a blue and white striped wall in the background. The giraffe is walking towards the camera, and its long neck and legs are prominently visible. The giraffe appears to be the main focus of the scene, with no other animals or people in the image.

(d) (102 words) The image features a large red double-decker bus driving down a city street, surrounded by tall buildings. The bus is positioned in the middle of the scene, with its height extending from the lower to the upper part of the image.
There are several people walking along the sidewalk, with some closer to the bus and others further away. A few cars can be seen on the street, with one car located behind the bus and another one in front of it. A traffic light is visible on the left side of the image, indicating the flow of traffic in the city.

(e) (78 words) The image captures a baseball game in progress, with a batter standing at home plate, holding a baseball bat and preparing to swing. The catcher and umpire are also present, with the catcher wearing a baseball glove.
There are several other people in the scene, including teammates and possibly coaches, scattered around the field. A baseball bat can be seen lying on the ground, and a sports ball, likely the baseball, is visible in the scene as well.

DC (a) (64 words) The image shows a small bathroom with a white toilet positioned in the middle of the room. The toilet is equipped with a handicap rail, making it accessible for people with disabilities. There is a toilet paper roll on the wall next to the toilet, and a towel is placed on the back of the toilet. The bathroom appears to be clean and well-maintained.

(b) (77 words) The image features a tall clock tower with two clocks on its side, one on the left and the other on the right. The clocks are positioned at different heights, with the one on the left being higher than the one on the right. The tower is part of a large building, and the clocks are visible from various angles. The sky in the background is clear and blue, providing a beautiful contrast to the clock tower.

(c) (52 words) The image features a giraffe walking in a dirt area near a building. The giraffe is the main focus of the scene, with its long neck and legs visible as it moves across the ground. The building appears to be a part of the giraffe's enclosure, providing a backdrop for the scene.

(d) (106 words) The image depicts a city street with a red double-decker bus driving down the road. The bus is surrounded by tall buildings, giving the impression of a busy urban environment. There are several cars on the street, including one behind the bus and others parked or driving in the same direction.
There are multiple traffic lights visible in the scene, with one near the left side of the street and others scattered throughout the area. A few pedestrians can be seen walking along the sidewalk, with one person carrying a handbag. The overall atmosphere of the image is bustling and lively, typical of a city street.

(e) (102 words) The image captures a baseball game in progress, with a batter holding a baseball bat and preparing to swing. The catcher and the umpire are also present, closely observing the batter's actions. The batter is wearing a green shirt and is positioned near the center of the scene. The catcher is located to the left of the batter, while the umpire stands to the left of the catcher.
In the background, there are two more people, possibly teammates or spectators, standing further away from the action. The scene is set on a baseball field, with a baseball glove visible in the foreground.

M3 (a) (34 words) The image shows a small bathroom with a white toilet. The toilet is situated in the corner of the room, and there is a sink nearby. The bathroom appears to be clean and well-maintained.

(b) (53 words) The image features a large clock tower with two clocks on its side. The clocks are visible on different sides of the tower, making it a prominent and functional structure. The tower is part of a building, and the clocks are placed at various heights, adding to the tower's overall height and prominence.

(c) (44 words) A giraffe is walking in a zoo enclosure, surrounded by a fence. It is standing near a building and appears to be looking at the camera. The giraffe is the main focus of the image, and it occupies a significant portion of the scene.

(d) (82 words) A city street with a red double-decker bus driving down the road. The bus is passing by a traffic light and a pedestrian crossing. There are several cars on the street, including one behind the bus and another one further down the road. A few people can be seen walking or standing on the sidewalk, with one person carrying a handbag. The scene captures the hustle and bustle of city life, with various modes of transportation and pedestrians going about their day.

(e) (96 words) In the image, a baseball player is at home plate, holding a baseball bat and preparing to swing. The catcher is positioned behind the batter, wearing a baseball glove, ready to catch the ball. The umpire is also present, closely observing the game.
There are several other people in the scene, some of whom are standing near the edge of the field, possibly waiting for their turn to play or watching the game. A bench can be seen in the background, likely for players to sit on when they are not actively participating in the game.

