LEARNING BRAIN REPRESENTATION WITH HIERACHICAL VISUAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding visual representations from brain signals has attracted significant attention in both neuroscience and artificial intelligence. However, the extent to which EEG signals contain actual visual information remains unclear. Current visual decoding approaches explore various EEG–image alignment strategies, yet most emphasize high-level semantic features while neglecting pixel-level details, thereby limiting our understanding of the human visual system. In this paper, we propose an EEG-image alignment strategy that leverages multiple pre-trained visual encoders with distinct inductive biases to capture hierarchical and multi-scale visual representations, while employing a contrastive learning objective to achieve effective alignment between EEG and visual embeddings. Furthermore, we introduce a Fusion Prior, which learns a stable mapping on large-scale visual data and subsequently matches EEG features to this pre-trained prior, thereby enhancing distributional consistency across modalities. Both quantitative and qualitative experiments demonstrate that our method achieves a strong balance between retrieval and reconstruction capabilities.

1 INTRODUCTION

With the rapid development of text-to-image generative models (Rombach et al., 2022; Zhang et al., 2023; Esser et al., 2024), reconstructing human visual stimuli from brain signals has become a prominent research focus in both neuroscience and artificial intelligence. Visual processing is a core function of the human brain. When visual stimuli is processed by the brain, the primary visual cortex initially deciphers basic pixel attributes such as color, edges, and textures, subsequently forwarding them to various higher-order visual cortices for further hierarchical processing (Blasdel & Lund, 1983; Tsumoto et al., 1978). These higher-level regions collaborate to synthesize and generalize visual data, resulting in semantic characteristics such as objects and environments, and thus formulating the essential processes underlying human visual perception of the external world. (Merigan & Maunsell, 1993).

To investigate these complex and dynamic relationships between the human visual system and brain representations, researchers commonly employ Functional magnetic resonance imaging (fMRI), Magnetoencephalography (MEG), and Electroencephalogram (EEG) for visual decoding and reconstruction (Zhang et al., 2025; Benchetrit et al., 2023). fMRI measures brain activity indirectly through blood-oxygen-level-dependent signals, offering high spatial resolution but limited temporal resolution, which makes it difficult to capture rapid neural dynamics (Logothetis et al., 2001). In contrast, EEG and MEG directly reflect the brain's electrophysiological activity. EEG provides high temporal resolution but suffers from low spatial resolution and a poor signal-to-noise ratio. MEG, while also offering millisecond-level temporal precision, provides comparatively better spatial resolution (Liu et al., 2023a; da Silva, 2013).

Previous research has explored decoding brain signals by aligning them with visual representations, enabling classification, retrieval, and reconstruction. Song et al. (2023) employed contrastive learning to maximize the similarity of matched brain–image pairs while minimizing that of mismatched ones. Li et al. (2024) proposed the Adaptive Thinking Mapper (ATM) to align brain signal features with CLIP-derived visual embeddings, combined with a two-stage multi-pipe strategy for brain-to-image generation. However, these approaches rely on direct alignment between brain signals and

image features, whereas the structural gap between the two modalities makes this strategy insufficient to capture the underlying shared representations.

Recently, several studies have attempted to improve direct alignment by introducing priors or enriching visual representations. Wu et al. (2025) introduced the Uncertainty-aware Blur Prior (UBP), which mitigates brain–image mismatches by blurring high-frequency image details. Zhang et al. (2025) extended CLIP-derived image embeddings with depth information to enhance brain–image alignment. However, these methods focus primarily on high-level semantic alignment while overlooking low-level pixel information. This oversight prevents a comprehensive understanding of the visual content encoded in brain signals and reduces interpretability.

To bridge the structural gap between the temporal dynamics of brain signals and the spatial hierarchies of images, we introduce Hierarchical Visual Fusion with Fusion Prior, a framework inspired by perceptual mechanisms of the human visual system. The framework integrates multiple pre-trained encoders to construct multiscale visual representations, ranging from pixel-level details to high-level semantics, and leverages contrastive learning to align brain and visual features. To address the limitations of CLIP and related encoders in capturing local and fine-grained information, we incorporate low-level visual features modeled by a Variational Autoencoder (VAE) into the fused representation. In addition, we pretrain a Fusion Prior on large-scale visual data to provide a stable mapping from fused features to diffusion conditions, which substantially improves retrieval accuracy, reconstruction fidelity, and interpretability. The key contributions of this work are as follows:

- We incorporate low-level visual information from a VAE upon semantic alignment, compensating for the limitations of CLIP-based encoders in modeling pixel-level details.
- We propose a Fusion Prior that learns a robust visual representation from large-scale data, providing a stable bridge for aligning brain signals to improve cross-modal consistency.
- Our method achieves the state-of-the-art performance in retrieval tasks with significant advancements over prior work, while delivering superior reconstruction quality.

