Novelty Assessment Report

Paper: Seeing Through the Brain: New Insights from Decoding Visual Stimuli with fMRI

PDF URL: https://openreview.net/pdf?id=88ZLp7xYxw

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-08

Abstract

Understanding how the brain encodes visual information is a central challenge in neuroscience and machine learning. A promising approach is to reconstruct visual stimuli—essentially images—from functional Magnetic Resonance Imaging (fMRI) signals. This involves two stages: transforming fMRI signals into a latent space and then using a pre-trained generative model to reconstruct images. The reconstruction quality depends on how similar the latent space is to the structure of neural activity and how well the generative model produces images from that space. Yet, it remains unclear which type of latent space best supports this transformation and how it should be organized to represent visual stimuli effectively.

We present two key findings. First, fMRI signals are more similar to the text space of a language model than to either a vision-based space or a joint text–image space. Second, text representations and the generative model should be adapted to capture the compositional nature of visual stimuli, including objects, their detailed attributes, and relationships. Building on these insights, we propose PRISM, a model that Projects fMRI sIgnals into a Structured text space as an interMediate representation for visual stimuli reconstruction. It includes an object-centric diffusion module that generates images by composing individual objects to reduce object detection errors, and an attribute–relationship search module that automatically identifies key attributes and relationships that best align with the neural activity.

Extensive experiments on real-world datasets demonstrate that our framework outperforms existing methods, achieving up to an 8% reduction in perceptual loss. These results highlight the importance of using structured text as the intermediate space to bridge fMRI signals and image reconstruction.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **reconstructing visual stimuli from fMRI brain signals**

A total of **50 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Generative Model Architectures for Image Reconstruction**
- **Semantic and Multimodal Integration Approaches**
- **Neural Encoding and Representation Learning**
- **Cross-Subject and Few-Shot Generalization**
- **Dynamic and Temporal Visual Reconstruction**
- **Specialized Decoding Tasks and Applications**
- **Methodological Advances and Optimization**
- **Comprehensive Analyses and Benchmarking**

Complete Taxonomy Tree

- reconstructing visual stimuli from fMRI brain signals Survey Taxonomy
- Generative Model Architectures for Image Reconstruction
  - Latent Diffusion Model-Based Reconstruction (7 papers)
  - [1] MindLDM: Reconstruct Visual Stimuli from fMRI Using Latent Diffusion Model (Jun-Hao Guo, 2024) View paper
  - [9] Natural scene reconstruction from fMRI signals using generative latent diffusion (Furkan Ozcelik, 2023) View paper
  - [14] Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors (Scotti, 2023) View paper
  - [21] NeuralDiffuser: Neuroscience-Inspired Diffusion Guidance for fMRI Visual Reconstruction (Haoyu Li, 2025) View paper
  - [24] Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion (Ozcelik, 2023) View paper
  - [29] Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs (Takagi, 2023) View paper
  - [42] NeuralDiffuser: Controllable fMRI Reconstruction with Primary Visual Feature Guided Diffusion (Li Haoyu, 2024) View paper
  - Generative Adversarial Network-Based Reconstruction (4 papers)
  - [5] Reconstructing natural scenes from fmri patterns using bigbigan (Mozafari, 2020) View paper
  - [20] Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning (Ziqi Ren, 2021) View paper
  - [36] Semantics-guided hierarchical feature encoding generative adversarial network for visual image reconstruction from brain activity (Lu Meng, 2024) View paper
  - [39] Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans (Furkan Ozcelik, 2022) View paper
  - Variational Autoencoder-Based Reconstruction (1 papers)
  - [48] Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex (Kuan Han, 2019) View paper

