Neuro-Vision to Language: Enhancing Brain Recording-based Visual Reconstruction and Language Interaction

Guobin Shen[1,2,3,4], **Dongcheng Zhao**[1,2,3], **Xiang He**[1,3,5], **Linghao Feng**[1,3], **Yiting Dong**[1,2,3,4], **Jihang Wang**[1,5], **Qian Zhang** [1,2,3,5], **and Yi Zeng**[1,2,3,4,5†]
[1] Brain-inspired Cognitive Intelligence Lab, Institute of Automation, Chinese Academy of Sciences,
[2] Center for Long-term Artificial Intelligence,
[3] Beijing Key Laboratory of Artificial Intelligence Safety and Superalignment,
[4] School of Future Technology, University of Chinese Academy of Sciences,
[5] School of Artificial Intelligence, University of Chinese Academy of Sciences
{shenguobin2021, zhaodongcheng2016, hexiang2021, fenglinghao2022, dongyiting2020, wangjihang2021, q.zhang, yi.zeng}@ia.ac.cn

Abstract

Decoding non-invasive brain recordings is pivotal for advancing our understanding of human cognition but faces challenges due to individual differences and complex neural signal representations. Traditional methods often require customized models and extensive trials, lacking interpretability in visual reconstruction tasks. Our framework integrates 3D brain structures with visual semantics using a *Vision Transformer 3D*. This unified feature extractor efficiently aligns fMRI features with multiple levels of visual embeddings, eliminating the need for subject-specific models and allowing extraction from single-trial data. The extractor consolidates multi-level visual features into one network, simplifying integration with Large Language Models (LLMs). Additionally, we have enhanced the fMRI dataset with diverse fMRI-image-related textual data to support multimodal large model development. Integrating with LLMs enhances decoding capabilities, enabling tasks such as brain captioning, complex reasoning, concept localization, and visual reconstruction. Our approach demonstrates superior performance across these tasks, precisely identifying language-based concepts within brain signals, enhancing interpretability, and providing deeper insights into neural processes. These advances significantly broaden the applicability of non-invasive brain decoding in neuroscience and human-computer interaction, setting the stage for advanced brain-computer interfaces and cognitive models.

1 Introduction

The decoding of non-invasive brain recordings, such as those obtained from fMRI [1], is a cornerstone of cognitive neuroscience [2; 3; 4]. This process offers unparalleled insights into the neural underpinnings of human cognition, contributing not only to fundamental scientific knowledge but also to advancements in clinical and technological applications. Despite its potential, the field faces significant challenges primarily due to the high variability of brain activity across individuals [5] and the complexity inherent in the neural representations of cognitive processes [6].

Brain decoding techniques have traditionally relied on customized, subject-specific models [7; 6; 8]. These models necessitate intricate and costly experimental setups, depending on multiple trials

†Corresponding Author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3113

to achieve reliable results. Such approaches, while helpful, are inherently limited in scalability and flexibility, hindering broader application and generalization across different populations and conditions.

Visual reconstruction [9] aims to recreate perceived visual stimuli from brain signals and is considered one of the benchmarks of brain decoding. However, this approach often struggles to accurately reproduce the visual experiences of individuals, generally lacking semantic precision and interpretability [8]. This inability to effectively decode and reconstruct signals restricts our understanding of how sensory information is processed. Recognizing these limitations, our study introduces language modalities as a critical enhancement designed to assess decoding performance more effectively and enrich brain-computer interfaces' interaction capabilities.

Addressing these multifaceted challenges, our research introduces the *Vision Transformer 3D* (ViT3D) [10] specifically tailored to the domain of visual reconstruction. Unlike traditional approaches that often reduce complex brain regions to one-dimensional vectors [11; 6; 12; 13], losing critical spatial structure information, our implementation of ViT3D preserves the three-dimensional structural integrity of the brain data. This adaptation enables an unprecedented enhancement in the extraction of visual semantic information, ensuring a deeper fidelity and richness in the decoded visual representations.

Our fMRI feature extractor includes a unified network backbone and two alignment heads for feature matching. This setup enables efficient, high-quality visual reconstructions across subjects from one experimental trial. By simply aligning the extractor's output with CLIP embeddings [14] and features of Variational Autoencoder (VAE) [15], our method eliminates the need for multiple, subject-specific models, substantially simplifying the decoding process. This straightforward and effective configuration reduces the resources required for brain decoding and showcases the potential for easy integration with Large Language Models (LLMs), enhancing its usability across various applications.

Moreover, our research delves into the integration of brain recordings with visual and linguistic data within a comprehensive multimodal framework using LLMs. This integration significantly improves visual reconstruction performance and introduces the groundbreaking capability for direct interaction through natural language. Our model facilitates diverse communication with brain data using natural language and precisely localizes linguistic concepts within brain signals. These advancements help deepen our understanding of the interactions between language, perception, and neural activity. Additionally, to bolster the development of these multimodal models, we have augmented the brain-recording visual dataset with natural language enhancements.

In summary, our contributions can be summarized as follows:

- Our fMRI feature extractor, based on *Vision Transformer 3D*, aligns fMRI features with visual embeddings at multiple levels, integrating 3D brain structures with visual semantics. This eliminates the need for subject-specific models and enables efficient data extraction from single trials, significantly reducing training costs and enhancing practical usability in real-world scenarios.
- We expanded the language dimension of our fMRI-visual dataset to build a multimodal large model capable of decoding fMRI data. This enhancement boosts brain decoding performance and broadens the application scope to include tasks like visual reconstruction, question-answering, and complex reasoning while also allowing precise localization and manipulation of language-based concepts within brain signals.
- Experimental results on the Natural Scenes Dataset (NSD) [16] for visual reconstruction and language interaction tasks demonstrate that our method surpasses existing models, effectively achieving concept localization and elimination.

2 Related Works

Non-invasive techniques such as functional magnetic resonance imaging (fMRI) are pivotal in providing direct insights into neural activities, significantly deepening our understanding of complex cognitive processes from neural network structures [17] to advanced image and language processing tasks [18; 19]. This section reviews key developments in fMRI-based brain decoding, particularly emphasizing the shift from simple subject-specific analyses to more complex, multimodal interpretations.

Visual Reconstruction from Brain Activity Visual reconstruction from brain activity involves translating brain recordings into the visual stimuli perceived by subjects. Early methods, like those developed by Horikawa *et al.* [18], relied on sparse linear regression to predict features extracted by convolutional neural networks from fMRI data. Recent advancements in generative artificial intelligence, particularly diffusion models [20], have propelled efforts to reconstruct visual stimuli directly from fMRI. For instance, Lin *et al.* [21] aligned fMRI data with image features and corresponding CLIP embeddings to facilitate image generation using fine-tuned StyleGAN [22]. Similarly, Takagi *et al.* [9] improved the quality of visual reconstructions by aligning fMRI with CLIP text embeddings and the latent spaces of diffusion models. Xia *et al.* [11] aligned fMRI data from dimensions of image CLIP features, depth, and color using T2I Adapters [23] for fine-grained conditional control. Despite these advancements, the complexity of such methods, involving multiple independent modules, complicates their integration with technologies like LLMs and restricts their generalizability across different subjects. We observed that some contemporary works also attempt cross-subject alignment; however, these methods either require subject-specific parameters [24] or face performance issues compared to subject-specific models [25].

fMRI Data Processing Efficiently processing fMRI data to extract visually relevant activities typically involves simplifying the data into one-dimensional vectors and selecting voxels most responsive to visual stimuli. Traditional methods utilize simple linear regression or fully connected networks to predict visual stimulus features [18; 21]. However, these methods often lose essential spatial structural information, which is critical given the individual differences in brain anatomy. To address these challenges, innovations such as Vision Transformer 3D (ViT3D) have been developed for managing data with intricate spatial structures [26; 27]. ViT3D segments 3D data into patches, preserving local spatial information within each patch and maintaining overall structural integrity through self-attention mechanism [28], thereby enhancing the performance of brain activity extraction [29].