3.3. Salient QA Examples

This section presents constructed Salient QA examples using images in Fig. 2 as references.

Reference	**Salient QA**
Fig. 2(a)	Q: Which description best fits the 'bathroom' in the image? A. The bathroom is narrow. B. The bathroom is huge. C. The bathroom is dirty. A: "A".

Fig. 2(b) Q: How is the weather? A. Cloudy. B. Sunny. C. Rainy. A: “B”.

Fig. 2(c) Q: Where is the animal? A: On the street. B. In a zoo. C. In the forest. A: “B”.

Fig. 2(d) Q: Which description best fits the ‘bus’ in the image? A. The bus is blue. B. The bus is double decker. C. The bus is green. A: “B”.

Fig. 2(e) Q: Which description best fits the ‘batter’ in the image? A. He is wearing a black shirt. B. He is wearing shorts. C. He is wearing a hat. A: “C”.

4. Additional Experiments

This section provides more results on concise captioning comparison (Sec. 4.1), detailed captioning comparison (Sec. 4.2), visual reconstruction (Sec. 4.3), denoiser training visualization (Sec. 4.4), and UMAP visualization (Sec. 4.5).

4.1. Concise Captioning Comparison

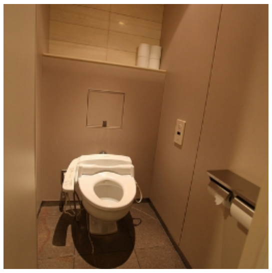
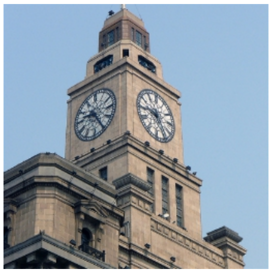
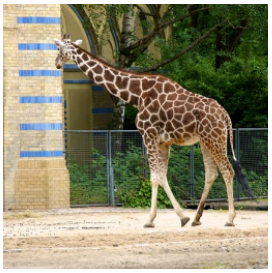
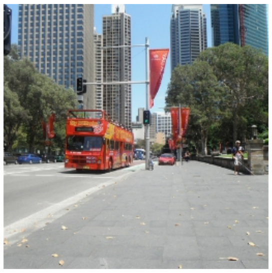
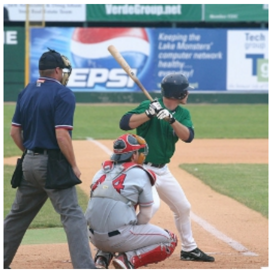
Tab. 5 provides the COCO captions alongside the concise captions predicted by our method, VINDEX (built upon Shikra [6]), as well as state-of-the-art baselines: SDRecon [41], BrainCap [12], OneLLM [13], MindEye2 [35], and UMBRAE [45].

4.2. Detailed Captioning Comparison

Below is a comparison of detailed captioning results using the four feature spaces—SE, ME, AF, and NF. Example images from the NSD test set are shown in Fig. 2. AF typically generates nonsensical outputs, such as garbled text and blank spaces, and therefore the results are not presented here.