2 Related Work

Brain Visual Decoding Neural decoding aims to infer human cognitive and perceptual states from brain signals such as EEG (Bai et al., 2023; Li et al., 2024), MEG (Cichy et al., 2016b), or fMRI (Kay et al., 2008; Takagi & Nishimoto, 2023b). Among these, visual decoding has become a particularly challenging and promising direction, mainly including tasks like image classification (Xu et al., 2024) , retrieval (Liu et al., 2023c) and reconstruction (Ozcelik & VanRullen, 2023; Takagi & Nishimoto, 2023a). A central focus has been on encoding EEG signals into effective representation vectors that capture temporal and frequency characteristics (Fu et al., 2025). To bridge the modality gap, CLIP-based models are commonly adopted as benchmarks (Liu et al., 2023c; Wang et al., 2024), while recent methods have explored strategies such as diffusion priors (Aggarwal et al., 2023) for enhancing semantic consistency in the generative space (Ozcelik & VanRullen, 2023; Takagi & Nishimoto, 2023b; Li et al., 2025), or bidirectional mappings to enforce cross-modal cycle consistency (Wei et al., 2024). At the same time, research on the image modality itself has explored ways to complement the limited semantic expressiveness of neural signals, including blurred preprocessing to suppress high-frequency noise (Li et al., 2024) and textual descriptions to enrich semantic guidance (Takagi & Nishimoto, 2023b). However, most existing approaches primarily emphasize high-level semantics without sufficiently capturing pixel-level, fine-grained representations, leaving notable gaps in the fidelity of generated or reconstructed images.

Hierarchical and Multiscale Visual Representations Recent advances in image-only representation learning emphasize multi-level semantics and dense structure within a single modality. Vision Transformers (Dosovitskiy et al., 2020) trained with self-supervision (e.g., token-level pretext objectives) yield strong global semantics without textual supervision (Bao et al., 2021; Xie et al., 2022; He et al., 2022; Caron et al., 2021; Oquab et al., 2023), while generative latent models such as VAEs and VQ-VAEs provide compact pixel-level codes with high reconstruction fidelity (Kingma & Welling, 2013; Higgins et al., 2017; Van Den Oord et al., 2017; Razavi et al., 2019). A complementary perspective from neuroscience links deeper network features to higher visual areas and early layers to fine spatial detail and rapid dynamics (Yamins et al., 2014; Cichy et al., 2016a), motivating the combination of coarse semantic abstractions with fine-grained local cues. In practice, however, many

decoding pipelines (Li et al., 2024; Wu et al., 2025; Zhang et al., 2025) instantiate a single semantic embedding space for simplicity and zero-shot transfer, which may underweight local structures that are important for faithful image reconstruction from neural signals.

Cross-modal Contrastive Learning Cross-modal contrastive learning (CMCL) aligns heterogeneous inputs within a shared embedding space by maximizing agreement between matched pairs under a temperature-scaled InfoNCE objective (Oord et al., 2018; Wu et al., 2018; He et al., 2020). Bi-encoder formulations with cosine similarity and (often) symmetric losses have become the default recipe for scalable pretraining and zero-shot transfer (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022; Li et al., 2022). Building on this recipe, large-scale vision–language systems such as CLIP/ALIGN demonstrate strong generalization across retrieval and classification benchmarks, and the paradigm extends beyond image–text to audio–visual (Arandjelovic & Zisserman, 2017; Morgado et al., 2021; Wu et al., 2022), video–language (Miech et al., 2019; 2020; Xu et al., 2021; Bain et al., 2021; Luo et al., 2022), and 3D–language (Xue et al., 2023; Zhang et al., 2022; Liu et al., 2023b) alignment. Despite this progress, CMCL typically assumes accurately paired data. At web scale, weak captions, temporal asynchrony, and domain shift impair alignment quality, motivating data curation, caption bootstrapping, and bridging/distillation strategies (Jia et al., 2021; Miech et al., 2019). Recent analyses (Liang et al., 2022; Wang et al., 2023) also reveal a modality gap between modalities in the shared space, which can complicate fine-grained alignment. In neural decoding, the brain–vision pairing is intrinsically scarce and noisy (limited trials, low SNR, trial-to-trial latency variability), so naively aligning brain signals to a single semantic-only space risks under-representing pixel-level structure and amplifying modality mismatch (Cichy et al., 2016a).

