Multigranular Evaluation for Brain Visual Decoding

Weihao Xia*, **Cengiz Oztireli**

University of Cambridge
wx258@cam.ac.uk

Abstract

Existing evaluation protocols for brain visual decoding predominantly rely on coarse metrics that obscure inter-model differences, lack neuroscientific foundation, and fail to capture fine-grained visual distinctions. To address these limitations, we introduce BASIC, a unified, multigranular evaluation framework that jointly quantifies structural fidelity, inferential alignment, and contextual coherence between decoded and ground-truth images. For the structural level, we introduce a hierarchical suite of segmentation-based metrics, including foreground, semantic, instance, and component masks, anchored in granularity-aware correspondence across mask structures. For the semantic level, we extract structured scene representations encompassing objects, attributes, and relationships using multimodal large language models, enabling detailed, scalable, and context-rich comparisons with ground-truth stimuli. We benchmark a diverse set of visual decoding methods across multiple stimulus-neuroimaging datasets within this unified evaluation framework. Together, these criteria provide a more discriminative, interpretable, and comprehensive foundation for evaluating brain visual decoding methods.

Code — https://github.com/weihaox/BASIC

Introduction

Recent advances in brain visual decoding (Takagi and Nishimoto 2023a; Ozcelik and VanRullen 2023; Scotti et al. 2023; Xia et al. 2024a) have achieved remarkable success in reconstructing visual stimuli from the neural activations. However, the evaluation protocols commonly used in this field remain limited in several critical aspects. First, current metrics often saturate across state-of-the-art models, limiting their discriminative capacity and obscuring substantive differences in decoded results. Second, these metrics often lack a neuroscientific foundation, failing to capture the perceptual validity of decoded outputs and their alignment with human-like perception. Third, prevailing evaluation strategies typically fail to reflect the multilevel and structured nature of visual perception, neglecting key components such as object semantics, scene understanding, and contextual reasoning. These limitations hinder rigorous benchmarking of brain decoding models and obscure the specific dimensions along which reconstructions succeed or fall short.

What should brain decoding recover. Brain visual decoding aims to reconstruct visual experiences from neural activations, recovering not only the appearance of stimuli but also their structure, semantics, and perceptual salience. A well-decoded reconstruction should reflect what the subject consciously perceived, preserving salient objects, their attributes, spatial configuration, and overall scene coherence. Since human visual perception is shaped by attention, context, and prior knowledge, brain decoding must align with the hierarchical nature of vision, spanning from low-level pixel patterns to high-level semantic understanding. Therefore, effective brain visual decoding requires both perceptual accuracy and semantic integrity. It should faithfully capture salient elements and spatial context in line with the subject's attention, while maintaining consistent inter-object relationships and scene-level coherence.

What should we measure in brain visual decoding. Current evaluation protocols, such as pixel-wise correlation or feature-based similarity, often fall short in capturing the full complexity of brain visual decoding. Low-level metrics tend to overlook scene semantics and perceptual plausibility, while high-level black-box measures conflate multiple alignment aspects into a single score, offering limited diagnostic insight. They struggle to determine whether the "decoded" details truly originates from brain signals, or if they are instead hallucinated constructs based on prototypical co-occurrences from pretrained generative models conditioned on scene or object labels. Moreover, existing metrics are often *saturated*, assigning uniformly high scores across diverse methods and thus failing to capture fine-grained distinctions in reconstruction quality. We argue that an effective brain visual decoding metric should be neuroscientifically interpretable, grounded in principles of human visual perception, and meet three key desiderata. First, it should be *multigranular*, capturing perceptual alignment across multiple abstraction levels – from basic object identification and segmentation to semantic and spatial reasoning about attributes and interactions. Second, it should be *semantically aligned* with human perception, reflecting the way humans interpret scenes, objects, and their relationships in a coherent and meaningful manner. Third, it should be *diagnosti-*

*Corresponding Author.

cally informative, offering interpretable feedback on what is correct, what is missing, and where the reconstruction fails. Specifically, it should localize and characterize semantic and structural errors in decoded outputs, such as object misidentification, incorrect attributes, or implausible interactions.

Our method: BASIC. To bridge these gaps, we introduce BASIC (Brain-Aligned Structural, Inferential, and Contextual similarity), a unified, multigranular evaluation framework for brain visual decoding. BASIC integrates structure matching and semantic reasoning to systematically quantify structural, inferential, and contextual alignment between decoded results and reference stimuli across diverse brain-to-vision tasks in a structured and interpretable manner.

BASIC decomposes evaluation into three complementary perspectives, each reflecting a core aspect of perceptual alignment in brain visual decoding: (a) *structural similarity* quantifies the reconstructed visual structures, capturing spatial organization and categorical boundaries. This consistency with reference stimuli is operationalized through a granularity-aware mask correspondence across the foreground, semantic, instance, and component levels; (b) *inferential similarity* measures semantic accuracy, evaluating whether the decoded image conveys the same entities and conceptual content as the reference. This is computed via structured comparisons across object categories, attributes, and inter-object relational graphs extracted using captions from multimodal large language models; (c) *contextual similarity* assesses perceptual and cognitive plausibility, examining whether the reconstructed scene forms an internally coherent and contextually appropriate whole. This is evaluated using MLLM-based scene reasoning to quantify narrative consistency and global scene coherence.

BASIC offers a comprehensive view of decoding performance across modalities (image, video, 3D) and neuroimaging types (fMRI, EEG). Our framework facilitates both quantitative comparisons and qualitative diagnostics, allowing for fine-grained benchmarking of brain decoding models across datasets. BASIC aims to (1) offer a more detailed and interpretable evaluation of brain decoding results, (2) quantify semantic plausibility in terms more closely aligned with human cognitive processes, and (3) facilitate comparison of decoding methods within a unified evaluation framework applicable across diverse stimulus-neuroimaging datasets. We hope our method contributes to establishing a more systematic foundation for brain visual decoding evaluation.