Feature	Detailed Captioning
SE	(a) (57 words) The image features a bathroom with a white toilet situated in the middle of the room. The toilet is surrounded by a tiled floor, which gives the bathroom a clean and modern appearance. The walls of the bathroom are also tiled, adding to the overall cohesive design. The bathroom appears to be well-maintained and ready for use.
	(b) (79 words) The image features a tall, old building with a clock tower on top. The clock is prominently displayed on the side of the building, making it a focal point. The building has a steeple-like structure, giving it a unique and historic appearance. The clock is positioned towards the center of the building, drawing attention to its presence. The overall architecture of the building suggests that it may be a church or a similar type of structure with historical significance.
	(c) (75 words) The image features a large giraffe standing in a grassy field. The giraffe is the main focus of the scene, and it appears to be walking through the field. The field is lush and green, providing a natural habitat for the giraffe. The giraffe’s long neck and legs are clearly visible, showcasing its unique features. The overall atmosphere of the image is serene and peaceful, capturing the beauty of the giraffe in its natural environment.
	(d) (115 words) The image depicts a busy city street scene with several cars and buses driving down the road. There are multiple cars in various positions, some closer to the foreground and others further back. A bus is also visible in the middle of the scene, adding to the traffic. In addition to the vehicles, there are several people walking along the sidewalk, going about their daily activities. Some of them are closer to the foreground, while others are further back in the scene. The street is lined with trees, providing a touch of greenery to the urban environment. The combination of the bustling traffic and the presence of pedestrians creates a lively atmosphere in the city.
	(e) (56 words) The image depicts a man wearing a baseball uniform, standing on a field with a baseball glove on his hand. He appears to be a baseball player, possibly waiting for a pitch or preparing to catch a ball. The scene takes place on a baseball field, with the man being the main focus of the image.
ME	(a) (40 words) The image features a white bathroom with a toilet and a sink. The toilet is located on the left side of the bathroom, while the sink is situated on the right side. The bathroom appears to be clean and well-maintained.
	(b) (86 words) The image features a large building with a clock tower, which is situated in the middle of a city. The clock tower is visible on the left side of the building, and the building itself is quite tall. The scene is set against a backdrop of trees, creating a picturesque view. The trees are scattered throughout the scene, with some located near the building and others further away. The combination of the clock tower, the building, and the trees creates a visually appealing and urban landscape.
	(c) (52 words) The image features a herd of zebras grazing in a grassy field. There are at least 13 zebras visible in the scene, scattered throughout the field. Some zebras are closer to the foreground, while others are further in the background. The zebras are peacefully eating grass, creating a serene and natural atmosphere.
	(d) (54 words) The image shows a city street with several cars parked along the side of the road. The cars are of various sizes and are parked in a row. The street appears to be empty, with no people visible in the scene. The cars are parked in a line, creating an organized and orderly appearance.

Table 5. **Concise Captioning Comparison.** Each image is shown with concise captions from SDRecon [41], BrainCap [12], OneLLM [13], MindEye2 [35], MEVOX [47], UMBRAE [45], and VINDEX (built upon Shikra [6]). Refer to Sec. 3.2 for image captions from COCO.

Image	Caption
	SDRecon: a small room in the white bathroom room fitted bathroom, hall room fitted modern bathroom BrainCap: a bathroom with a sink and a toilet OneLLM: A bathroom with a white toilet and a white sink. MindEye2: a bathroom with a toilet and a sink. MEVOX: A bathroom with a toilet and a sink. UMBRAE-S: A bathroom with a toilet, sink and mirror. UMBRAE: A bathroom with a toilet, sink and shower. VINDEX-S: A bathroom with a toilet, a shower and a tub. VINDEX: A white bathroom with a toilet and a shower.
	SDRecon: a bathroom BrainCap: a clock on the side of a tower. OneLLM: A large clock tower sitting on top of a building. MindEye2: a clock tower with a tower in the background. MEVOX: A tall building with a clock on the top. UMBRAE-S: An old building with a clock on the top. UMBRAE: A clock tower that has two clocks sits in the sky. VINDEX-S: A clock tower on a building with a steeple on top. VINDEX: A large building with a clock on the top.
	SDRecon: a wild park in the woods with two cars parked BrainCap: a group of trees and a zebra OneLLM: a fire truck parked in front of a building. MindEye2: a giraffe standing in a field. MEVOX: A zebra standing in the middle of a lush green field. UMBRAE-S: A giraffe is standing in a grassy field. UMBRAE: A giraffe is standing in a grassy field. VINDEX-S: Two giraffes standing next to each other in the grass. VINDEX: A giraffe standing in a grassy field next to trees.
	SDRecon: a street in downtown, the building is located at the center of this street BrainCap: a city street with a lot of cars and a bus. OneLLM: A long red train car sits on train tracks. MindEye2: a street with a car and a bus. MEVOX: A bus driving down a street next to a tall building. UMBRAE-S: A street with cars parked on the side of it. UMBRAE: A city street filled with lots of traffic. VINDEX-S: A street view of cars driving down the street. VINDEX: A city street filled with lots of traffic.
	SDRecon: two men on the field, one is wearing a baseball bat and the other player BrainCap: a young man holding a baseball bat. OneLLM: A group of young men standing on a tennis court. MindEye2: a baseball player is standing in front of a batter. MEVOX: A man holding a tennis racket on a tennis court. UMBRAE-S: A baseball player getting ready to bat. UMBRAE: A group of men playing baseball on a field. VINDEX-S: A baseball player is getting ready to hit the ball. VINDEX: A baseball player getting ready to swing at the ball.