- Semantic and Multimodal Integration Approaches
  - Vision-Language Model Integration (3 papers)
  - [3] Decoding visual neural representations by multimodal learning of brain-visual-linguistic features (Changde Du, 2023) View paper
  - [32] BrainCLIP: Bridging Brain and Visual-Linguistic Representation via CLIP for Generic Natural Visual Stimulus Decoding from fMRI (Liu Yu-long, 2023) View paper
  - [44] BrainCLIP: Brain Representation via CLIP for Generic Natural Visual Stimulus Decoding (Yongqiang Ma, 2025) View paper
  - Language-Centric Semantic Decoding ★ (6 papers)
  - [0] Seeing Through the Brain: New Insights from Decoding Visual Stimuli with fMRI (Anon et al., 2026) View paper
  - [8] Improved image reconstruction from brain activity through automatic image captioning (Fatemeh Kalantari, 2025) View paper
  - [10] Beyond brain decoding: Visual-semantic reconstructions to mental creation extension based on fmri (H Jing, 2025) View paper
  - [19] Mindsemantix: Deciphering brain visual experiences with a brain-language model (Ren ZiQi, 2024) View paper
  - [26] UniBrain: Unify Image Reconstruction and Captioning All in One Diffusion Model from Human Brain Activity (Weijian Mai, 2023) View paper
  - [49] Brain Captioning: Decoding human brain activity into images and text (Ferrante, 2023) View paper
  - Compositional and Attribute-Based Reconstruction (2 papers)
  - [13] Dream: Visual decoding from reversing human visual system (Weihao Xia, 2024) View paper
  - [25] Rethinking visual reconstruction: Experience-based content completion guided by visual cues (J Chen, 2023) View paper
- Neural Encoding and Representation Learning
  - Hierarchical and Attention-Based Encoding (3 papers)
  - [15] Attention module improves both performance and interpretability of fourâ□□dimensional functional magnetic resonance imaging decoding neural network (Zhoufan Jiang, 2022) View paper
  - [16] Functional diversity of visual cortex improves constraint-free natural image reconstruction from human brain activity (Lingxiao Yang, 2023) View paper
  - [40] MoRE-Brain: Routed Mixture of Experts for Interpretable and Generalizable Cross-Subject fMRI Visual Decoding (Wei, 2025) View paper
  - Specialized Encoding Architectures (4 papers)
  - [11] Decoding visual fMRI stimuli from human brain based on graph convolutional neural network (Lu Meng, 2022) View paper
  - [28] Neuro-Vision to Language: Image Reconstruction and Language enabled Interaction via Brain Recordings (Shen, 2024) View paper
  - [33] Accurate reconstruction of image stimuli from human functional magnetic resonance imaging based on the decoding model with capsule network architecture (Kai Qiao, 2018) View paper
  - [43] Neuro-Vision to Language: Enhancing Brain Recording-based Visual Reconstruction and Language Interaction (Yiting Dong, 2024) View paper
  - Neuroscience-Inspired Encoding Models (2 papers)
  - [34] Generative decoding of visual stimuli (E Miliotou, 2023) View paper
  - [47] Neural encoding and decoding with deep learning for dynamic natural vision (Wen Haiguang, 2018) View paper
- Cross-Subject and Few-Shot Generalization (3 papers)
  - [4] Reconstructing Retinal Visual Images from 3T fMRI Data Enhanced by Unsupervised Learning (Yujian Xiong, 2024) View paper
  - [22] Retrieving and reconstructing conceptually similar images from fMRI with latent diffusion models and a neuro-inspired brain decoding model (Matteo Ferrante, 2024) View paper
  - [37] MindShot: A Few-Shot Brain Decoding Framework via Transferring Cross-Subject Prior and Distilling Frequency Domain Knowledge (Jiang Shuai, 2024) View paper
- Dynamic and Temporal Visual Reconstruction (3 papers)
  - [7] Reconstructing visual experiences from brain activity evoked by natural movies (Shinji Nishimoto, 2011) View paper
  - [23] Making Your Dreams A Reality: Decoding the Dreams into a Coherent Video Story from fMRI Signals (Fu, 2025) View paper
  - [30] Cinematic mindscapes: High-quality video reconstruction from brain activity (Chen Zijiao, 2023) View paper
- Specialized Decoding Tasks and Applications
  - Mental Imagery and Internal Representation Decoding (1 papers)
  - [27] NSD-Imagery: A benchmark dataset for extending fMRI vision decoding methods to mental imagery (Kneeland, 2025) View paper
  - Non-Visual Stimulus Decoding (1 papers)
  - [6] Neural Evidence for Tonal Prediction: Multivariate Decoding of Predicted Tone Categories Using Functional Magnetic Resonance Imaging Data (Shun Liu, 2026) View paper
  - Emotion and High-Level Semantic Decoding (1 papers)
  - [35] Decoding facial emotion from activity in the human visual cortex using functional magnetic resonance imaging (Qiang Yang, 2019) View paper
- Methodological Advances and Optimization (2 papers)
  - [17] Optimized AI-based neural decoding from BOLD fMRI signal for analyzing visual and semantic ROIs in the human visual system (Lorenzo Veronese, 2025) View paper
  - [41] Brain Decoding and Reconstruction of concepts of visual stimuli from fMRI through deep diffusion models (Matteo Ferrante, 2024) View paper
- Comprehensive Analyses and Benchmarking (8 papers)
  - [2] Mind reader: Reconstructing complex images from brain activities (Lin Sikun, 2022) View paper
  - [12] Visual Image Reconstruction from Brain Activity via Latent Representation (Kamitani Yukiyasu, 2025) View paper
  - [18] Decoding Visual Experience and Mapping Semantics through Whole-Brain Analysis Using fMRI Foundation Models (Wang Yanchen, 2024) View paper
  - [31] A Survey on fMRI-based Brain Decoding for Reconstructing Multimodal Stimuli (Liu Pengyu, 2025) View paper
  - [38] for Reconstructing Visual Stimuli (S Battula, 2024) View paper
  - [45] Decoding visual consciousness from human brain signals (Haynes, 2009) View paper
  - [46] Perception-to-image: Reconstructing natural images from the brain activity of visual perception (Wei Huang, 2020) View paper
  - [50] Scaling laws for decoding images from brain activity (Banville, 2025) View paper