Related Work

Brain visual decoding. The task of brain visual decoding aims to reconstruct perceived visual stimuli, such as images, videos, or 3D shapes, from recorded neural activations. Recent progress in this domain has been closely tied to advancements in computational modeling frameworks. Early methods primarily involved training neural networks from scratch to learn mappings between brain activity and visual features. However, these models often suffered from limited fidelity and exhibited artifacts and poor visual quality in the reconstructed outputs. Recent developments have led to significant improvements due to the emergence of multimodal generative models (Radford et al. 2021; Rombach et al. 2022; Xu et al. 2023) and large-scale brain-stimulus datasets (Wen et al. 2018; Allen et al. 2022; Gao et al. 2024; Guo et al. 2025). These resources have facilitated a new generation of decoding paradigms that leverage intermediate representations from pretrained generative models and align them with neural responses through various strategies, including linear regression (Ozcelik and VanRullen 2023; Takagi and Nishimoto 2023a), diffusion priors (Scotti et al. 2023, 2024), or feature-wise reconstruction (Xia et al. 2024a; Xia and Öztireli 2025b,a). Recent efforts have also focused on eliminating dependence on subject-specific encoders (Scotti et al. 2024; Wang et al. 2024a; Xia et al. 2024a; Tian et al. 2025; Gong et al. 2025), demonstrating promising progress toward subject-independent brain decoding. Besides the commonly used fMRI-image Natural Scenes Dataset (NSD) (Allen et al. 2022), a growing body of research has explored alternative combinations of neuroimaging modalities and stimuli, such as fMRI-video (Wen et al. 2018), fMRI-3D (Gao et al. 2024), EEG-image (Grootswagers et al. 2022), EEG-video (Liu et al. 2024b), and EEG-3D (Guo et al. 2025) decoding. These explorations broaden the scope of brain decoding and offer new insights into the neural representation of dynamic and immersive visual experiences.

Brain decoding evaluation. The evaluation metrics for visual brain decoding lack a universally accepted standard. Different stimulus-neuroimaging combinations employ varying evaluation protocols. For instance, the following eight metrics are commonly used on NSD: PixCorr, SSIM (Wang et al. 2004), AlexNet (Krizhevsky, Sutskever, and Hinton 2017)-2/5, Inception (Szegedy et al. 2016), CLIP (Radford et al. 2021), EffNet (Tan and Le 2019), and SwAV (Caron et al. 2020). The first four metrics are considered low-level, focusing on perceptual similarity and pixel-wise or structural correspondence. In contrast, the latter four are used to evaluate semantic decoding or high-level representations, emphasizing the overall theme or content of the reconstructed images. For other stimulus-neuroimaging setups, evaluation protocols remain inconsistent. Metrics such as n-way classification accuracy (Guo et al. 2025), CLIP-based Pearson correlation (Gong et al. 2024), mask-matching ratios (Li et al. 2025), and modality-specific evaluation metrics (Gao et al. 2025) have all been reported in the recent literature. However, prior metrics overlook the hierarchical nature of perception and struggle to distinguish models with subtle differences. In contrast, our BASIC framework offers a unified, multigranular evaluation across structural, inferential, and contextual dimensions, enabling more nuanced and interpretable comparisons.

BASIC: Evaluating Brain Visual Decoding

BASIC provides a versatile evaluation framework for brain visual decoding across diverse combinations of visual stimuli and neuroimaging modalities. This section details the evaluation dimensions and the two complementary modules of our BASIC metric: BASIC-H integrates the inferential and contextual dimensions of BASIC into a unified mea-

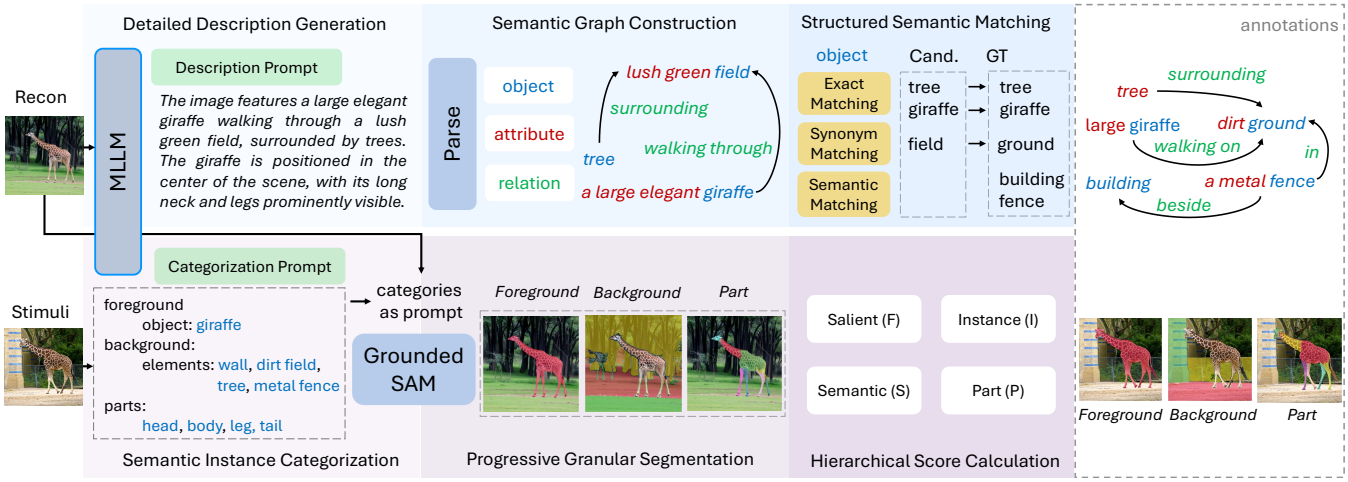


Figure 1: BASIC evaluates decoded reconstructions along two axes: high-level semantic (BASIC-H) and low-level structural (BASIC-L) similarities. For the semantic axis (inferential and contextual), we extract and compare structured representations from reconstructed and ground-truth images. For the structural axis, we compute mask-based matching across fine-grained segmentation types of identified scenes and objects: salient, semantic, instance, and parts.

sure of high-level semantic correspondence, while BASIC-L quantifies low-level structural alignment.