Multimodal Integration with Brain Signals The utilization of language as a medium for representation allows for the expression of complex concepts with precision and abstraction. The advent of LLMs has showcased their potential to act as bridges across different modalities, enhancing interactions with visual and audio data through natural language [30; 31]. For example, approaches like those by Défossez *et al.* [32], which align brain recordings with spoken language to decode speech non-invasively, have demonstrated the effectiveness of combining brain recordings with LLMs. However, these approaches are often limited by the specificity of the fMRI feature extractors used, which can restrict the scalability and the size of the models employed. By integrating our specially designed cross-subject fMRI feature extractor, we enhance visual reconstruction quality and enable complex reasoning and direct interaction with model outputs using natural language, achieving precise localization of open natural language concepts within the brain.

3 Methodology

Our approach is designed to tackle the key challenges encountered in the visual reconstruction of brain activity and the integration of LLMs with multimodal data. Traditional brain decoding methods, especially those involving fMRI, often struggle with the complexity of accurately reconstructing visual stimuli and generalizing these models across different subjects. Furthermore, while LLMs hold significant potential for enhancing interactions across various cognitive models, their integration with neuroimaging data has been hindered by the need for non-scalable or efficient customized, subject-specific models.

In the following sections, we detail our methodology's components, as seen in Fig. 1. We describe the architecture of our feature preprocessor that maintains the spatial structure of fMRI data, our unified fMRI feature extractor, and the integration strategies for the network with LLMs. We also elaborate on the multimodal interaction techniques that enable direct and meaningful communication between the computational model and the neural representations and the implementation details for visual reconstruction.

3.1 fMRI Feature Preprocessor

fMRI quantifies changes in the blood-oxygen-level-dependent (BOLD) signals to characterize neural activity. The BOLD signal for a given subject can be represented as a 3D matrix $b_{\text{origin}} \in \mathbb{R}^{X_s \times Y_s \times Z_s}$, where s indexes the subject, accounting for inter-individual variability. Traditional preprocessing methods typically involve masking voxels sensitive to the specific task, followed by flattening the

 https://doi.org/10.52202/079017-3113

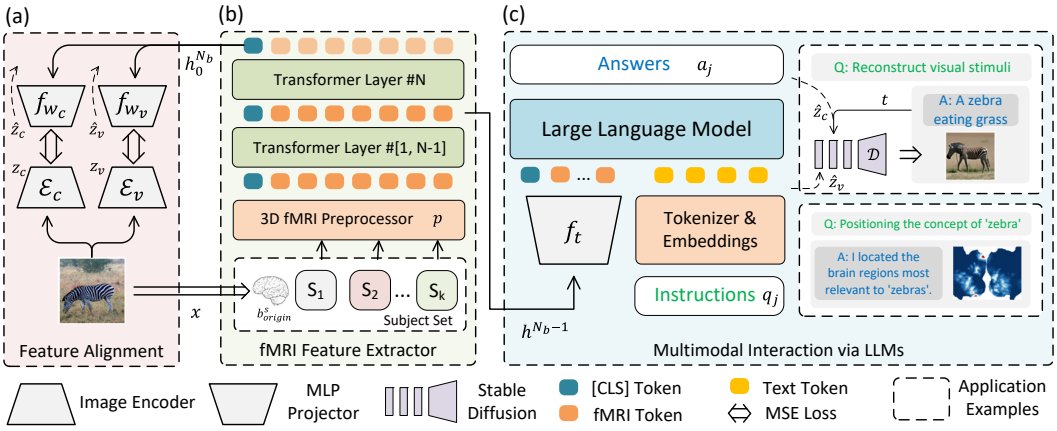

Figure 1: Overview of the integrated multimodal framework combining fMRI feature extraction with LLMs for interactive communication and reconstruction. The architecture comprises: **(a)** a dual-stream pathway for feature alignment with VAE and CLIP embeddings. **(b)** A 3D fMRI preprocessor p, and an fMRI feature extractor. **(c)** A multimodal LLM integrated with fMRI. The extracted features are then fed into an LLM for processing natural language instructions and generating responses or visual reconstructions.

remaining data into a 1D vector. For subject s, the processed fMRI signal can thus be represented as $b_s \in \mathbb{R}^{1\times N_s}$, where N_s denotes the number of voxels selected after masking.

However, this approach results in a loss of spatial structural information, complicating alignment across different subjects. We propose a feature extraction method that preserves spatial structure, as shown in Fig. 2. Starting with the original BOLD signal b_{origin}, we first use trilinear interpolation to resize the data to a uniform dimension, ensuring maximal spatial consistency across subjects' brains while not introducing subject-specific parameters. After resizing, the normalized data undergoes a patching process where it is divided into smaller cubic segments, each of dimension $C = r^3$. This step retains the local spatial features within each segment, preserving the 3D structure crucial for accurate analysis. Finally, patches containing non-task-relevant information are filtered out to reduce computational load. This results in patched data with dimensions $\mathbb{R}^{N\times C}$, where N is the number of patches retained that contain meaningful information. The entire preprocessing operation can be summarized as a mapping:

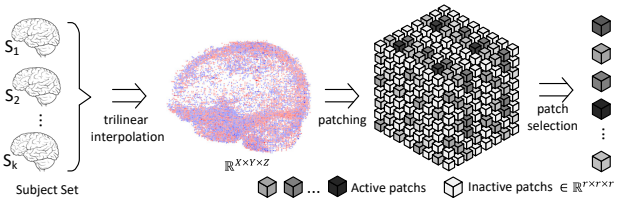

Figure 2: Description of fMRI data preprocessing. First align the data of different subjects, then patch them, and finally remove activity-irrelevant patches.

$$\left\{p : b^s_{\text{origin}} \mapsto b \mid b^s_{\text{origin}} \in \mathbb{R}^{X_s\times Y_s\times Z_s}, b \in \mathbb{R}^{N\times C}\right\}. \tag{1}$$

In Eq. 1, p is the preprocessing function applies resizing, patching, and masking to transform the original fMRI data into a structured format. This function provides a uniform representation of the BOLD signals across different subjects, ensuring that the spatial structure is preserved and capable of integration with Transformer architectures.