Adapters for Image Diffusion Models Adapters have emerged as parameter-efficient modules that extend pretrained diffusion models with controllability and editing while largely freezing base weights, offering a unifying recipe across tasks and modalities (Wang et al., 2025). T2I-Adapter (Mou et al., 2024) learns lightweight branches that align external control signals (e.g., edges, depth, sketches) with internal features of a frozen text-to-image model, enabling accurate and composable multi-condition control. ControlNet (Zhang et al., 2023) freezes the pretrained backbone and adds zero-initialized side networks to inject spatial conditions without destabilizing the original prior. IP-Adapter (Ye et al., 2023) decouples cross-attention to integrate image prompts alongside text, delivering strong multimodal conditioning with 22M trainable parameters while keeping the diffusion backbone frozen.

3 METHOD

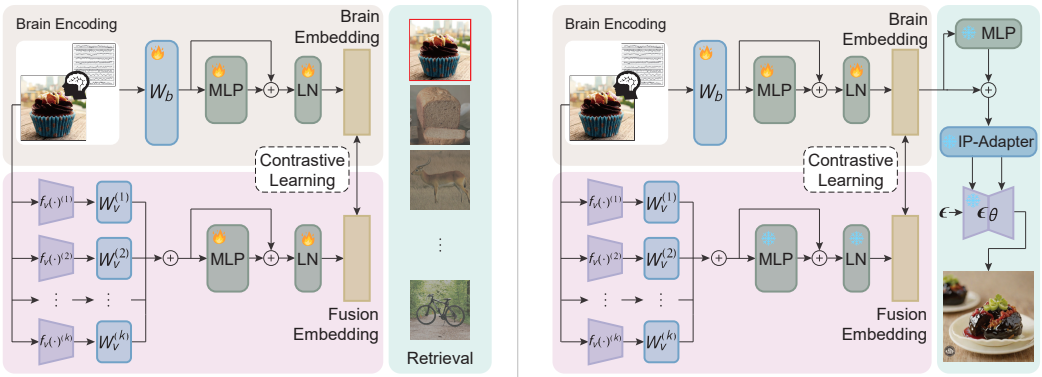
3.1 PROBLEM STATEMENT

The goal of visual decoding is to retrieve or reconstruct the visual information corresponding to recorded brain signals. We denote paired brain signals and visual images as $(x_v, x_b) \in \mathcal{D}$, where $x_v \in \mathbb{R}^{H\times W\times 3}$ represents the visual stimulus, with H and W denoting the image height and width, respectively. $x_b \in \mathbb{R}^{C\times T}$ represents the brain signals recorded under the same stimulus, where C corresponds to the number of electrode channels and T indicates the length of time samples.

3.2 ALIGNING BRAIN SIGNALS WITH HIERARCHICAL VISUAL REPRESENTATIONS

Directly aligning brain signals with visual representations may fail to capture the intrinsic multiscale nature of the visual information, thereby limiting alignment performance. While high-level semantic features in the visual modality are crucial for category recognition and abstract understanding, low-level features provide complementary structural information and pixel-level details, which are indispensable for improving reconstruction quality. Inspired by this visual perception mechanism (Blasdel & Lund, 1983; Tsumoto et al., 1978), we integrate multiple pretrained visual encoders to separately extract high-level semantic features and low-level pixel features, and align them with brain signal embeddings through a contrastive learning objective to construct a unified hierarchical visual representation.

Hierarchical visual representations As depicted in Fig. 1-(a), we devise a multi-head encoder structure to obtain hierarchical visual representations ranging from high-level visual semantics (e.g., objects, scenes, and relations) to low-level visual features (e.g., colors, textures, and layouts). We



(a) Brain-to-image retrieval.

(b) Brain-to-image reconstruction.

Figure 1: **Learning pipelines**. Left: Retrieval objective that aligns the brain embedding z_b with the fused visual embedding z_f (HVF over K pretrained encoders) using a symmetric InfoNCE; evaluation is nearest-neighbor retrieval in the fused space. Right: Reconstruction pipeline with a frozen, pretrained fusion prior—HVF plus a Conditioning Adapter (MLP projector + IP-Adapter with decoupled cross-attention). We contrastively align z_b to the frozen z_f, project to z_c, and inject z_c into a frozen SDXL UNet to synthesize the image. Visual encoders and the UNet are frozen; only the brain side is updated during alignment.

apply K pretrained encoders (K=3 by default) to the image x_v, yielding $z_v^{(k)} = f_v^{(k)}(x_v)$ for k=$1, \ldots, K$. For high-level visual semantics, we integrate multiple CLIP encoders and use a single global token from each (i.e., [CLS] token for ViT-based models and the pooled projection for ResNet-based models). For low-level visual features, the VAE encoder outputs a latent of shape $[H/8, W/8, 4]$, which we flatten into a vector of length $(H/8)(W/8) \times 4 = HW/16$, preserving local structure and visual detail.