Evaluation Dimension

To gain deeper insight into how brain decoding models represent visual semantics across different levels of abstraction, we develop a structured evaluation protocol grounded in key conceptual dimensions of human visual perception: scenes, objects, attributes, and relations. Each dimension is further divided into subcategories as outlined in Table 1.

Scene reflects global properties, including overall layout, geometric structure, event context, and stylistic tone. This probes the holistic configuration and contextual coherence.

Object evaluates recognition and differentiation of objects in the scene, including both categorical accuracy (e.g., "cat" vs. "dog") and semantic granularity—from object universality (e.g., identifying any "car") to specificity (e.g., distinguishing an "ambulance" from a "sedan").

Attribute covers appearance cues (e.g., color, texture, material) and spatial properties (e.g., location, count). We also extract symbolic visual information such as text using optical character recognition.

Relation assesses interactions between entities, including inter-object and object-scene relationships, as well as spatial and part-whole relations (e.g., "wheel is part of car"), physical interactions (e.g., "man holding umbrella"), and dynamic cues such as posture, expression, and motion.

Camera captures the photographer's perspective and natural conditions during the shot, including viewpoint (e.g., frontal, top-down), motion (e.g., zoom, pan), and lighting.

These dimensions are motivated by principles in visual neuroscience and cognitive psychology (Desimone et al. 1984; Puce et al. 1996), and aligned with the structure of scene understanding in large multimodal models (Dong et al. 2024; Lu et al. 2025; Liu et al. 2023). These dimensions are then organically integrated into structural, inferential,

Dims.	Details
scene	layout, geometry, event, action, style
object	category, universality, differentiation
attribute	appearance, position, quantity, symbols and text
relation	spatial or part-whole interaction, kinematics
camera	lighting, camera angle, camera motion

Table 1: Evaluation dimension and details.

and contextual similarity components within our framework, converted into two sub-indicators, as outlined below.

BASIC-H

BASIC-H quantifies high-level semantic correspondence by integrating inferential and contextual similarities, which respectively capture complementary aspects of conceptual accuracy and narrative coherence, for a holistic evaluation of reconstructed scene semantics. Specifically, BASIC-H operates in an automated pipeline with three steps: (1) detailed description generation, where state-of-the-art MLLMs produce semantic-rich captions for both reconstructed and reference images; (2) semantic graph construction, which parses captions into semantic graphs representing objects, attributes, and relations; and (3) structured semantic matching, which computes semantic correspondace using symbolic concept matching and embedding-based similarities. See Fig. 1 for the method overview.

Detailed description generation. To support multigranular evaluation, we first prompt MLLMs to produce semantically rich descriptions from images. This *description prompt* extracts detailed information across key semantic components corresponding to the semantic dimensions outlined in Table 1, including scene context, objects, object attributes, inter-object relations, and camera viewpoint (when applicable). Recent studies indicate that state-of-the-art MLLMs

Method	Object			Attribute			Relation			BASIC-H
	P	R	F1	P	R	F1	P	R	F1	
NSD (Allen et al. 2022)										
SDRecon (Takagi and Nishimoto 2023a)	55.59	53.05	53.79	10.12	38.73	14.96	40.15	38.71	39.06	35.31
BrainDiffuser (Ozcelik and VanRullen 2023)	57.87	59.11	58.09	13.35	45.82	19.43	43.20	44.42	43.50	39.71
MindEye (Scotti et al. 2023)	62.94	60.64	61.26	18.14	51.17	25.06	49.98	48.42	48.84	44.30
DREAM (Xia et al. 2024b)	**65.63**	63.06	63.56	18.97	50.68	25.92	53.45	**53.21**	52.91	46.37
MindEye2 (Scotti et al. 2024)	62.57	62.12	61.72	17.86	50.16	24.71	49.72	49.17	49.07	44.39
MindBridge (Wang et al. 2024a)	59.00	58.35	58.19	13.70	45.69	19.49	45.75	45.78	45.43	40.16
UMBRAE (Xia et al. 2024a)	62.00	61.86	61.44	17.51	51.32	24.49	48.91	48.71	48.45	44.06
NeuroPictor (Huo et al. 2024)	63.00	61.05	61.38	17.92	50.13	24.66	49.62	49.08	48.98	44.21
NeuroVLA (Shen et al. 2024)	65.36	**65.03**	**64.57**	**21.27**	**53.73**	**28.65**	**53.80**	52.86	**52.95**	**47.88**
SepBrain (Wang et al. 2024b)	62.10	60.19	60.57	16.62	48.71	23.31	48.03	47.55	47.44	43.04
UniBrain (Wang et al. 2024b)	59.03	58.21	58.07	13.27	43.89	19.02	45.55	45.58	45.25	39.89
STTM (Liu et al. 2025)	64.50	62.50	62.88	19.53	51.70	26.64	51.17	50.28	50.36	45.88
MindTuner (Gong et al. 2025)	62.55	62.64	61.95	18.06	49.18	24.73	50.22	50.10	49.80	44.63
BrainGuard (Tian et al. 2025)	63.63	62.36	62.43	18.66	52.40	25.84	51.25	50.72	50.60	45.43
EEG-Things (Grootswagers et al. 2022)										
ATM (Li et al. 2024)	**44.54**	**44.51**	**43.77**	**9.65**	**38.48**	**14.46**	**36.74**	**36.63**	**36.28**	**30.55**
CognitionCapturer (Zhang et al. 2025)	42.20	43.99	42.59	7.56	34.56	11.67	34.31	35.74	34.70	28.64
CC2017 (Wen et al. 2018)										
MinD-Video (Chen, Qing, and Zhou 2023)	47.56	45.05	45.34	6.95	29.27	10.57	32.42	31.10	31.32	28.63
NeuroClips (Gong et al. 2024)	**62.82**	**61.68**	**61.28**	**17.12**	**50.82**	**24.30**	**53.91**	**56.33**	**54.42**	**45.12**
DecoFuse (Li et al. 2025)	47.35	47.12	46.74	6.26	28.12	9.77	32.75	32.18	32.11	29.03
SEED-DV (Liu et al. 2024b)										
EEG2Video (Liu et al. 2024b)	66.05	65.75	64.36	25.43	51.85	32.20	54.63	56.02	54.59	49.54
fMRI-Shape (Gao et al. 2024)										
MinD-3D (Gao et al. 2024)	**41.52**	36.42	37.59	13.52	44.00	19.15	42.49	41.12	41.26	30.95
MinD-3D++ (Gao et al. 2025)	41.35	**37.92**	**38.12**	**20.03**	**61.89**	**27.80**	**44.82**	**43.39**	**43.57**	**35.08**
EEG-3D (Guo et al. 2025)										
Neuro-3D (Guo et al. 2025)	35.79	32.86	34.26	6.32	27.92	10.32	23.41	29.87	26.24	23.08