(e) (69 words) The image features a tennis court with a tennis player in action. The player is holding a tennis racket and is in the middle of a swing, likely returning a volley. The tennis player is positioned towards the left side of the court. The court is surrounded by a fence, and there are several people in the background, possibly watching the game or waiting for their turn to play.

NF (a) (97 words) The image features a bathroom with a white toilet situated in the corner of the room. The toilet is positioned under a window, which allows natural light to enter the space. The bathroom also has a sink, which is located towards the right side of the room.

There are two bottles in the bathroom, one placed near the sink and the other closer to the toilet. Additionally, there is a cup on the left side of the room, and a bowl can be seen near the sink. The overall appearance of the bathroom is clean and well-organized.

(b) (79 words) The image features a large, old-fashioned clock tower with a steeple, situated on top of a building. The clock is prominently displayed on the tower, making it a focal point of the scene. The tower is surrounded by a group of trees, creating a picturesque setting.

There are several people in the scene, with some standing closer to the clock tower and others further away. They appear to be enjoying the view of the tower and the surrounding environment.

(c) (46 words) The image features a large giraffe standing in a grassy field, surrounded by trees. The giraffe appears to be walking through the grass, possibly in search of food. The scene is set in a natural environment, with the giraffe being the main focus of the image.

(d) (104 words) The image depicts a city street with a white car parked on the side of the road. The car is positioned near the center of the scene, and it appears to be a compact vehicle. There are several other cars parked along the street, with some closer to the foreground and others further in the background.

In addition to the cars, there are two people visible in the scene. One person is standing near the left side of the image, while the other person is located closer to the center. The street is lined with trees, providing a pleasant atmosphere for the city setting.

(e) (101 words) The image features a baseball field with several baseball players standing on the field. There are at least nine people visible in the scene, with some of them closer to the foreground and others further in the background. A baseball glove can be seen on the ground, indicating that the players are either preparing for a game or have just finished one.

The players are spread out across the field, with some standing closer to the center and others near the edges. The overall atmosphere of the scene suggests that the players are engaged in a casual or recreational baseball game.

4.3. Visual Reconstruction

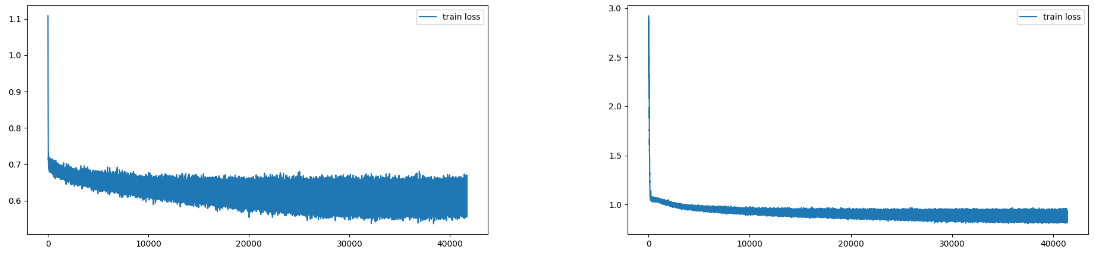
This paper explores fMRI-based multimodal interaction using MLLMs, focusing on perception over reconstruction. We address four tasks: Concise Captioning, Descriptive Captioning, and Concept Localization, and Complex Reasoning. Despite this focus, the generated multimodal explanations [36, 45] are shown to improve reconstruction performance in Tab. 6.