3.2 Dual-Stream fMRI Feature Extractor

Visual reconstruction tasks typically begin by mapping the processed BOLD signal, b, to various estimated visual feature spaces, represented as $\{\hat{z}_1, \hat{z}_2, \ldots\}$, where z indicates different levels of visual features and $\hat{z}$ denotes the visual features estimated from the BOLD signals. Subsequently, these features are used to reconstruct the image, represented as $\hat{x}$, from the visual features. To enhance the quality of the reconstructed images, complex feature extractor designs are required,

which increases the preprocessing and computational overhead. However, thanks to the design of our fMRI feature extractor, we can achieve efficient visual reconstruction using a single network backbone, as shown in Fig. 1(b).

Specifically, the patched features obtained from Eq. 1 are directly processed through a Transformer Encoder $\mathcal{E}_b$ to extract features, obtaining the hidden states from the last layer, $h^{N_b} = \mathcal{E}_b(b)$, where N_b represents the number of layers of the encoder. These outputs are then aligned with the visual stimulus's CLIP embeddings $z_c = \mathcal{E}_c(x)$ and VAE features $z_v = \mathcal{E}_v(x)$, where x represents the visual stimulus. The loss function used to train the fMRI feature extractor can be expressed as:

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{(b,x)\sim P(B,X)} \left[\left\| f_{w_c}(h_0^{N_b}) - \mathcal{E}_c(x) \right\|_2^2 + \alpha \left\| f_{w_v}(h_0^{N_b}) - \mathcal{E}_v(x) \right\|_2^2 \right]. \tag{2}$$

In Eq. 2, the expectation $(b, x) \sim P(B, X)$ averages the alignment loss across samples from the $\mathbf{B}$ (fMRI signals) and $\mathbf{X}$ (corresponding visual stimuli). Here, $h_0^{N_b}$ represents the output from the first token of the last hidden state layer of the encoder $\mathcal{E}_b$. The functions f_{w_c} and f_{w_v} are two-layer perceptrons designed to align the extracted fMRI features with these embeddings. The hyperparameter α balances the losses between the alignments of CLIP and VAE features. Through this dual-stream configuration, we create a backbone network that incorporates different levels of visual features. Using only a simple MSE loss function, we achieve high-quality visual reconstruction.

3.3 Multimodal Interaction with fMRI

Our feature extractor architecture, equipped with a single backbone network, is adept at encapsulating various feature levels, making it highly suitable for integration with LLMs. Inspired by advancements such as those in LLaVA [33], we utilize the penultimate hidden states of our network, h^{N_b-1}, as multimodal tokens of fMRI data. A two-layer perceptron f_t projects this state to the same dimension as the text embeddings, resulting in the fMRI embeddings $t = f_t(h^{N_b-1})$. Considering a sequence of question-answer pairs related to the fMRI data $[q_0, a_0, q_1, a_1, \ldots, q_L, a_L]$, the training objective is formulated as:

$$\max p_\theta(\mathbf{A}|\mathbf{Q}, t) = \max \prod_{j=0}^{L} p_\theta(a_j|q_j, a_{j-1}, \ldots, q_0, t). \tag{3}$$

Eq. 3 describes the probability of generating a sequence of answers $\mathbf{A}$ given a sequence of questions $\mathbf{Q}$ and the derived fMRI embeddings t. Each answer a_j is conditionally dependent on all preceding questions and answers, as well as the context embeddings derived from fMRI data, where θ represents the trainable parameters of LLMs.

To effectively couple fMRI data with language models, annotated data is essential. Although the NSD [16] uses labeled visual stimuli from the COCO dataset [34], semantic mismatches often occur due to modifications such as image cropping performed when displaying images to subjects (Appendix A.1). Moreover, the original captions in NSD are not detailed enough to capture nuanced semantic information. Recognizing the importance of comprehensive linguistic annotations, we have constructed a diverse instructional dataset that includes various textual data: brief descriptions, detailed descriptions, continuous dialogues, complex reasoning tasks, instruction reconstruction, and concept localization.

3.4 Interaction and Reconstruction via LLMs

Our fine-tuned model can understand information embedded within fMRI data and adhere to human instructions. Interactions and explanations of visual stimuli content occur through natural language. A typical dialogue format is structured as follows: `<human>:[image] [instruction] <bot>:[answer]`. Here, `[instruction]` denotes a natural language instruction, which during inference is tokenized and embedded, while `[image]` acts as a placeholder, replaced by the fMRI data embedding t. The model then responds based on the directive and the embedded fMRI data. `[answer]` represents the response generated by the LLMs.

After instruction-based fine-tuning, the model communicates directly via natural language and supports visual reconstruction and location identification of concepts expressed in natural language, as shown in Fig. 3. These are facilitated respectively through Stable UnCLIP [20] for visual reconstruction and GradCAM [35] for concept localization.

https://doi.org/10.52202/079017-3113

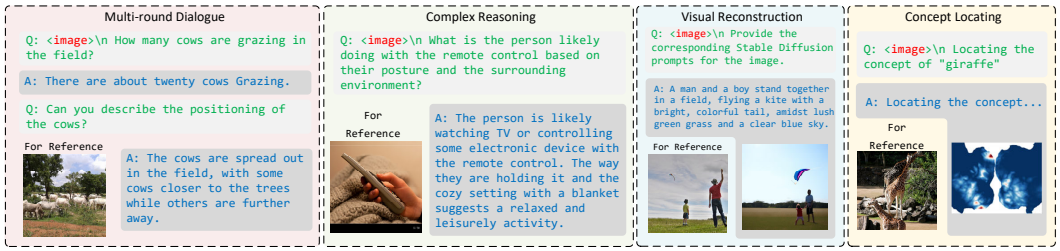

Figure 3: Demonstration of the model's capabilities for engaging in multi-round dialogue, complex reasoning, visual reconstruction, and concept location tasks using fMRI data.

For visual reconstruction, the LLM initially generates a reconstruction prompt a_r, which, combined with the latent representations $\hat{z}_v$ and $\hat{z}_c$ from the fMRI feature extractor, results in the generation of an image. This process can be formalized as:

$$\hat{x} = \mathcal{D}((1-\beta)\hat{z}_v + \beta\sigma \mid \hat{z}_c, a_r), \quad \sigma \in \mathcal{N}(0,1). \tag{4}$$

In Eq. 4, $\mathcal{D}$ represents the frozen UnCLIP, used for visual reconstruction, where $\hat{z}_c$ and a_r serve as conditional information during the denoising process, and $\hat{z}_v$ acts as the initial latent representation of the image. The hyperparameter β is used to introduce noise into the latent space prior, balancing low-level features brought by the prior and high-level information controlled by the diffusion conditions during the denoising process. For concept localization in natural language, LLMs activate the feature extractor using keywords from the instructions, enabling precise location identification of the discussed concepts.

4 Experiments

4.1 Implementation Details

Dataset and Preprocessing: We utilized the Natural Scenes Dataset (NSD) [16], containing high-resolution 7Tesla fMRI scans and corresponding visual stimuli from COCO [34]. The dataset involved eight subjects, but analyses focused on the four (`subj01`, `subj02`, `subj05` and `subj07`) who completed all sessions. Modifications like cropping necessitated re-annotation of images using BLIP2 [36] for captions and DETR [37] for bounding boxes to maintain consistency. fMRI data was standardized to dimensions $83 \times 104 \times 81$ using trilinear interpolation and segmented into $14 \times 14 \times 14$ patches.