Table 2: BASIC-H scores across stimulus-neuroimaging datasets. This metric – along with sub-indicators for Precision (P), Recall (R), and F1, each evaluated over Objects, Attributes, and Relations – provides a quantitative assessment of the high-level semantic correspondence between the reconstructions and the original stimuli. **Best** and second best are highlighted.

can generate captions on par with human experts, albeit with occasional hallucinations (Lu et al. 2025; Dong et al. 2024).

Semantic graph construction. Given detailed captions, we first segment them into individual sentences using NLTK (Loper and Bird 2004). We then extract key visual elements – objects, attributes, and relations – using a T5-based factual parser (Li et al. 2023). To handle objects mentioned across multiple sentences in the caption, we consolidate references to the same object into subcaptions and bind attributes to the correct instances, avoiding indiscriminate merging. For directional relations (e.g., subject-object pairs), we concatenate relevant sentences to maintain relational coherence (Lu et al. 2025; Dong et al. 2024).

Structured semantic matching. The matching process aligns the extracted objects, attributes, and relations in three steps: (a) *exact matching*, which identifies direct lexical matches between visual elements from the reconstruction and the reference stimuli; (b) *synonym matching*, which aligns semantically equivalent terms, such as "building" and "edifice" for objects; "bright blue" and "azure" for attributes; "next to" and "beside" for relations; and (c) *semantic matching*, which computes cosine similarity for any remaining unmatched elements based on contextual meaning. This multi-step process ensures that all visual elements, whether directly mentioned or linguistically varied, are appropriately matched across decoded and reference images.

Finally, we compute semantic alignment scores based on the matching results. To mitigate the effects of omissions and hallucinations in MLLM outputs, we evaluate precision, recall, and F1 scores for each type of visual element (object, attribute, relation), based on the matching outcomes. The overall BASIC-H score is computed as a weighted sum of the three F1 scores, with weights of α_1, α_2, α_3 assigned to objects, attributes, and relations, respectively. Precision penalizes hallucinations by measuring the proportion of correctly matched elements among all generated ones, whereas recall captures omissions by evaluating how completely the reference content is recovered. Their combination in F1 score enables a balanced assessment of both over- and under-generation issues common in MLLM outputs.