Table 6. **Quantitative Visual Reconstruction Evaluation.** We report metrics following prior studies [30, 34].

Method	#Models	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Mind-Reader [23]	4	-	-	-	-	78.2%	-	-	-
SDRecon [41]	4	-	-	83.0%	83.0%	76.0%	77.0%	-	-
Brain-Diffuser [30]	4	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423
MindEye [34]	4	**.309**	.323	94.7%	**97.8%**	93.8%	94.1%	.645	.367
DREAM [46]	4	.288	.338	**95.0%**	97.5%	94.8%	95.2%	.638	.413
MindBridge [43]	1	.151	.263	87.7%	95.5%	92.4%	94.7%	.712	.418
UMBRAE [45]	1	.283	.328	93.9%	96.7%	93.4%	94.1%	.700	.393
NeuroVLA [36]	1	.265	**.357**	93.1%	97.1%	**96.8%**	**97.5%**	**.633**	**.321**
VINDEX	1	.203	.317	93.5%	96.9%	93.5%	95.1%	.658	.403

4.4. Denoiser as a Training Stabilizer

We conduct an ablation study on the denoiser architecture and weights in the main paper, working together with regression loss to improve performance. Here, we provide further analysis on how the denoiser stabilizes training. As shown in Fig. 3, the vanilla regression loss decreases but exhibits significant oscillation. Incorporating diffusion loss stabilizes the training process and leads to faster convergence.



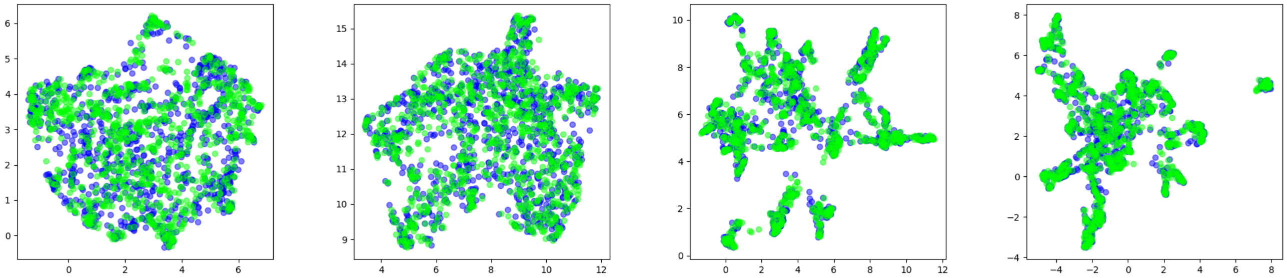
(a) (b)

Figure 3. **Denoiser as a Training Stabilizer.** (a) The vanilla regression loss decreases but exhibits significant oscillation; (b) The training process becomes less oscillatory with the incorporation of diffusion loss, which stabilizes the training and accelerates convergence.

4.5. UMAP Visualization

To better understand the brain-feature alignment, we apply UMAP [27] for dimensionality reduction, projecting the predicted features from brain encoders and target vision encoders into a 2D space. The well-aligned features should form cohesive clusters, while misaligned features are expected to be disjointed [34].

Fig. 4 shows the UMAP visualization for predicted features from brain encoders B-CLIP224, B-CLIP336, B-DINO224, and B-SIGLIP384 (fMRI as input), along with the corresponding ground truth features from target vision encoders (associated visual stimuli as input). Refer to Tab. 2 for more details on pre-trained brain encoders.



(a) B-CLIP224 (b) B-CLIP336 (c) B-DINO224 (d)B-SIGLIP384

Figure 4. **UMAP Comparison.** The UMAP visualizations depict predicted features (blue) from brain encoders B-CLIP224, B-CLIP336, B-DINO224, and B-SIGLIP384 using brain inputs, along with the corresponding image features (green) extracted from target vision encoders using associated images as input. For further details on the brain encoders, please refer to Tab. 2.