Architecture and Training: Our architecture integrates CLIP ViT-L/14 [14] and an AutoencoderKL [15] for image feature extraction, aligned with fMRI data processed through a 16-layer Transformer Encoder [28]. This setup employed two perceptrons (f_{w_c}, f_{w_v}) to align features with CLIP and VAE, respectively. Training involved a multi-stage approach where the LLM was initially frozen, followed by a fine-tuning stage for both the LLM and the Transformer Encoder. For visual reconstructions, the model utilized UnCLIP-2 [20] with β set to 0.93, and concept localization was achieved using GradCAM [35]. For more details on the dataset and experimental setup, please refer to Appendices A, B, and C. Additional experimental results can be found in Appendix D.

4.2 Captioning and Question Answering

Tab. 1 shows the performance of our method on multimodal language tasks. With the introduction of LLMs, we have expanded the task forms to include brain captions, detailed descriptions, and complex reasoning, as illustrated in Fig. 3. Our approach has demonstrated superior performance on the majority of metrics for the brain captioning task. Notably, our model can generalize across subjects without the need to train separate models for each subject or introduce subject-specific parameters.

Given the semantic mismatch between captions and images in the original NSD (Section 3.3), we reran the experiment using BLIP2 [36]-generated captions as ground truth. The results, shown in Tab. 1, show significant improvements when evaluated against BLIP2-generated captions, confirming the effectiveness of our model in the brain captioning task and the reasonableness of the task setting.

Beyond brain captioning, we have incorporated tasks for detailed description and complex reasoning. Our model also achieved the best performance on these two tasks, suggesting that it can generate

Table 1: Quantitative analysis of brain captioning, detailed descriptions, and complex reasoning tasks. Some results are derived from UMBRAE [24]. Results are compared to those from other studies, with best, second, and third highlighted. Underline indicates the best result under identical conditions, while * denotes results obtained using BLIP2-generated captions as ground truth.

Method	# Models	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S
		Brain Caption								
SDRecon [9]	4	36.21	17.11	7.72	3.43	10.03	25.13	13.83	5.02	61.07
OneLLM [38]	4	47.04	26.97	15.49	9.51	13.55	35.05	22.99	6.26	54.80
UniBrain [39]	4	–	–	–	–	16.90	22.20	–	–	–
BrainCap [40]	4	55.96	36.21	22.70	14.51	16.68	40.69	41.30	9.06	64.31
UMBRAE [24]	1	57.84	38.43	25.41	17.17	18.70	42.14	53.87	12.27	66.10
Our Method	1	57.19	37.17	23.78	15.85	18.60	36.67	49.51	12.39	65.49
w/o ViT3D	1	52.91	32.18	15.64	8.49	14.07	23.25	39.64	8.34	56.92
Our Method*	1	64.26	51.44	47.70	32.17	20.41	52.61	83.94	18.27	68.72
w/o ViT3D*	1	58.87	42.11	29.48	21.39	15.85	38.48	56.37	11.27	64.35
		Detail Description								
Our Method	1	38.91	24.02	15.24	12.41	18.44	27.83	42.58	18.41	56.16
w/o ViT3D	1	33.57	18.95	11.09	6.13	15.56	23.80	20.23	16.21	51.47
		Complex Reasoning								
Our Method	1	65.41	59.61	50.68	36.46	34.46	62.60	217.83	60.29	80.96
w/o ViT3D	1	60.36	47.81	39.76	30.57	24.37	45.39	150.67	52.13	73.26

not only simple captions but also detailed descriptions and perform complex reasoning. The model's performance increases on complex reasoning tasks, possibly due to the richer semantic information in the questions, which our model captures more effectively. An ablation study was also conducted, revealing a noticeable performance drop in multimodal language tasks when the structural-preserving features of fMRI data were not extracted using ViT3D. Instead, the fMRI data were flattened and patched, similar to methods used in other literature, while maintaining the same fMRI feature extractor structure. This underlines the effectiveness of ViT3D and the capability of our model in multimodal tasks.

4.3 Visual Reconstruction

While our primary objective extends beyond mere visual decoding from fMRI data, visual reconstruction offers a tangible demonstration of a model's comprehension of fMRI data. Therefore, we conducted visual reconstruction experiments and compared our results with those from other studies. The quantitative evaluation highlights our method's proficiency.

Tab. 2 showcases that our model competes with or surpasses traditional subject-specific frameworks on several metrics. Notably, it excels in high-level feature matching, demonstrating the model's ability to effectively leverage LLMs for interpreting complex visual data. The robust performance across various visual stimuli confirms our model's comprehensive understanding of fMRI data. Experiments without key components like LLM and VAE features highlight the significance of each element in our method, which is crucial for achieving state-of-the-art results. Moreover, we have conducted single-trial experiments, opting to use only the first visual stimulus, similar to the approach of MindEye [6], rather than averaging signals from three identical stimuli, which typically escalates data collection costs. Even under these more stringent conditions, our model shows only a slight decrease in performance, enhancing its feasibility for practical applications. Visual reconstruction examples are provided in Fig. 4, illustrating the effectiveness of our approach.

The balance between noise introduction and feature preservation in fMRI data visual reconstruction is governed by the hyperparameter β. Fig. 5 presents a detailed ablation study on how different β values impact various metrics. Fig. 5(a) shows that Pixel Correlation (PixCorr) peaks at intermediate β values, indicating the optimal balance between injected noise and retained prior. The integration of the LLMs does not significantly influence low-level feature capture. In Fig. 5(b), increasing β enhances CLIP accuracy, with LLM integration having a substantial effect on capturing high-level features. Fig. 5(c) indicates the features of the 5 th layer of AlexNet, similar to CLIP features, effectively represent the similarity between reconstructed images and visual stimuli, capturing high-level features accurately. Additionally, Fig. 5(d) illustrates that both the Structural Similarity Index (SSIM) and

https://doi.org/10.52202/079017-3113

Table 2: Quantitative evaluation on visual reconstruction. Performance metrics are reported across different levels of features, with the best, second and third scores highlighted. The underline indicates the best result under the same conditions. Our method achieves state-of-the-art results using a single model trained across subjects (# Models = 1).

Method	# Models	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
Mind-Reader [21]	4	–	–	–	–	78.2%	–	–	–
Takagi *et al* [9]	4	–	–	83.0%	83.0%	76.0%	77.0%	–	–
Gu *et al* [41]	4	.150	.325	–	–	–	–	.862	.465
Brain-Diffuser [42]	4	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423
MindEye [6]	4	.309	.323	94.7%	97.8%	93.8%	94.1%	.645	.367
DREAM [11]	4	.288	.338	95.0%	97.5%	94.8%	95.2%	.638	.413
MindBridge [25]	1	.151	.263	87.7%	95.5%	92.4%	94.7%	.712	.418
UMBRAE [24]	1	.283	.328	93.9%	96.7%	93.4%	94.1%	.700	.393
Our Method	1	.265	.357	93.1%	97.1%	96.8%	97.5%	.633	.321
w/o LLM	1	.263	.369	92.0%	97.1%	94.2%	96.1%	.680	.328
w/o VAE feature	1	.093	.263	84.5%	90.6%	93.6%	95.7%	.684	.398
w/o C_Subj	4	.241	.356	88.1%	95.7%	92.1%	95.1%	.631	.347
w/o C_Subj & ViT3D	4	.164	.273	86.7%	91.4%	89.3%	91.8%	.731	.417
		Single Trial							
MindEye [6]	4	.255	.308	91.6%	95.9%	91.3%	91.6%	.691	.398
Our Method	1	.257	.336	91.2%	96.3%	94.6%	95.3%	.671	.324

Figure 4: Visual reconstruction results showcasing the comparison between (a) using the average signal from all trials and (b) using the first visual stimulus.