Narrative

Core task: reconstructing visual stimuli from fMRI brain signals. The field has evolved into several major branches that reflect different strategic emphases. Generative Model Architectures focus on leveraging modern diffusion and GAN-based frameworks to produce high-fidelity images directly from neural data, as seen in works like MindLDM[1] and Brain Diffuser[24]. Semantic and Multimodal Integration Approaches emphasize bridging brain signals with language or conceptual representations, often using pretrained vision-language models to guide reconstruction through semantic priors. Neural Encoding and Representation Learning investigates how to best map fMRI voxels to latent feature spaces that capture hierarchical visual information. Cross-Subject and Few-Shot Generalization tackles the challenge of training models that work across individuals or with limited data, while Dynamic and Temporal Visual Reconstruction extends methods to video or time-varying stimuli. Specialized Decoding Tasks explore applications such as emotion recognition or retinal imaging, and Methodological Advances refine optimization strategies and model architectures. Comprehensive Analyses and Benchmarking provide systematic evaluations across datasets and methods.

Within the Semantic and Multimodal Integration branch, a particularly active line of work explores language-centric decoding, where textual descriptions or captions serve as intermediate representations to constrain and enrich image generation. Seeing Through Brain[0] exemplifies this direction by integrating linguistic semantic information to improve reconstruction fidelity and interpretability. Nearby efforts such as Image Captioning[8] and Brain Captioning[49] similarly leverage natural language as a bridge between neural activity and visual content, while Visual Semantic Reconstructions[10] and MindSemantix[19] explore how semantic embeddings can guide generative models. Compared to purely image-driven approaches like BigBiGAN Reconstruction[5] or Multimodal Brain Visual[3], these language-centric methods trade some direct pixel-level control for enhanced semantic coherence and cross-modal alignment, reflecting an ongoing tension between low-level perceptual accuracy and high-level conceptual fidelity.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. Improved image reconstruction from brain activity through automatic image captioning

Authors: Fatemeh Kalantari, Karim Faez, Hamidreza Amindavar, Soheila Nazari, H. Amindavar, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Significant progress has been made in the field of image reconstruction using functional magnetic resonance imaging (fMRI). Certain investigations reconstructed images with visual information decoded from brain signals, yielding insufficient accuracy and quality. The combination of semantic information in the reconstruction was recommended to improve performance. However, this issue continues to come across numerous difficulties. To address such problems, we proposed an approach that combines se...

Relationship Analysis

Both papers belong to the Language-Centric Semantic Decoding category, using text representations as intermediate steps for fMRI-to-image reconstruction. They overlap in decoding fMRI signals into textual descriptions (captions) before generating images, and both leverage vision-language models (VLMs) for semantic guidance. However, the original paper (PRISM) emphasizes structured, object-centric text descriptions with automatic attribute-relationship search and compositional diffusion generation, while the candidate paper uses BLIP-generated captions combined with visual features and applies latent diffusion models conditioned on decoded semantic features without explicit compositional object modeling.

2. Beyond brain decoding: Visual-semantic reconstructions to mental creation extension based on fmri

Authors: H Jing, D Jiang, Y Ma, H Hua | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ neural decoding model that achieves unity between LLM and the multimodal visual decoding of the human brain. Excitingly, we realized a creative extension based on the visual â¦

Relationship Analysis

Both papers belong to the Language-Centric Semantic Decoding category, using text representations as intermediate steps for fMRI-to-image reconstruction. The original paper (PRISM) focuses on projecting fMRI signals into structured text space with object-centric descriptions and compositional generation, while the candidate paper (NeuroCreat) emphasizes multimodal integration with LLMs for semantic extraction and extends to creative tasks like video generation from visual working memory. The key difference is that PRISM concentrates on structured compositional text representations for accurate reconstruction, whereas NeuroCreat leverages LLMs for broader multimodal tasks including creative extensions beyond reconstruction.

3. Mindsemantix: Deciphering brain visual experiences with a brain-language model

Authors: Ren ZiQi, Li Jie, Ziqi Ren, Xue, Xuetong, et al. (16 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Deciphering the human visual experience through brain activities captured by fMRI represents a compelling and cutting-edge challenge in the field of neuroscience research. Compared to merely predicting the viewed image itself, decoding brain activity into meaningful captions provides a higher-level interpretation and summarization of visual information, which naturally enhances the application flexibility in real-world situations. In this work, we introduce MindSemantix, a novel multi-modal fram...

Relationship Analysis

Both papers belong to the Language-Centric Semantic Decoding category, using text representations as intermediate steps for fMRI-to-image reconstruction. They overlap in leveraging language models to decode fMRI signals into textual descriptions that guide image generation, with both employing diffusion models for final reconstruction. However, the original paper (PRISM) focuses on structured, object-centric text descriptions with attribute-relationship search and compositional generation, while the candidate paper (MindSemantix) constructs an end-to-end Brain-Language Model integrating LLMs directly into brain activity comprehension via a Brain Q-Former architecture for multi-modal alignment.