Method	F (Foreground) IoU	F (Foreground) AP	B (Binary) IoU	B (Binary) AP	S (Semantic) IoU	S (Semantic) AP	I (Instance) IoU	I (Instance) AP	P (Part) IoU	P (Part) AP	BASIC-L
NSD (Allen et al. 2022)											
SDRecon (Takagi and Nishimoto 2023b)	9.03	13.06	38.90	49.97	15.08	21.43	17.84	1.07	4.38	6.79	11.81
BrainDiffuser (Ozcelik and VanRullen 2023)	17.96	20.21	38.98	45.85	18.66	20.78	20.09	1.94	7.86	7.82	16.65
MindEye (Scotti et al. 2023)	19.20	22.79	45.53	55.17	18.73	21.94	20.36	1.99	7.49	8.26	17.03
DREAM (Xia et al. 2024b)	23.62	26.10	46.03	57.10	21.15	24.13	21.41	2.32	9.22	8.79	19.57
MindEye2 (Scotti et al. 2024)	25.29	26.27	47.93	57.52	24.33	25.68	24.09	3.45	12.32	11.03	22.16
MindBridge (Wang et al. 2024a)	16.24	19.04	40.51	48.95	16.87	19.81	18.61	1.54	6.30	6.78	15.00
UMBRAE (Xia et al. 2024a)	21.33	24.62	40.96	48.53	18.94	21.16	20.29	2.15	8.43	8.47	17.89
NeuroPictor (Huo et al. 2024)	**29.45**	**31.29**	47.79	56.48	**27.97**	**29.57**	**26.47**	**4.08**	**17.17**	**15.84**	**25.88**
NeuroVLA (Shen et al. 2024)	16.03	20.85	44.09	54.42	13.84	17.54	17.57	1.34	4.38	5.57	13.54
SepBrain (Wang et al. 2024b)	21.22	23.72	45.06	53.47	20.88	23.32	21.98	2.51	8.79	8.98	18.84
UniBrain (Wang et al. 2024b)	15.24	18.02	37.48	45.09	14.99	17.75	17.45	1.06	5.54	6.07	13.79
STTM (Liu et al. 2025)	27.31	29.44	**49.35**	**59.05**	24.61	26.37	24.19	3.50	12.55	11.65	22.90
MindTuner (Gong et al. 2025)	15.38	16.66	40.74	49.02	21.46	24.50	21.49	1.97	8.16	8.13	16.98
BrainGuard (Tian et al. 2025)	25.90	28.07	49.22	58.99	23.34	25.21	23.98	3.29	10.82	10.32	21.76
EEG-Things (Grootswagers et al. 2022)											
ATM (Li et al. 2024)	**13.87**	**18.77**	**39.46**	**50.14**	**22.85**	**34.69**	**22.02**	**2.15**	11.13	17.07	**17.60**
CognitionCapturer (Zhang et al. 2025)	10.26	12.71	33.31	40.67	21.37	29.84	20.75	1.37	**13.08**	**17.26**	16.22
CC2017 (Wen et al. 2018)											
MinD-Video (Chen, Qing, and Zhou 2023)	12.82	24.20	44.89	53.50	20.29	33.57	19.87	3.53	6.83	**14.69**	15.25
NeuroClips (Gong et al. 2024)	**24.37**	**31.59**	**65.51**	**73.39**	**28.23**	**36.00**	**28.74**	**7.73**	**9.82**	13.14	**23.52**
DecoFuse (Li et al. 2025)	12.32	18.07	49.87	59.46	22.33	34.63	20.87	2.74	4.06	9.80	15.31
SEED-DV (Liu et al. 2024b)											
EEG2Video (Liu et al. 2024b)	27.77	32.82	57.26	69.97	22.93	31.96	23.41	3.51	3.10	7.10	20.54
fMRI-Shape (Gao et al. 2024)											
MinD-3D (Gao et al. 2024)	**5.69**	**8.10**	**5.69**	**8.10**	**19.91**	29.35	**19.38**	1.45	**15.96**	25.30	**14.72**
MinD-3D++ (Gao et al. 2025)	3.80	5.23	3.80	5.23	16.03	**35.83**	17.96	**1.83**	14.94	**36.19**	12.62
EEG-3D (Guo et al. 2025)											
Neuro-3D (Guo et al. 2025)	2.39	3.65	2.39	3.65	13.77	25.92	12.25	1.02	12.33	18.56	9.69

Table 3: BASIC-L scores across datasets. This metric evaluates low-level structural correspondence between reconstructed and reference images at four granularities: foreground saliency, semantic consistency, instance separation, and part-level delineation.

BASIC-L

BASIC-L evaluates the structural correspondence between the reconstructed and reference images across four granularities: foreground saliency, semantic consistency, instance separation, and part-level delineation. The process first extracts key hierarchical visual components and organizes them into natural language expressions in a progressively granular, whole-to-part format using structured prompts. It then applies referring expression comprehension to localize specific objects in the scene based on these expressions. The evaluation follows two steps:

Semantic instance categorization. We prompt MLLMs to generate structured semantic annotations from images. These annotations include object-centric identifications that specify spatial roles (e.g., foreground or background), semantic categories (e.g., "dog", "tree", "airplane"), and part-level component decompositions (e.g., "head", "trunk", "fuselage"). This process incorporates a multi-level decomposition of visual elements, capturing a hierarchical view of the scene components and facilitating further grounded segmentation and multigranular structural alignment.

Progressive granular segmentation. The categories from these structured annotations are then used as prompts for referring expression comprehension methods (Ren et al. 2023; Kirillov et al. 2023), which produces segmentation masks for each identified object and component. The segmentation results are decided by the text and box thresholds, where only the highest-similarity boxes exceeding the box threshold and words with similarity scores above the text threshold are considered as predicted labels (Liu et al. 2024a).

Hierarchical score calculation. We compare predicted masks against those from the reference image across four granularities: salient (foreground categories, F), binary (foreground and background categories, B), semantic (all distinct categories, S), instance (individual instances of all identified object categories, I), and part (subobject components, mostly salient foreground objects, P), computing intersection-over-union (IoU) and average precision (AP).

We aggregate saliency, semantic, instance, and part-level segmentation scores into the overall BASIC-L metric via a weighted sum of respective IoUs using weights of β_1, β_2, β_3, β_4 for each granularity. This scheme prioritizes global layout and object-level coherence while accounting for fine-

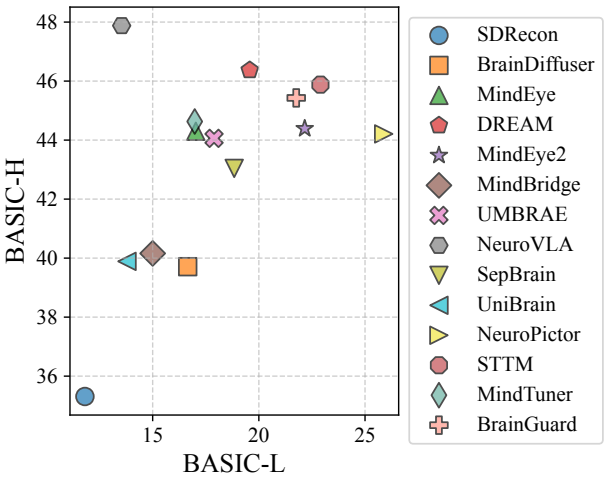


Figure 2: BASIC performance.

grained details, reflecting the limited granularity of information captured during the brain data acquisition experiments.