CLIP scores benefit from optimally chosen β values, with LLM integration enhancing overall image quality and semantic accuracy. Appropriately adjusting β helps balance the representation of different feature levels in the reconstructed images. Fig. 6 provides examples of visual reconstructions at various β values, demonstrating the model's enhanced capabilities.

4.4 Concept Localization

To further our understanding of semantic concept localization within brain signals, we capitalized on the alignment between our fMRI encoder and CLIP features, which were developed during the training phase. Building on this, we devised a method to localize concepts within brain signals. Specifically, we first fine-tuned Language Models (LLMs) to extract the target concepts from natural language. These concepts, once encoded through the CLIP text encoder, served as targets for GradCAM, which facilitated the localization of the concept within brain signals. To enhance the precision of our localization, we trained three models with varying patch sizes $(14, 12, 10)$ and utilized the penultimate layers of all models to extract semantic features. Fig. 7 illustrates the brain signal localization results for different semantic information, indicating our method's capacity to discriminate the positions of various semantics within brain signals for the same visual stimulus.

https://doi.org/10.52202/079017-3113

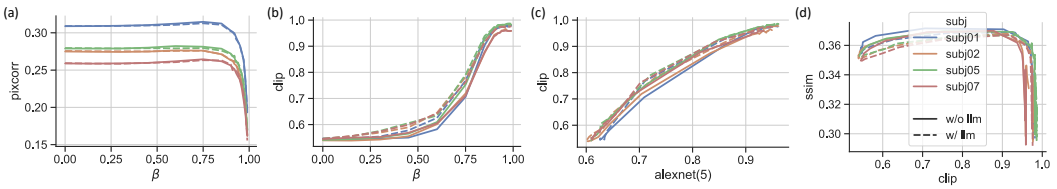

Figure 5: Ablation analysis of the hyperparameter β on visual reconstruction performance.

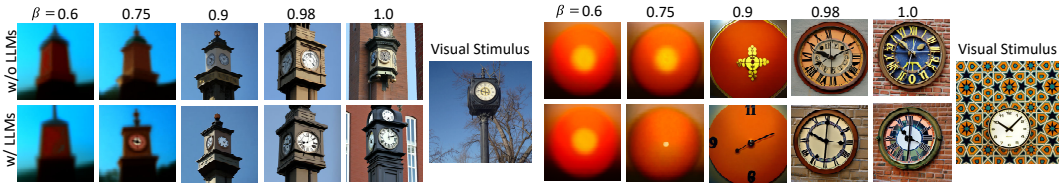

Figure 6: Visualization of the impact of β on visual reconstruction.

To validate the efficacy of our method, we conducted an ablation study on the semantic concepts. After localizing the concepts in the original brain signals, we zeroed out the signals in the identified voxels and performed feature extraction and visual reconstruction using the modified brain signals. As depicted in Fig. 8, the removal of neural activity in specific brain regions associated with certain semantic concepts resulted in the omission of the corresponding semantics in the visual reconstruction. This substantiates the validity of our concept localization method within brain signals and demonstrates our approach's capacity for extracting and modifying semantic information in brain activity, which is pivotal for comprehending semantic information processing in the brain.

5 Conclusion

Our study has successfully developed and validated a novel brain decoding framework that leverages the capabilities of Vision Transformer 3D in conjunction with fMRI data, enhanced by the integration of LLMs. This approach has demonstrated a notable improvement in the reconstruction of visual stimuli from brain signals, offering a more precise and interpretable understanding of the underlying neural mechanisms. The experimental results confirmed the robustness of our model in performing various cognitive tasks, including captioning, question-answering, and visual reconstruction, all from single-trial fMRI data. By enabling accurate localization of linguistic concepts within the brain, our work has potential applications in developing brain-computer interfaces and advanced cognitive modeling. Conclusively, this research contributes to the broader endeavor of decoding and interpreting brain activity, with significant implications for neuroscience and technology interface development. The fusion of advanced AI models with neuroimaging opens new avenues for exploring the intricacies of human cognition and the seamless integration of technology with neural processes.

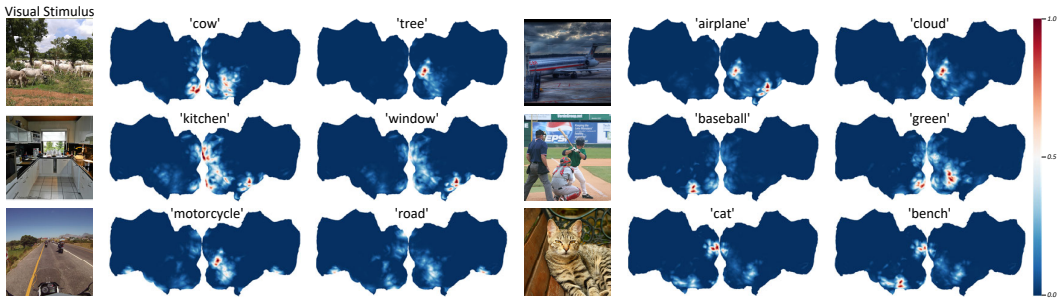

Figure 7: Differential heatmaps of neural activity representing various semantic information for the same visual stimulus.

https://doi.org/10.52202/079017-3113

Figure 8: Validation of concept localization by semantic signal nullification and its effect on visual reconstruction.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Beijing Major Science and Technology Project (Grant No. Z241100001324005).

References

[1] Nikos K Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008.

[2] Yiheng Hu and Qing Yu. Spatiotemporal dynamics of self-generated imagery reveal a reverse cortical hierarchy from cue-induced imagery. *Cell Reports*, 42(10), 2023.

[3] Russell A Poldrack. The future of fmri in cognitive neuroscience. *Neuroimage*, 62(2):1216–1220, 2012.

[4] Russell A Poldrack. The role of fmri in cognitive neuroscience: where do we stand? *Current opinion in neurobiology*, 18(2):223–227, 2008.

[5] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. 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, 2023.

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

[7] Andrew Luo, Margaret Marie Henderson, Michael J Tarr, and Leila Wehbe. Brainscuba: Fine-grained natural language captions of visual cortex selectivity. In *The Twelfth International Conference on Learning Representations*, 2023.

[8] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mindgpt: Interpreting what you see with non-invasive brain recordings. *arXiv preprint arXiv:2309.15729*, 2023.