5. Discussion

Our method explores fMRI-based multimodal interaction using MLLMs but has certain limitations. First, we primarily focus on representative feature spaces to validate the essence of our idea rather than exhaustively experimenting with all existing vision encoders. Second, we rely solely on fMRI data [1] without extending our analysis to other neuroimaging techniques such as EEG or MEG. EEG, being more portable and cost-effective, enables large-scale data collection, albeit with lower spatial resolution. Incorporating EEG experiments could further explore alignment with image features while benefiting from larger data scales. Third, we align brain signals with image features, serving as a simple approach for zero-shot learning. However, as shown in Tab. 2 and Tab. 7, token numbers and feature dimensions increase when using vision encoders with higher image resolutions, imposing significant computational costs during training. Furthermore, results in the main paper indicate that aligning brain signals with higher-resolution visual features does not necessarily yield better understanding or reconstruction. Instead, aligning with tokens directly offers a more direct manner, enabling the use of token pruning and merging techniques [4, 18] to dynamically reduce token count based on the task, thereby alleviating the computational burden.

Table 7. **Detailed on pre-trained Vision Experts.** These vision models are trained and specialized for specific tasks, and it has been shown that MLLMs using these task-specific vision encoders achieve optimal performance within their pre-training domains.

Vision Expert	Task	Size	#Dimension	Link
CLIP [31]	Image-text Contrastive	448	1024	openai/clip-vit-large-patch14-336
DINOv2 [29]	Visual Grounding	448	1024	facebookresearch/dinov2
SAM [19]	Image Segmentation	1024	1024	facebook/sam-vit-large
EVA-02 [11]	Object Detection	1024	1024	EVA-02-Large
ConvNeXt [44]	Image Classification	1024	1024	laion/CLIP-convnext-xxlarge

Table 8. **Details on Hybrid MLLMs with Mixtured Vision Encoders**.

Method	**Encoders**	**MLLM**
EAGLE [38]	CLIP, ConvNeXt, EVA-02, Pix2Struct, DINOv2, SAM	LLaVA-1.5/Qwen2.5
MERV [8]	DINOv2, ViViT, SigLIP, LanguageBind	LLaMA2/3
MoVA [52]	CLIP, DINOv2, CoDETR, SAM, Pix2Struct, *etc.*	Vicuna/Llama3/Yi
MoME [37]	CLIP, DINOv2, Pix2Struct	Vicuna-v1.5
LEO [2]	InternViT, SAM	InternVL

Visual Encoders Beyond the vision experts used in the main paper, there are several other pre-trained visual encoders that can be incorporated into the toolbox. These encoders, trained on various tasks and resolutions, allow us to explore the distinct advantages of different experts. We compile a set of vision experts, including: (1) Constrastive Vision-Language Alignment: CLIP [31] and ConvNeXt [44] from OpenCLIP [33]. (2) Visual Grounding: DINOv2 [29] using self-supervised learning. (3) Object-Centric Training: EVA-02 [11] and CoDETR [51], pre-trained on detection datasets. (4) Optical Character Recognition (OCR): Pix2Struct [20]. (5) Segmentation: SAM [19]. (6) Video-Language Pretraining: LanguageBind [50]. (7) Vision Foundation Model: InternViT [7]. The detailed task (taxonomy), input image size (resolution), and checkpoint for each vision encoder are in Tab. 7. Preliminary results from recent hybrid multimodal models [2, 8, 38, 52] indicate that MLLMs using these task-specific vision encoders achieve optimal performance within their pre-training domains. For example, EVA-02 [11] excels in the the visual question answering benchmark GQA [17] and the object hallucination evaluation benchmark POPE [22]. Both CLIP [31] and ConvNeXt [44] perform well across several benchmarks, benefiting from large-scale image-text pair training using contrastive loss. In contrast, while Pix2Struct excels at text recognition, it shows limited capability in object recognition and general VQA tasks. DINOv2 [29] and SAM [19], trained via self-supervised learning and semantic segmentation, respectively, struggle with text recognition tasks.