4. UniBrain: Unify Image Reconstruction and Captioning All in One Diffusion Model from Human Brain Activity

Authors: Weijian Mai, Zhang Zhijun, Zhijun Zhang | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Image reconstruction and captioning from brain activity evoked by visual stimuli allow researchers to further understand the connection between the human brain and the visual perception system. While deep generative models have recently been employed in this field, reconstructing realistic captions and images with both low-level details and high semantic fidelity is still a challenging problem. In this work, we propose UniBrain: Unify Image Reconstruction and Captioning All in One Diffusion Mode...

Relationship Analysis

Both papers belong to the Language-Centric Semantic Decoding category, using text representations as intermediate steps for fMRI-to-image reconstruction. They overlap in leveraging language models to bridge fMRI signals and visual reconstruction, with both employing diffusion models for generation. The key difference is that the original paper (PRISM) focuses on structured, object-centric text descriptions with explicit attribute-relationship search and compositional generation, while the candidate paper (UniBrain) unifies both image reconstruction and captioning tasks in a single diffusion framework using versatile diffusion with multi-modal CLIP conditions.

5. Brain Captioning: Decoding human brain activity into images and text

Authors: Ferrante, Matteo, Ozcelik, Furkan, Matteo Ferrante, et al. (15 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Every day, the human brain processes an immense volume of visual information, relying on intricate neural mechanisms to perceive and interpret these stimuli. Recent breakthroughs in functional magnetic resonance imaging (fMRI) have enabled scientists to extract visual information from human brain activity patterns. In this study, we present an innovative method for decoding brain activity into meaningful images and captions, with a specific focus on brain captioning due to its enhanced flexibili...

Relationship Analysis

Both papers belong to the Language-Centric Semantic Decoding category, using text representations as intermediate steps for fMRI-to-image reconstruction. The original paper (PRISM) focuses on structured, object-centric text descriptions with compositional attributes and relationships, employing a language model (T5/LLaMA3) as the primary latent space and an object-centric diffusion module for generation. The candidate paper (Brain Captioning) emphasizes generating natural language captions using the GIT (Generative Image-to-text Transformer) model and incorporates depth estimation with ControlNet for image reconstruction, representing a more holistic captioning approach rather than structured compositional descriptions.

Contributions Analysis

Overall novelty summary. The paper proposes PRISM, which projects fMRI signals into a structured text space to reconstruct visual stimuli. It sits within the Language-Centric Semantic Decoding leaf of the taxonomy, which contains six papers including the original work. This leaf is part of the broader Semantic and Multimodal Integration Approaches branch, indicating a moderately populated research direction. The approach differs from sibling methods by emphasizing structured compositional text representations rather than direct caption generation or unstructured semantic embeddings, positioning it at the intersection of language-based decoding and compositional modeling.

The taxonomy reveals that Language-Centric Semantic Decoding is one of three subtopics under Semantic and Multimodal Integration, alongside Vision-Language Model Integration (three papers) and Compositional and Attribute-Based Reconstruction (two papers). The paper's emphasis on structured text space connects it to both neighboring directions: it shares the language-centric philosophy with sibling papers like Seeing Through Brain and Brain Captioning, while its compositional modeling approach aligns with the attribute-based reconstruction leaf. This positioning suggests the work bridges two related but distinct research threads within the semantic integration paradigm.

Among thirty candidates examined, the PRISM framework contribution shows one refutable candidate from ten examined, while the other two contributions (text space alignment and compositional modeling) show no clear refutations from their respective ten-candidate searches. The limited refutation suggests that while the overall framework may have some overlap with prior work, the specific claims about text space superiority and compositional structure appear less directly challenged within the examined literature. However, the modest search scope means these findings reflect top-thirty semantic matches rather than exhaustive coverage of the field's approximately fifty papers.

Based on the limited search scope, the work appears to occupy a recognizable but not heavily saturated niche within language-centric fMRI decoding. The single refutable pair among thirty candidates suggests moderate novelty, though the analysis cannot rule out additional overlaps beyond the top-ranked semantic matches. The positioning between language-based and compositional approaches may offer differentiation, but definitive assessment would require broader literature coverage beyond the examined subset.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: fMRI signals align better with language model text space than vision-based spaces

Description: The authors demonstrate through empirical analysis using CKA, CCA, and Generalization Gap metrics that fMRI signals exhibit stronger alignment with pure text embeddings from language models compared to vision model representations or joint vision-language spaces, challenging the assumption that vision-based representations are essential for visual stimuli reconstruction.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

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

URL: View paper

Brief Assessment

Enhanced Language Interaction[43] focuses on integrating fMRI features with visual embeddings using a vision transformer and LLMs for multimodal tasks, not on comparing alignment between fMRI signals and different representation spaces (text vs. vision).

2. Modality-agnostic fmri decoding of vision and language

URL: View paper

Brief Assessment

Modality Agnostic[65] focuses on modality-agnostic decoding (predicting stimuli regardless of presentation modality) rather than comparing alignment strength between fMRI and different representation spaces for reconstruction tasks. Their finding that language models perform comparably to multimodal models for decoding does not directly challenge the original paper's claim about alignment superiority for reconstruction purposes.