[9] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[10] Jean Lahoud, Jiale Cao, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Ming-Hsuan Yang. 3d vision with transformers: A survey. *arXiv preprint arXiv:2208.04309*, 2022.

[11] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8226–8235, 2024.

[12] Jingyuan Sun, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens. Contrast, attend and diffuse to decode high-resolution images from brain activities. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Andrew Luo, Maggie Henderson, Leila Wehbe, and Michael Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] 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.

[17] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

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

[19] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

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

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

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

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.

[24] Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision (ECCV)*, 2024.

[25] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. *arXiv preprint arXiv:2404.07850*, 2024.

[26] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[27] Wang Wenxuan, Chen Chen, Ding Meng, Yu Hong, Zha Sen, and Li Jiangyun. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer*, pages 109–119, 2021.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

 https://doi.org/10.52202/079017-3113

[29] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022.

[30] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[32] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[34] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

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

[38] 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. *arXiv preprint arXiv:2312.03700*, 2023.

[39] Weijian Mai and Zhijun Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428*, 2023.

[40] 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.

[41] Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert R Sabuncu. Decoding natural image stimuli from fmri data with a surface-based convolutional network. In *Medical Imaging with Deep Learning*, 2023.

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

[43] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[44] Meta. Meta llama 3. `https://github.com/meta-llama/llama3`, 2024.

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[47] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[48] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[49] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[50] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

 https://doi.org/10.52202/079017-3113

A Dataset

A.1 Natural Scenes Dataset

We conducted our experiments on the Natural Scenes Dataset (NSD) [16], which consists of high-resolution 7Tesla fMRI scans collected from eight healthy adult participants. Our analysis focused on the four subjects (`subj01`, `subj02`, `subj05`, and `subj07`) who completed all data collection sessions. Participants were exposed to thousands of natural scene images from the COCO dataset [34] during the sessions. However, the NSD required preprocessing to correct temporal resampling for slice timing differences and spatial interpolation to adjust for head motion and spatial distortion. We processed the scans in a 1.8-mm native volume space, particularly targeting the "nsdgeneral" region known for its high sensitivity to visual stimuli. This area, predominantly covering the posterior cortex, is crucial for targeted analysis of visual processing.

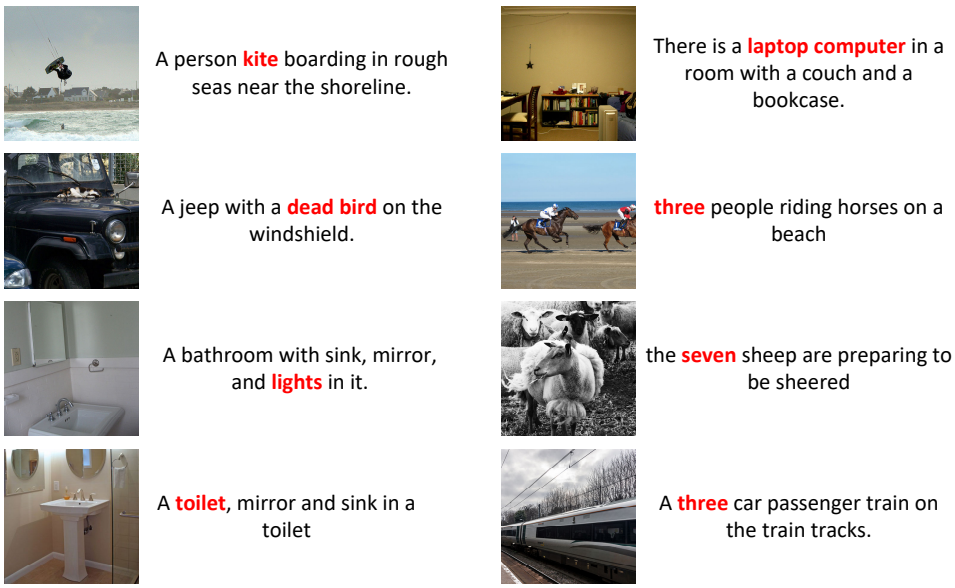

Figure 9: Examples of some images and corresponding captions from the NSD dataset. Due to some image operations such as cropping, there is a mismatch between the original captions and the instance bounding boxes.

Our testing protocols included using the average response from three trials associated with each image to enhance reliability, a common practice in recent studies, as well as assessing each response separately to provide a more challenging and practically relevant test scenario. This approach allowed us to rigorously evaluate our method under realistic and diverse conditions, ensuring thorough validation against established benchmarks. Modifications such as cropping when presenting images to participants led to mismatches between the original captions and instance bounding boxes, as illustrated in Fig. 9. To ensure data consistency, we re-annotated the cropped images, generating eight captions for each image using BLIP2 [36].

A.2 Language Extension

To ensure compatibility between fMRI data and Large Language Models (LLMs) and to enable instruction-following and diversified interactions, we extended the Natural Scenes Dataset (NSD) with natural language annotations. This extension includes seven types of dialogues: brief descriptions, detailed descriptions, continuous dialogues, complex reasoning tasks, instruction reconstruction, and concept localization.

We first generated concise descriptions of the visual stimuli using BLIP2 and integrated these with the original COCO dataset captions to create brief descriptions of the images. Subsequently, DETR [37] was employed to generate bounding boxes for these images. We then combined the image captions and bounding box information as inputs and utilized `GPT-3.5-turbo-0125` to generate various

forms of dialogues. These dialogues were manually adjusted for format and content to ensure consistency and relevancy. For specifics on the prompts used during generation, please refer to the supplementary document.

Furthermore, we constructed a multimodal fine-tuning dataset based on the generated dialogues. For each piece of fMRI data, we created corresponding language extensions. Commands for brief and detailed descriptions are illustrated in Tab. 3 and 4, respectively. We randomly selected questions and corresponding answers from these to construct Q&A pairs, enhancing the model's ability to engage in meaningful dialogue based on the visual content. For continuous dialogues and complex reasoning, we used dialogues generated by `GPT-3.5-turbo-0125`. For instruction reconstruction, the commands are shown in Tab. 5, which aim to have the model generate detailed descriptions of visual stimuli using concise expressions. For concept localization, the commands are shown in Tab. 6, used to extract concepts mentioned in the prompts and visualize the model's attention using grad-CAM [35].

- "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."
- "Provide a brief description of the image."
- "Describe the image briefly."
- "Summarize 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 photoś 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 3: The list of instructions for brief description.

B fMRI Data Preprocessing

We initially preprocess the fMRI data to ensure consistency. Specifically, we resize the data to uniform dimensions using trilinear interpolation. The BOLD signal dimensions for `subj01` are used as the standard, set at $83 \times 104 \times 81$. After applying zero-padding to the edges, we divide the data into $14 \times 14 \times 14$ patches to preserve local information. We then employ a mask created from the union of the "nsdgeneral" regions across all subjects in the NSD dataset, which helps in eliminating information unrelated to the visual stimuli. This process results in data formatted as $N \times C$, where N is the number of retained patches, and C represents the size of each patch.

During the training of our feature extractors, we enhance data variability by applying MixUp [43] to different fMRI responses from the same subject for the same visual stimuli. MixUp coefficients are generated using a uniform distribution, with the mixing ratio $\lambda \sim U(0, 1)$, to blend features from different trials. This technique aids in developing a robust model by exposing it to interpolated data points, fostering generalization across varied neural responses.