Experiments

Experimental Setup

Implementation details. We use LLaVA-1.6-13B (Liu et al. 2023) as the default MLLM for detailed description generation and semantic instance categorization. Ground truth captions are obtained from human experts through error correction, missing element addition, and hallucination removal, based on GPT-4o-generated captions, for better alignment with human judgment. For multigranular segmentation, we employ Grounded-SAM2 (Ren et al. 2023), an open-source referring expression comprehension tool. The default weights for BASIC-H and BASIC-L are set to 4:4:2 and 3:2.5:2.5:2, respectively, reflecting the empirically informed importance of each sub-indicator in capturing semantic and structural alignment. Experiments are conducted on an NVIDIA A100 GPU. To ensure broad applicability and methodological consistency, we refrain from using MLLMs or segmentation methods specifically designed for video or 3D data. Instead, we retain an image-based framework with automatic selection of representative video frames and rendered 3D views that best capture key semantics and structure from stimuli and decoded reconstructions.

Experimental Results

While previous metrics suffer from score saturation and fail to effectively distinguish between method performances (see supplementary material), our metrics enable more nuanced evaluation. They offer clearer differentiation across objects, attributes, and relations for high-level semantics, as well as across salient, semantic, instance, and part-level granularity for low-level structure. Detailed descriptions of each dimension are provided in Table 2 and Table 3.

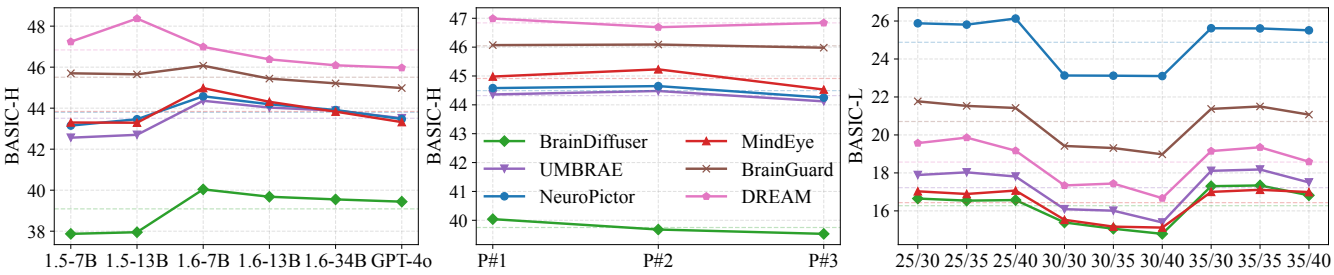
Semantics: object, attribute, and relation. Table 2 presents a comparison of object, attribute, and relation prediction performance across several datasets. While prior metrics often produce saturated or undifferentiated scores across models, our proposed multigranular evaluation reveals meaningful variations in object coherence, attribute accuracy, and relational plausibility, offering a more fine-grained diagnostic lens. Here, for methods on NSD (Allen et al. 2022), NeuroVLA (Shen et al. 2024), DREAM (Xia et al. 2024b), and STTM (Liu et al. 2025) achieve the highest BASIC-H scores, reflecting their strength in modeling rich visual semantics. This likely stems from detailed caption generation, complex visual semantics modeling, cross-subject training, or their combinations. A closer comparison between NeuroVLA and DREAM shows that while both models reconstruct correct instances (achieving higher precision) across the three semantic dimensions, NeuroVLA misses fewer relevant instances (better recall). In contrast, SDRecon (Takagi and Nishimoto 2023b), BrainDiffuser (Ozcelik and VanRullen 2023), and UniBrain (Wang et al. 2024b) lag behind with BASIC-H scores. These methods struggle notably with the identification of object categories, attributes, and relations, indicating limitations in handling semantics or generalizing across visual categories.

Structure: salient, semantic, instance, and part. Table 3 reports performance across salient, binary, semantic, instance, and part segmentation matching scores. For decoding methods on NSD, NeuroPictor (Huo et al. 2024) leads with the highest BASIC-L score, driven by its superior performance in instance and part segmentation. This reflects its ability to reconstruct fine-grained visual details and delineate object boundaries from neural signals. STTM (Liu et al. 2025) and MindEye2 (Scotti et al. 2024) perform competitively. SDRecon (Takagi and Nishimoto 2023b) and NeuroVLA (Shen et al. 2024) show relatively poor spatial structural fidelity. BrainGuard (Tian et al. 2025) achieves competitive scores in salient and semantic categories but performs less satisfactorily in instance and part segmentation. This indicates that while the model captures the overall structure of salient objects and categories, it faces challenges in distinguishing between multiple instances within the same category and identifying sub-component relationships.

Fig. 2 provides performance of visual decoding models on NSD in terms of structural alignment (BASIC-L) and semantic alignment (BASIC-H). Each point represents a model, revealing trade-offs between spatial fidelity and semantic coherence. Models closer to the top-right demonstrate stronger performance across both dimensions. For structural alignment, NeuroPictor (Huo et al. 2024), STTM (Liu et al. 2025), and MindEye2 (Scotti et al. 2024) achieve the highest BASIC-L scores, reflecting better spatial reconstruction. In contrast, semantic alignment (BASIC-H) is best captured by NeuroVLA (Shen et al. 2024), DREAM (Xia et al. 2024b), and STTM (Liu et al. 2025), indicating superior semantic preservation.