3. Modeling the human visual system: Comparative insights from response-optimized and task-optimized vision models, language models, and different readout â¦

URL: View paper

Brief Assessment

Human Visual System[62] examines alignment between fMRI and different model representations but focuses on comparative performance across visual hierarchy regions rather than challenging the novelty of using text embeddings for fMRI alignment. The candidate explores vision vs. language model effectiveness across cortical regions, not the foundational claim of text-space superiority.

4. Beyond brain decoding: Visual-semantic reconstructions to mental creation extension based on fmri

URL: View paper

Brief Assessment

Visual Semantic Reconstructions[10] focuses on multimodal brain decoding with LLMs and visual working memory for creative generation tasks, not on comparative alignment analysis between fMRI signals and different representation spaces (text vs. vision).

5. Brainchat: Decoding semantic information from fmri using vision-language pretrained models

URL: View paper

Brief Assessment

Brainchat[63] focuses on aligning fMRI with vision-language model embeddings (CoCa) for semantic decoding tasks, not on comparing pure text embeddings versus vision embeddings. The candidate does not challenge the original paper's empirical finding about text space superiority.

6. Language models align with brain regions that represent concepts across modalities

URL: View paper

Brief Assessment

Language Models Align[59] focuses on identifying brain regions that represent concepts consistently across modalities (text, images, word clouds) and measuring alignment with those regions. The original paper focuses on visual stimuli reconstruction quality and demonstrates that text embeddings are more effective than vision embeddings for this reconstruction task. These are distinct research objectives with different methodologies and applications.

7. Brain-streams: fMRI-to-image reconstruction with multi-modal guidance

URL: View paper

Brief Assessment

Brain Streams[52] focuses on extracting multi-modal guidance from different brain regions (ventral visual cortex for semantic, early visual cortex for perceptual) rather than comparing alignment between fMRI signals and different representation spaces (text vs. vision embeddings).

8. Llm4brain: Training a large language model for brain video understanding

URL: View paper

Brief Assessment

Llm4brain[61] focuses on video-stimulated fMRI semantic decoding using LLMs for cross-subject reconstruction, not on comparing alignment between fMRI signals and different embedding spaces (text vs. vision). The candidate does not challenge the original paper's empirical finding about relative alignment strengths.

9. Mind2word: Towards generalized visual neural representations for high-quality video reconstruction

URL: View paper

Brief Assessment

Mind2word[60] focuses on video reconstruction from fMRI using text embeddings as an intermediate representation, but does not provide comparative analysis between text space, vision space, and joint vision-language spaces using metrics like CKA, CCA, or generalization gap to demonstrate alignment superiority.

10. BrainChat: Interactive Semantic Information Decoding from fMRI Using Large-Scale Vision-Language Pretrained Models

URL: View paper

Brief Assessment

BrainChat Interactive[64] focuses on aligning fMRI with both pretrained image and text embeddings to create a unified representation for interactive captioning and question answering, rather than comparing alignment strengths across different representation spaces as the original paper does.

Contribution 2: PRISM framework for fMRI-to-image reconstruction via structured text space

Description: The authors introduce PRISM, a novel framework that maps fMRI signals into a structured text space capturing compositional visual information (objects, attributes, and relationships). The framework includes an object-centric diffusion module for compositional image generation and an attribute-relationship search module that automatically identifies brain-aligned attributes and relationships using a vision-language model.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. NeuroAdapter: Visual Reconstruction with Masked Brain Representation

URL: View paper

Brief Assessment

NeuroAdapter[54] takes a fundamentally different approach by directly encoding neural data to condition diffusion without intermediate text/image representations, whereas PRISM explicitly maps fMRI to structured text space with object-centric descriptions. The candidate does not demonstrate prior work using structured text space as an intermediate representation.

2. Mind reader: Reconstructing complex images from brain activities

URL: View paper

Prior Art Analysis

Mind Reader[2] demonstrates prior work that maps fMRI signals to text space as an intermediate representation for image reconstruction. The candidate paper explicitly states they use text modality and map fMRI signals to CLIP text embeddings, then use these embeddings to condition a generative model for reconstruction. This directly challenges the novelty claim of using structured text space as an intermediate representation, as Mind Reader[2] already employed text-based intermediate representations from language models for fMRI-to-image reconstruction before the original paper.

Evidence

Evidence 1 - **Rationale**: Both papers use text space as an intermediate representation for fMRI-to-image reconstruction, demonstrating that the core concept of using text space as a bridge was already established in Mind Reader[2]. - **Original**: we proposeprism, a model thatprojects fmri signals into astructured text space as an intermediate representation for visual stimuli reconstruction. - **Candidate**: we propose to map fmri to a well-aligned space shared by image and text, and use conditional generative models to reconstruct seen images from representations in that space.