C Architecture and Training Details

CLIP ViT-L/14 [14] and Autoencoder KL [15] were used as feature extractors for images, aligning with fMRI data. For the fMRI data, we employed a 16-layer Transformer Encoder [28] with a hidden size of 768 to extract features, using the class token from the last layer as the output. Two

 https://doi.org/10.52202/079017-3113

- "Describe the following image in detail.",
- "Provide a detailed description of the given image.",
- "Give an elaborate explanation of the image you see.",
- "Share a comprehensive rundown of the presented image.",
- "Offer a detailed description of the image.",
- "Describe the image in detail.",
- "Offer a thorough analysis of the image.",
- "Provide a detailed explanation of the subsequent 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.",
- "Explain the image in detail.",
- "Examine the image closely and share its details.",
- "Write an exhaustive depiction of the given image.",

Table 4: The list of instructions for detailed description.

- "Provide the corresponding Stable Diffusion prompts for the image.",

Table 5: The list of instructions for instruction reconstruction.

- "Locating the concept of "<object>",

Table 6: The list of instructions for concept localization.

two-layer perceptrons with a hidden dimension of 1024, f_{w_c} and f_{w_v}, were used to align with CLIP and VAE features, respectively, with the hyperparameter α set to $1/64$. The learning rate was set to 5×10^{-4}, training for 30 epochs. For multimodal interaction, the aforementioned Transformer Encoder was frozen, using its second-to-last layer's hidden states as the fMRI token, processed through a two-layer perceptron f_{w_t}, to interact with Llama-3-8B [44]. Training was divided into two stages: initially, the LLM was frozen, tuning only f_{w_t}, and in the second stage, both the LLM and f_{w_t} were fine-tuned simultaneously. The learning rate was set to 2×10^{-5}, training for one epoch. In the visual reconstruction phase, the LLM called upon UnCLIP-2 [20] for visual reconstruction, with the hyperparameter β set to 0.93. For concept location, the LLM first extracted keywords and then localized them using GradCAM [35], visualizing the results.

All experiments were conducted on a server equipped with 8 NVIDIA A100 GPUs, each with 80 GB of memory. Training of the feature extractor was performed on a single GPU with a batch size set to 32, and the training duration across subjects was approximately 8 hours. For fine-tuning with LLMs, all 8 GPUs were utilized. During the phase where only f_{w_t} was fine-tuned, the batch size was set to 32, and the training time was about 4 hours. In the phase where LLMs were unfrozen for joint fine-tuning, the batch size was adjusted to 24, extending the training time to approximately 36 hours.

C.1 Metrics for Visual Reconstruction

For evaluating visual decoding performance, we adhere to the established suite of eight metrics, commonly utilized in the field [41; 6; 11; 9]. The metrics are divided into two categories: low-level and high-level. Low-level metrics include Pixelwise Correlation (PixCorr) and Structural Similarity Index Metric (SSIM) [45], as well as AlexNet(2) and AlexNet(5) [46], which measure the fidelity of reconstructed images against ground truth. High-level metrics comprise Inception [47], CLIP [14], EfficientNet-B (EffNet-B) [48], and SwAV-ResNet50 (SwAV) [49], which evaluate the semantic accuracy of the reconstructions.

Following protocols from previous research [11], we downsampled the generated images from a resolution of 512×512 to 425×425, which matches the resolution of ground truth images in the NSD Dataset. This resolution adjustment was specifically for PixCorr and SSIM evaluations. For other metrics, the generated images were processed according to the input requirements of each respective model.

Two-way identification tests were conducted following the methodology of Ozcelik and VanRullen [42]. For each model, we calculated the Pearson correlation between embeddings of the ground truth image and its reconstruction, as well as between the ground truth image and another random reconstruction from the test set. A test was marked as correct if the correlation with the ground truth was higher than with the unrelated reconstruction. Performance for each test sample was averaged over all possible pairwise comparisons with the other 981 reconstructions to eliminate any bias from random selection. This resulted in 982 averaged percentage correct outputs, which were then averaged to derive the final metrics presented in Tab 2.

In addition to the established metrics, we introduced a testing protocol by utilizing only the first fMRI record for each visual stimulus to construct a single-trial test [6]. This approach presents a more stringent and practical challenge, reflecting a scenario closer to real-world applications where each neural response is unique and not averaged over multiple instances.

D More Results

D.1 Visual Reconstruction

Fig. 10 shows some randomly selected visual reconstruction results of different subjects under the single-trial condition. Consistent with the results in the main article, our method produces consistent visual reconstruction results across different subjects. This shows that our method has good generalization performance across different subjects. But there are also some erroneous reconstruction results.

Table 7: Performance comparison on visual reconstruction tasks when using different LLMs and different instructions. The underline indicates the best result under the same conditions.

Method	Instruction	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
w/o LLM	-	.263	<u>.369</u>	92.0%	97.1%	94.2%	96.1%	.680	.328
w/ Vicuna-13B	briefly descriptions	.259	.351	91.5%	96.6%	95.1%	96.7%	.641	.337
w/ Vicuna-13B	instruction reconstruction	.257	.361	92.1%	96.9%	96.2%	97.1%	<u>.628</u>	.331
w/ Llama-3-8B	briefly descriptions	.261	.354	92.7%	96.8%	96.4%	97.0%	.637	.327
w/ Llama-3-8B	instruction reconstruction	<u>.265</u>	.357	<u>93.1%</u>	<u>97.1%</u>	<u>96.8%</u>	<u>97.5%</u>	.633	<u>.321</u>

Tab. 7 shows the performance comparison of visual reconstruction tasks when using different large language models (LLMs) and different generation instructions. We compared the two models Vicuna-13B [50] and Llama-3-8B [44]. The results indicate that our method is compatible with various LLMs and performs better in visual reconstruction tasks when using more powerful LLMs. When using Llama-3-8B, the best performance was achieved across the majority of metrics. Additionally, we explored the impact of different generation instructions on the visual reconstruction task. The results show that using instructions specifically designed for visual reconstruction can improve the performance of the visual reconstruction task. We believe this is because these instructions enable the LLMs to generate more detailed descriptions of visual stimuli, thereby enhancing the performance of the visual reconstruction task.

https://doi.org/10.52202/079017-3113

Figure 10: More visual reconstruction results for different subjects under single-trial conditions.

D.2 Language Interaction

We show some examples of multimodal interactions from fMRI data, as shown in Fig. 11. The model conducted multimodal interactions based on fMRI signals and generated different forms of dialogue, including brief descriptions, detailed descriptions, and complex reasoning.