Discussion

Toward open, stable, and versatile evaluation. Open: Our evaluation pipeline is designed to be model-agnostic, avoiding reliance on proprietary or task-specific components



(a) MLLM Choice (b) Prompting Strategy (c) Threshold (Box / Text)

Figure 3: BASIC demonstrates stable and consistent performance in method evaluation across variations in (a) MLLMs (Liu et al. 2023), (b) prompting strategies, and (c) thresholds for box and text (Ren et al. 2023).

to ensure broad applicability and reproducibility. While models like GPT-4o are shown to have stronger multimodal performance in captioning benchmarks (Lu et al. 2025; Dong et al. 2024), they are not suitable for open, large-scale benchmarking due to API restrictions and cost. We use LLaVA-1.6-13B, which balances captioning accuracy and computational efficiency. **Stable:** The methods under our metrics demonstrate stable and consistent performance despite variations in (a) MLLMs, (b) prompting strategies, and (c) thresholds for box and text, as shown in Fig. 3. The relative ranking and discriminative power across decoding methods remain consistent, although absolute scores may vary slightly. The stable BASIC-H results across different MLLMs and prompts indicate that our metric reliably captures semantic performance differences among methods. Similarly, the BASIC-L scores provide consistent structural evaluation across varying text and box threshold settings, where only the highest similarity boxes exceeding the box threshold and words with similarity scores above the text threshold are considered predicted labels. **Versatile:** For video and 3D data evaluation, instead of using modality-specific tools, we maintain the same image-based pipeline, with automatic selection of representative video frames and 3D-rendered views to ensure compatibility across formats without compromising evaluation quality.

Toward informative diagnostic insight. This multigranular, interpretable feedback provides fine-grained diagnostic insights into brain-based visual decoding and can help uncover blind spots. Decoding methods that rely on pretrained models benefit from powerful generative capabilities but are also susceptible to systematic biases, often introducing hallucinated details such as prototypical co-occurrence patterns. For instance, *savannah* may be infered when the decoded scene label is *giraffe* even though the original stimulus was a *zoo*. These hallucinations tend to affect high-level contextual information—such as scene types or typical object-environment associations—whereas low-level attributes (e.g., blue) and spatial relationships (e.g., to the left) are less likely to be inferred without support from brain-derived signals. BASIC-H mitigates this confound and allows us to better isolate the contribution of genuine brain-derived information and more accurately assess the true decoding performance. BASIC-H penalizes misidentifications, such as hallucinated objects, incorrect attributes, and implausible relationships. These semantic-level discrepancies are directly reflected in the precision and recall of object categories, attributes, and relations, offering interpretable signals beyond opaque scores from pretrained networks. The segmentation-based BASIC-L scores also serve diagnostic purposes, measuring structure correspondance across four granularities: salient, semantic, instance and part.

Towards cross-dataset performance dissection. Our evaluation follows a unified protocol across visual modalities, enabling cross-dataset dissection. The fMRI-based image reconstructions (Ozcelik and VanRullen 2023; Scotti et al. 2023) consistently outperform EEG-based counterparts (Li et al. 2024; Zhang et al. 2025), attributable to intrinsic limitations of EEG in spatial resolution and information. The fMRI-to-video reconstructions (Gong et al. 2024; Chen, Qing, and Zhou 2023; Li et al. 2025) currently struggles to preserve the structure of salient objects, but performs comparably to image modalities in terms of instance semantic identification and structural preservation. The high performance in EEG-to-video decoding (Liu et al. 2024b) appears to reflect the simplicity of stimuli – short clips with prominent foregrounds and minimal background clutter. Future work explore more semantically rich and visually complex video scenarios. For 3D reconstruction, both fMRI-based and EEG-based visual decoding methods (Gao et al. 2024, 2025; Guo et al. 2025) struggle to recover even basic semantics, despite simple, canonical structures of the target object categories. The incapability in semantic identification and structure preservation highlights the need for decoding models with stronger semantic and geometric priors.

Conclusion

We present a brain-based visual decoding evaluation framework that captures the multigranular nature of human visual perception. By integrating mask-based segmentation alignment with structured object-attribute-relation similarity, our approach enhances performance discriminability, neuroscientific validity, and semantic interpretability. The resulting BASIC metrics provide a comprehensive assessment across structural precision, inferential accuracy, and contextual coherence. We benchmark diverse brain visual decoding models across six major stimulus-neuroimaging datasets under a unified evaluation protocol, establishing the first standardized, interpretable, and extensible benchmark for this task.

Acknowledgements

This work was supported by a UKRI Future Leaders Fellowship [grant number G104084].

References

Allen, E. J.; St-Yves, G.; Wu, Y.; Breedlove, J. L.; Prince, J. S.; Dowdle, L. T.; Nau, M.; Caron, B.; Pestilli, F.; Charest, I.; et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1): 116–126.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, volume 33, 9912–9924.

Chen, Z.; Qing, J.; and Zhou, J. H. 2023. Cinematic mindscapes: High-quality video reconstruction from brain activity. In *NeurIPS*, 24841–24858.

Desimone, R.; Albright, T. D.; Gross, C. G.; and Bruce, C. 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8): 2051–2062.

Dong, H.; Li, J.; Wu, B.; Wang, J.; Zhang, Y.; and Guo, H. 2024. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*.

Gao, J.; Fu, Y.; Wang, Y.; Qian, X.; Feng, J.; and Fu, Y. 2024. Mind-3d: Reconstruct high-quality 3d objects in human brain. In *ECCV*, 312–329.

Gao, J.; Fu, Y.; Wang, Y.; Qian, X.; Feng, J.; and Fu, Y. 2025. MinD-3D++: Advancing fMRI-Based 3D Reconstruction with High-Quality Textured Mesh Generation and a Comprehensive Dataset. *TPAMI*, 47(12): 11802–11816.

Gong, Z.; Bao, G.; Zhang, Q.; Wan, Z.; Miao, D.; Wang, S.; Zhu, L.; Wang, C.; Xu, R.; Hu, L.; Liu, K.; and Zhang, Y. 2024. NeuroClips: Towards high-fidelity and smooth fMRI-to-video reconstruction. In *NeurIPS*, 51655–51683.