Evidence 2 - **Rationale**: Mind Reader[2] explicitly maps fMRI to text embeddings (hcap from CLIP text encoder), demonstrating prior use of text space as an intermediate representation for reconstruction. - **Original**: our first finding (section 3.1) shows that fmri signals align more closely with the text space of lm, motivating the use of solely text as a bridge for reconstruction. - **Candidate**: we choose to use the roi with the widest region coverage, and the lengthn of xfmri ranges from 12682 to 17907 for different brains in the nsd dataset. our goal is to train two mapping models,fmi and fmc in fig. 1 (collectively denoted asfm), that encodes xfmri 2 rn to himg =c img(ximg) 2 r512 and hca...

Evidence 3 - **Rationale**: Both papers compare different modalities and conclude that text modality is beneficial for fMRI-to-image reconstruction, showing Mind Reader[2] already explored this design choice. - **Original**: to identify the most effective intermediate space, we compare fmri signals with representations from pre-trained vision, language, and vision-language models using established metrics - **Candidate**: we find that incorporating an additional text modality is beneficial for the reconstruction problem compared to directly translating brain signals to images. therefore, the modalities involved in our method are: (i) voxel-level fmri signals, (ii) observed images that trigger the brain signals, and (i...

Evidence 4 - **Rationale**: Mind Reader[2] already used vision-language models (CLIP) to create aligned latent spaces and condition generative models on these embeddings, establishing the framework of using VLMs for fMRI-to-image reconstruction before the original paper. - **Original**: we develop two core modules: an object-centric diffusion module that adapts the diffusion model to generate images by composing individual objects, and an attribute-relationship search module that uses a vision-language model (vlm) to automatically identify object attributes and relationships aligne... - **Candidate**: we leverage an aligned vision-language latent space pre-trained on massive datasets. instead of training models from scratch to find a latent space shared by the three modalities, we encode fmri signals into this pre-aligned latent space. then, conditioned on embeddings in this space, we reconstruct ...

3. MindShot: Multi-Shot Video Reconstruction from fMRI with LLM Decoding

URL: View paper

Brief Assessment

MindShot Video[57] focuses on multi-shot video reconstruction from fMRI using shot boundary prediction and keyframe caption decoding, not static image reconstruction via structured text space with object-centric diffusion and attribute-relationship search modules as in PRISM.

4. Generative multimodal decoding: Reconstructing images and text from human fMRI

URL: View paper

Brief Assessment

Generative Multimodal[56] focuses on multimodal decoding using image captioning models (GIT) and diffusion models with depth estimation, not on structured text space with compositional object-attribute-relationship representations as in PRISM. The candidate uses natural language captions as conditioning but does not employ object-centric structured descriptions or attribute-relationship search modules.

5. Neural encoding and decoding with distributed sentence representations

URL: View paper

Brief Assessment

Distributed Sentence[53] focuses on neural encoding and decoding of sentence representations using distributed semantic models to predict fMRI patterns, not on reconstructing visual images from fMRI signals through structured text spaces and diffusion models.

6. Coherent Language Reconstruction from Brain Recordings with Flexible Multi-Modal Input Stimuli

URL: View paper

Brief Assessment

Coherent Language[58] focuses on brain-to-text decoding across multiple modalities (fMRI, EEG, MEG) using multimodal alignment for language generation, not fMRI-to-image reconstruction via structured text space as an intermediate representation.

7. Alleviating the semantic gap for generalized fmri-to-image reconstruction

URL: View paper

Brief Assessment

Semantic Gap[55] focuses on alleviating semantic gaps between training and testing data using CLIP-based semantic expansion and structural information, not on structured text space as an intermediate representation for compositional visual information (objects, attributes, relationships) as in the original paper.

8. MindLDM: Reconstruct Visual Stimuli from fMRI Using Latent Diffusion Model

URL: View paper

Brief Assessment

MindLDM[1] uses CLIP text feature space as an intermediate representation but does not employ structured, compositional text descriptions with object-level attributes and relationships. The candidate focuses on aligning fMRI features to CLIP text space without the object-centric diffusion or attribute-relationship search modules that define PRISM's structured approach.

9. Brain-streams: fMRI-to-image reconstruction with multi-modal guidance

URL: View paper

Brief Assessment

Brain Streams[52] uses multi-modal guidance (textual, visual, and layout) from different brain regions but does not employ structured text space with compositional elements (objects, attributes, relationships) as an intermediate representation like PRISM.

10. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction

URL: View paper

Brief Assessment

Mindtuner[51] focuses on cross-subject visual decoding using visual fingerprint modeling and fMRI-to-text alignment via image as pivot modality, not on structured text space with compositional visual information (objects, attributes, relationships) as intermediate representation for single-subject reconstruction.

Contribution 3: Structured text representations improve reconstruction through compositional modeling

Description: The authors establish that adapting both the text space and generative model to explicitly represent the compositional structure of visual perception—distinguishing objects, their attributes, and inter-object relationships—leads to improved reconstruction quality compared to unified holistic representations.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. MARS: Paying More Attention to Visual Attributes for Text-Based Person Search

URL: View paper

Brief Assessment

MARS[71] focuses on text-based person search using attribute-noun chunks for person retrieval, not on compositional visual reconstruction from fMRI signals or structured text representations for generative modeling.