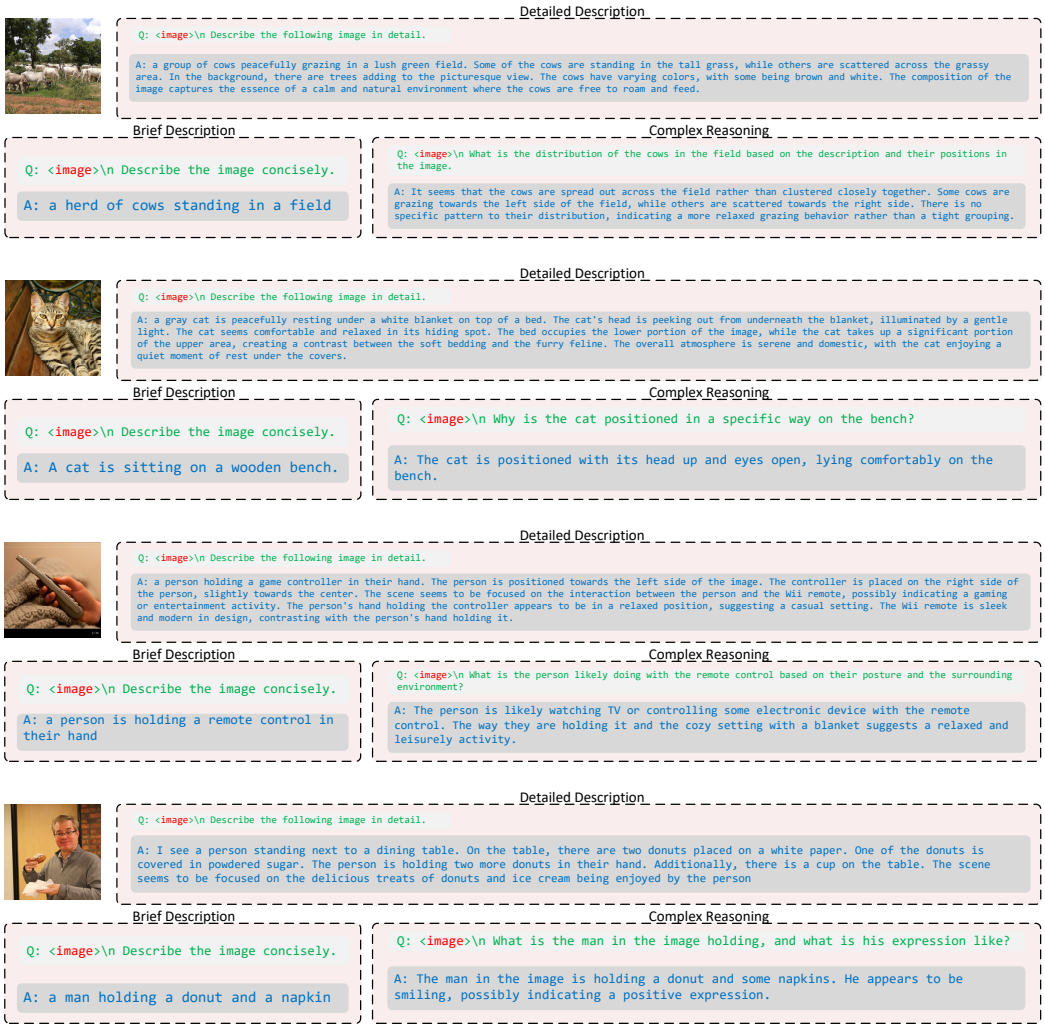

Figure 11: Some examples of multimodal interactions from fMRI data. The images are for reference only. The model performs multi-modal interaction based on fMRI signals and generates different forms of dialogue, including brief descriptions, detailed descriptions, and complex reasoning.

D.3 Concept Localization

In the section, we explored the neural correlates of visual stimuli by mapping the captions derived from the visual content directly onto brain signals. This process involved using GradCAM [35] to generate heatmaps that visually represent the regions of the brain activated in response to specific elements of the visual stimuli. These heatmaps provide a compelling visualization of how different concepts associated with the images are processed across various areas of the brain.

Fig. 12 displays heatmaps that localize the brain activity corresponding to the captions of visual stimuli. These images are crucial for understanding the distribution and intensity of neural responses as they relate to the cognitive processing of visual information. By analyzing these heatmaps, we can infer which areas of the brain are most involved in the interpretation and semantic processing of the stimuli, providing insights into the underlying mechanisms of visual perception and cognitive response.

The localization of these concepts within the brain not only aids in validating theoretical models of brain function but also enhances our understanding of the cognitive processes involved in visual perception. Such detailed mappings are instrumental in advancing our knowledge of the brain's architecture and its functional connectivity in response to complex stimuli.

https://doi.org/10.52202/079017-3113

Figure 12: Heatmaps illustrating brain regions activated by specific captions derived from visual stimuli, demonstrating the spatial distribution of neural responses.

E Limitations

While our study introduces several innovative approaches to the decoding of non-invasive brain recordings and extends the capabilities of visual reconstruction using advanced computational models, there are several limitations that should be acknowledged:

Generalization across Diverse Populations: Our method was validated primarily using the Natural Scenes Dataset (NSD), which consists of data from a limited number of subjects. Although we demonstrate robustness across these subjects, the generalizability of our findings to broader populations remains an area for further investigation. Differences in neural anatomy and functional organization across individuals that are not represented in the NSD could affect the efficacy and accuracy of our model in wider applications.

Computational Complexity and Resource Requirements: The implementation of our framework, particularly the integration with Large Language Models (LLMs) and advanced image processing techniques like ViT3D, requires substantial computational resources. This might restrict the utility of our approach in environments with limited computational capacity or in real-time applications where rapid processing is crucial.

Challenges in Real-World Application: While the single-trial test setting introduced in our study adds a layer of practical relevance by evaluating model performance in a more realistic scenario, it also presents challenges. The variability in single-trial fMRI responses, which can be influenced by numerous uncontrollable factors such as minor head movements or physiological fluctuations, may lead to inconsistencies in decoding accuracy. This variability emphasizes the need for further refinement of noise reduction and signal processing techniques.

Ethical Considerations: The development and application of brain decoding technologies raise ethical questions, particularly concerning privacy and consent. As our methods advance and potentially become capable of decoding more detailed and sensitive information from brain data, ensuring ethical deployment and the protection of participant data becomes paramount.

These limitations highlight the need for continuous improvement and careful consideration in the deployment of brain decoding technologies. Addressing these challenges through further research and development will be crucial for realizing the full potential of non-invasive brain decoding in both scientific research and clinical applications.

 https://doi.org/10.52202/079017-3113

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Appendix E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The manuscript does not contain theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Appendices A, B and C we describe in detail the settings required to reproduce the experiment, and for the datasets and key code that will be released with this manuscript, we have provided it in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-3113

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code and language extensions for the NSD dataset are provided in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed training details and evaluation metrics are provided in Appendices B and C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The results of the manuscript do not include error bars, mainly for the following reasons: 1. In order to make a fair comparison with other methods, other literature in this field does not report error bars. 2. Multiple experiments involving LLMs are computationally expensive. 3. The results of this manuscript have been verified on different subjects, and the results are consistent.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed hardware facilities and computational overhead are provided in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This manuscript does not involve direct human subjects; all data used are from publicly available datasets that comply with the ethical standards of NIPS.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Detailed social impact is discussed in the Conclusion and Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

https://doi.org/10.52202/079017-3113

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no relevant risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We do cite all the existing assets in our paper as well as in our codebase.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

 https://doi.org/10.52202/079017-3113

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.