Hybrid Multimodal Models. Besides the feature spaces discussed in the main paper, there are also other hybrid MLLMs, especially those based on a mixture of vision experts. MLLMs that utilize these task-specific vision encoders deliver optimal performance within their respective pre-training domains. We provide a reference list in Tab. 8 for these hybrid MLLMs.

From the brain results presented in the paper, we observe that aligning with multiple vision encoders does not only bring performance gains but also increases training complexity. Brain signals struggle with the precise location of concepts, especially for small, inconspicuous objects [45]. It is predictable that aligning with SAM features [19] may not effectively support decoding brain signals into clear segmentation results. These findings from both research directions inspire us to further explore aligning brain signals with foundation vision models such as ConvNeXt [44] or InternViT [7], which support universal perception, rather than aligning with multiple vision experts specialized in individual vision tasks.

References

[1] 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. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. 1, 3, 11

[2] Mozhgan Nasr Azadani, James Riddell, Sean Sedwards, and Krzysztof Czarnecki. Leo: Boosting mixture of vision encoders for multimodal large language models. *arXiv preprint arXiv:2501.06986*, 2025. 12

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[4] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *ICLR*, 2023. 11

[5] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *ICLR*, 2025. 1, 2, 3, 4, 6

[6] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3, 4, 8, 9

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 12

[8] Jihoon Chung, Tyler Zhu, Max Gonzalez Saez-Diez, Juan Carlos Niebles, Honglu Zhou, and Olga Russakovsky. Unifying specialized visual encoders for video language models. *arXiv preprint arXiv:2501.01426*, 2025. 1, 12

[9] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024. 5

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[11] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 2024. 12

[12] Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023. 8, 9

[13] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *CVPR*, 2024. 8, 9

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Workshop on Deep Learning*, 2014. 1

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 3

[17] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 12

[18] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *WACV*, pages 1383–1392, 2024. 11

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, 2023. 12

[20] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *ICML*, 2023. 12

[21] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NeurIPS*, 37:56424–56445, 2024. 3

[22] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*, 2023. 12

[23] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind Reader: Reconstructing complex images from brain activities. *NeurIPS*, 35: 29624–29636, 2022. 10

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 1, 6

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 3, 4, 6

[26] Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. In *CVPR*, 2025. 5

[27] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 11

[28] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171, 2021. 2, 3

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2023. 1, 2, 3, 12

[30] Furkan Ozcelik and Rufin VanRullen. Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. 10

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 12

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks Track*, 2022. 12

[34] Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In *NeurIPS*, 2023. 1, 3, 10, 11

[35] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, and Tanishq Mathew Abraham. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *ICML*, 2024. 1, 8, 9

[36] Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. In *NeurIPS*, pages 98083–98110, 2024. 10

[37] Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. In *NeurIPS*, pages 42048–42070, 2025. 1, 12

[38] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *ICLR*, 2025. 12

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[40] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 2, 3

[41] Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. *arXiv preprint arXiv:2306.11536*, 2023. 1, 8, 9, 10

[42] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pages 9568–9578, 2024. 1, 2, 3, 6

[43] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11333–11342, 2024. 10

[44] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, 2023. 12

[45] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *ECCV*, pages 242–259, 2024. 1, 2, 3, 8, 9, 10, 12

[46] Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *WACV*, pages 8226–8235, 2024. 1, 10

[47] Weihao Xia and Cengiz Öztireli. Mevox: Multi-task vision experts for brain captioning. In *CVPR Workshop*, 2025. 9

[48] Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for mllms. In *NeurIPS*, pages 33108–33140, 2025. 1, 2, 3, 4, 6

[49] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 1, 2, 3

[50] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2024. 12

[51] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *ICCV*, pages 6748–6758, 2023. 12

[52] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. In *NeurIPS*, 2024. 1, 12