2. Learning visual composition through improved semantic guidance

URL: View paper

Brief Assessment

Visual Composition[67] focuses on improving multimodal embeddings for image retrieval through recaptioning and contrastive learning, not on fMRI-to-image reconstruction or compositional text representations for neural decoding.

3. Evaluating Text-to-Visual Generation with Image-to-Text Generation

URL: View paper

Brief Assessment

Text to Visual[66] focuses on evaluating text-to-visual generation models using VQA-based alignment metrics, not on fMRI-to-image reconstruction or compositional visual perception modeling from neural signals.

4. Generating semantically precise scene graphs from textual descriptions for improved image retrieval

URL: View paper

Brief Assessment

Semantic Scene Graphs[72] focuses on parsing natural language descriptions into scene graphs for image retrieval tasks, not on fMRI-to-image reconstruction or neural decoding. The paper addresses a different problem domain (text-to-graph parsing for retrieval) rather than brain activity decoding with compositional visual representations.

5. Evaluating and Improving Compositional Text-to-Visual Generation

URL: View paper

Brief Assessment

Compositional Generation[70] focuses on evaluating text-to-image generation models using compositional prompts with attributes and relationships, not on fMRI-to-image reconstruction or neural signal decoding using structured text representations.

6. Attribute-Centric Compositional Text-to-Image Generation

URL: View paper

Brief Assessment

Attribute Centric[68] focuses on text-to-image generation with attribute-centric compositional modeling to handle underrepresented attribute compositions in training data. The original paper addresses fMRI-to-image reconstruction using structured text as an intermediate representation. These are fundamentally different tasks with different technical approaches and objectives.

7. SceneEval: Evaluating semantic coherence in text-conditioned 3D indoor scene synthesis

URL: View paper

Brief Assessment

SceneEval[75] focuses on evaluating text-conditioned 3D indoor scene synthesis by measuring fidelity to explicit user requirements (object counts, attributes, relationships) and implicit expectations (collision, navigability). It does not address compositional visual reconstruction from fMRI signals or structured text representations for neural decoding, which is the core contribution of the original paper.

8. Componerf: Text-guided multi-object compositional nerf with editable 3d scene layout

URL: View paper

Brief Assessment

Componerf[74] focuses on compositional 3D scene generation from text using multiple NeRFs with spatial layouts, not on fMRI-to-image reconstruction or brain activity decoding. The domains and technical approaches are fundamentally different.

9. Text2scene: Generating compositional scenes from textual descriptions

URL: View paper

Brief Assessment

Text2scene[73] focuses on generating visual scenes from text descriptions using compositional object-attribute modeling, but targets scene synthesis rather than fMRI-to-image reconstruction. The domains and technical approaches differ fundamentally.