Gong, Z.; Zhang, Q.; Bao, G.; Zhu, L.; Xu, R.; Liu, K.; Hu, L.; and Miao, D. 2025. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *AAAI*, 14247–14255.

Grootswagers, T.; Zhou, I.; Robinson, A. K.; Hebart, M. N.; and Carlson, T. A. 2022. Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1): 3.

Guo, Z.; Wu, J.; Song, Y.; Bu, J.; Mai, W.; Zheng, Q.; Ouyang, W.; and Song, C. 2025. Neuro-3D: Towards 3D visual decoding from EEG signals. In *CVPR*, 23870–23880.

Huo, J.; Wang, Y.; Wang, Y.; Qian, X.; Li, C.; Fu, Y.; and Feng, J. 2024. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *ECCV*, 56–73.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment anything. In *ICCV*, 4015–4026.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

Li, C.; Huo, J.; Gong, W.; Fu, Y.; Xue, X.; and Feng, J. 2025. DecoFuse: Decomposing and Fusing the" What"," Where", and" How" for Brain-Inspired fMRI-to-Video Decoding. *arXiv preprint arXiv:2504.00432*.

Li, D.; Wei, C.; Li, S.; Zou, J.; Qin, H.; and Liu, Q. 2024. Visual decoding and reconstruction via eeg embeddings with guided diffusion. In *NeurIPS*, 102822–102864.

Li, Z.; Chai, Y.; Zhuo, T. Y.; Qu, L.; Haffari, G.; Li, F.; Ji, D.; and Tran, Q. H. 2023. Factual: A benchmark for faithful and consistent textual scene graph parsing. In *ACL Findings*, 6377–6390.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. In *NeurIPS*, 34892–34916.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 38–55.

Liu, X.-H.; Liu, Y.-K.; Wang, Y.; Ren, K.; Shi, H.; Wang, Z.; Li, D.; Lu, B.-L.; and Zheng, W.-L. 2024b. EEG2video: Towards decoding dynamic visual perception from EEG signals. In *NeurIPS*, 72245–72273.

Liu, Y.; Ma, Y.; Zhu, G.; Jing, H.; and Zheng, N. 2025. See through their minds: Learning transferable neural representation from cross-subject fMRI. In *AAAI*, 5730–5738.

Loper, E.; and Bird, S. 2004. Nltk: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*.

Lu, F.; Wu, W.; Zheng, K.; Ma, S.; Gong, B.; Liu, J.; Zhai, W.; Cao, Y.; Shen, Y.; and Zha, Z.-J. 2025. Benchmarking Large Vision-Language Models via Directed Scene Graph for Comprehensive Image Captioning. In *CVPR*, 19618–19627.

Ozcelik, F.; and VanRullen, R. 2023. Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1): 15666.

Puce, A.; Allison, T.; Asgari, M.; Gore, J. C.; and McCarthy, G. 1996. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of neuroscience*, 16(16): 5205–5215.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2023. Grounded sam: Assembling open-world models for diverse visual tasks. In *ICCV Demo Track*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Scotti, P. S.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Cohen, E.; Dempster, A. J.; Verlinde, N.; Yundler, E.;

Weisberg, D.; et al. 2023. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. In *NeurIPS*, 24705–24728.

Scotti, P. S.; Tripathy, M.; Villanueva, C. K. T.; Kneeland, R.; Chen, T.; Narang, A.; Santhirasegaran, C.; Xu, J.; Naselaris, T.; Norman, K. A.; and Abraham, T. M. 2024. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *ICML*, 44038–44059.

Shen, G.; Zhao, D.; He, X.; Feng, L.; Dong, Y.; Wang, J.; Zhang, Q.; and Zeng, Y. 2024. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. In *NeurIPS*, 98083–98110.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Takagi, Y.; and Nishimoto, S. 2023a. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, 14453–14463.

Takagi, Y.; and Nishimoto, S. 2023b. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. *arXiv preprint arXiv:2306.11536*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 6105–6114.

Tian, Z.; Quan, R.; Ma, F.; Zhan, K.; and Yang, Y. 2025. BrainGuard: Privacy-Preserving Multisubject Image Reconstructions from Brain Activities. In *AAAI*, 14414–14422.

Wang, S.; Liu, S.; Tan, Z.; and Wang, X. 2024a. Mindbridge: A cross-subject brain decoding framework. In *CVPR*, 11333–11342.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4): 600–612.

Wang, Z.; Zhao, Z.; Zhou, L.; and Nachev, P. 2024b. UniBrain: A Unified Model for Cross-Subject Brain Decoding. *arXiv preprint arXiv:2412.19487*.

Wen, H.; Shi, J.; Zhang, Y.; Lu, K.-H.; Cao, J.; and Liu, Z. 2018. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12): 4136–4160.

Xia, W.; de Charette, R.; Oztireli, C.; and Xue, J.-H. 2024a. UMBRAE: Unified Multimodal Brain Decoding. In *ECCV*, 242–259.

Xia, W.; de Charette, R.; Öztireli, C.; and Xue, J.-H. 2024b. DREAM: Visual Decoding from Reversing Human Visual System. In *WACV*, 8226–8235.

Xia, W.; and Öztireli, C. 2025a. Exploring The Visual Feature Space for Multimodal Neural Decoding. In *ICCV*, 4370–4379.

Xia, W.; and Öztireli, C. 2025b. MEVOX: Multi-Task Vision Experts for Brain Captioning. In *CVPRW*.

Xu, X.; Wang, Z.; Zhang, E.; Wang, K.; and Shi, H. 2023. Versatile diffusion: Text, images and variations all in one diffusion model. In *ICCV*, 7754–7765.

Zhang, K.; He, L.; Jiang, X.; Lu, W.; Wang, D.; and Gao, X. 2025. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *AAAI*, 14486–14493.