10. TextPSG: Panoptic Scene Graph Generation from Textual Descriptions

URL: View paper

Brief Assessment

TextPSG[69] focuses on panoptic scene graph generation from image captions for scene understanding, not fMRI-to-image reconstruction. The original paper addresses neural decoding and visual stimulus reconstruction from brain signals using structured text as an intermediate representation, which is a fundamentally different problem domain.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Seeing Through the Brain: New Insights from Decoding Visual Stimuli with fMRI View paper
- [1] MindLDM: Reconstruct Visual Stimuli from fMRI Using Latent Diffusion Model View paper
- [2] Mind reader: Reconstructing complex images from brain activities View paper
- [3] Decoding visual neural representations by multimodal learning of brain-visual-linguistic features View paper
- [4] Reconstructing Retinal Visual Images from 3T fMRI Data Enhanced by Unsupervised Learning View paper
- [5] Reconstructing natural scenes from fmri patterns using bigbigan View paper
- [6] Neural Evidence for Tonal Prediction: Multivariate Decoding of Predicted Tone Categories Using Functional Magnetic Resonance Imaging Data View paper
- [7] Reconstructing visual experiences from brain activity evoked by natural movies View paper
- [8] Improved image reconstruction from brain activity through automatic image captioning View paper
- [9] Natural scene reconstruction from fMRI signals using generative latent diffusion View paper
- [10] Beyond brain decoding: Visual-semantic reconstructions to mental creation extension based on fmri View paper
- [11] Decoding visual fMRI stimuli from human brain based on graph convolutional neural network View paper
- [12] Visual Image Reconstruction from Brain Activity via Latent Representation View paper
- [13] Dream: Visual decoding from reversing human visual system View paper
- [14] Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors View paper
- [15] Attention module improves both performance and interpretability of fourâ□□dimensional functional magnetic resonance imaging decoding neural network View paper
- [16] Functional diversity of visual cortex improves constraint-free natural image reconstruction from human brain activity View paper
- [17] Optimized AI-based neural decoding from BOLD fMRI signal for analyzing visual and semantic ROIs in the human visual system View paper
- [18] Decoding Visual Experience and Mapping Semantics through Whole-Brain Analysis Using fMRI Foundation Models View paper
- [19] Mindsemantix: Deciphering brain visual experiences with a brain-language model View paper
- [20] Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning View paper
- [21] NeuralDiffuser: Neuroscience-Inspired Diffusion Guidance for fMRI Visual Reconstruction View paper
- [22] Retrieving and reconstructing conceptually similar images from fMRI with latent diffusion models and a neuro-inspired brain decoding model View paper
- [23] Making Your Dreams A Reality: Decoding the Dreams into a Coherent Video Story from fMRI Signals View paper
- [24] Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion View paper
- [25] Rethinking visual reconstruction: Experience-based content completion guided by visual cues View paper
- [26] UniBrain: Unify Image Reconstruction and Captioning All in One Diffusion Model from Human Brain Activity View paper
- [27] NSD-Imagery: A benchmark dataset for extending fMRI vision decoding methods to mental imagery View paper
- [28] Neuro-Vision to Language: Image Reconstruction and Language enabled Interaction via Brain Recordings View paper
- [29] Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs View paper
- [30] Cinematic mindscapes: High-quality video reconstruction from brain activity View paper
- [31] A Survey on fMRI-based Brain Decoding for Reconstructing Multimodal Stimuli View paper
- [32] BrainCLIP: Bridging Brain and Visual-Linguistic Representation via CLIP for Generic Natural Visual Stimulus Decoding from fMRI View paper
- [33] Accurate reconstruction of image stimuli from human functional magnetic resonance imaging based on the decoding model with capsule network architecture View paper
- [34] Generative decoding of visual stimuli View paper
- [35] Decoding facial emotion from activity in the human visual cortex using functional magnetic resonance imaging View paper
- [36] Semantics-guided hierarchical feature encoding generative adversarial network for visual image reconstruction from brain activity View paper
- [37] MindShot: A Few-Shot Brain Decoding Framework via Transferring Cross-Subject Prior and Distilling Frequency Domain Knowledge View paper
- [38] for Reconstructing Visual Stimuli View paper
- [39] Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans View paper
- [40] MoRE-Brain: Routed Mixture of Experts for Interpretable and Generalizable Cross-Subject fMRI Visual Decoding View paper
- [41] Brain Decoding and Reconstruction of concepts of visual stimuli from fMRI through deep diffusion models View paper
- [42] NeuralDiffuser: Controllable fMRI Reconstruction with Primary Visual Feature Guided Diffusion View paper
- [43] Neuro-Vision to Language: Enhancing Brain Recording-based Visual Reconstruction and Language Interaction View paper
- [44] BrainCLIP: Brain Representation via CLIP for Generic Natural Visual Stimulus Decoding View paper
- [45] Decoding visual consciousness from human brain signals View paper
- [46] Perception-to-image: Reconstructing natural images from the brain activity of visual perception View paper
- [47] Neural encoding and decoding with deep learning for dynamic natural vision View paper
- [48] Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex View paper
- [49] Brain Captioning: Decoding human brain activity into images and text View paper
- [50] Scaling laws for decoding images from brain activity View paper
- [51] Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction View paper
- [52] Brain-streams: fMRI-to-image reconstruction with multi-modal guidance View paper
- [53] Neural encoding and decoding with distributed sentence representations View paper
- [54] NeuroAdapter: Visual Reconstruction with Masked Brain Representation View paper
- [55] Alleviating the semantic gap for generalized fmri-to-image reconstruction View paper
- [56] Generative multimodal decoding: Reconstructing images and text from human fMRI View paper
- [57] MindShot: Multi-Shot Video Reconstruction from fMRI with LLM Decoding View paper
- [58] Coherent Language Reconstruction from Brain Recordings with Flexible Multi-Modal Input Stimuli View paper
- [59] Language models align with brain regions that represent concepts across modalities View paper
- [60] Mind2word: Towards generalized visual neural representations for high-quality video reconstruction View paper

- [61] Llm4brain: Training a large language model for brain video understanding View paper
- [62] Modeling the human visual system: Comparative insights from response-optimized and task-optimized vision models, language models, and different readout â¦ View paper
- [63] Brainchat: Decoding semantic information from fmri using vision-language pretrained models View paper
- [64] BrainChat: Interactive Semantic Information Decoding from fMRI Using Large-Scale Vision-Language Pretrained Models View paper
- [65] Modality-agnostic fmri decoding of vision and language View paper
- [66] Evaluating Text-to-Visual Generation with Image-to-Text Generation View paper
- [67] Learning visual composition through improved semantic guidance View paper
- [68] Attribute-Centric Compositional Text-to-Image Generation View paper
- [69] TextPSG: Panoptic Scene Graph Generation from Textual Descriptions View paper
- [70] Evaluating and Improving Compositional Text-to-Visual Generation View paper
- [71] MARS: Paying More Attention to Visual Attributes for Text-Based Person Search View paper
- [72] Generating semantically precise scene graphs from textual descriptions for improved image retrieval View paper
- [73] Text2scene: Generating compositional scenes from textual descriptions View paper
- [74] Componerf: Text-guided multi-object compositional nerf with editable 3d scene layout View paper
- [75] SceneEval: Evaluating semantic coherence in text-conditioned 3D indoor scene synthesis View paper