We fuse features with a post-norm residual Hierarchical Visual Fuser (HVF). For each encoder, a learned linear map $W_v^{(k)} \in \mathbb{R}^{d_k \times d}$ aligns the embedding to the shared dimension d=1024:

$$\bar{z}_v = \sum_{k=1}^{K} z_v^{(k)} W_v^{(k)}, \tag{1}$$

The aligned features are fused with a residual Multi-Layer Perceptron (MLP), and we have

$$z_f = \mathrm{ResBlock}(\bar{z}_v) = \mathrm{LayerNorm}\big(\bar{z}_v + \phi_v(\bar{z}_v)\big), \tag{2}$$

where ϕ_v denotes a two-layer MLP with hidden size $d_v = 1024$ and GELU activation.

Contrastive learning objective For the brain modality, we adopt an MLP-based Brain Projection (MBP) network that projects the EEG signal to an embedding. We first align the preprocessed signal to the visual embedding width using a learned linear projection $W_b \in \mathbb{R}^{CT \times d}$. We then reuse the same architecture as Eq. (2) to produce a d-dimensional embedding compatible with z_f with a hidden size of d that

$$\bar{z}_b = x_b W_b, \qquad z_b = \mathrm{ResBlock}(\bar{z}_b) = \mathrm{LayerNorm}\big(\bar{z}_b + \phi_b(\bar{z}_b)\big), \tag{3}$$

where ϕ_b denotes a two-layer MLP with hidden size $d_b = 1024$ and GELU activation.

We employ a CLIP-style InfoNCE loss (Oord et al., 2018) to align brain and visual embeddings. Given N paired samples, we compute cosine-similarity logits with a trainable temperature τ:

$$\hat{z}_b^{(i)} = \frac{z_b^{(i)}}{\|z_b^{(i)}\|_2}, \quad \hat{z}_f^{(i)} = \frac{z_f^{(i)}}{\|z_f^{(i)}\|_2}, \quad s_{ij} = \frac{\hat{z}_b^{(i)\top} \hat{z}_f^{(j)}}{\tau}, \tag{4}$$

where $\|\cdot\|_2$ is L2 norm. The learning objective is defined as:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{2N}\left(\sum_{i=1}^{N} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{N} \exp(s_{ij})} + \sum_{i=1}^{N} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{N} \exp(s_{ji})}\right), \tag{5}$$

3.3 PRETRAINED FUSION PRIOR FOR RECONSTRUCTION

While the above contrastive learning aligns brain signals with hierarchical visual representations, directly feeding these fused representations into a pretrained diffusion model for reconstruction often results in unstable outputs. The core issue is the absence of a stable conditioning prior: brain-driven features do not yet match the distribution expected by the generative model, leading to noisy or misaligned guidance. To address this, we introduce the fusion prior to learn a robust mapping from fused visual features to diffusion conditions.

Fusion prior pretraining As depicted in Fig. 1-(b), we first feed the fused visual representation z_f from the HVF into an additional projector to obtain z_c:

$$z_c = z_f + \phi_c(z_f), \tag{6}$$

where $z_f, z_c \in \mathbb{R}^d$ and both ϕ_v in Eq. 2 and ϕ_c denotes a two-layer MLP with hidden size $d_c = 4096$ and GELU activation. The IP-Adapter (Ye et al., 2023) then injects z_c into a frozen SDXL (Podell et al., 2023) UNet via cross-attention. Given noisy latent x_t at timestep t, the whole network δ is trained to predict the noise ϵ with

$$\mathcal{L}_{\text{prior}} = \| \epsilon - \delta(x_t, t, z_c) \|_2^2, \tag{7}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is the diffusion target and $\mathcal{L}_{\text{prior}}$ is the loss function.

During pretraining on large-scale visual data, the UNet backbone remains frozen, while the HVF and the projector are trained from scratch, the IP-Adapter is initialized from pretrained weights[1] to accelerate convergence. Text prompts are left empty, ensuring the model learns a text-free mapping from fused visual features to diffusion conditions.

Brain-to-fusion alignment Once the HVF is pretrained, we freeze it and update only the brain encoder using the same loss function $\mathcal{L}_{\text{contrastive}}$ as in Eq. (5), which ensures that brain-derived embeddings are projected into a stable, pretrained fusion space. This prevents representational drift and yields robust reconstruction when passed to the diffusion model.

Full pipeline for reconstruction In all, training uses two stages and inference one. (i) *Prior pretraining:* for input images x_v, extract $\{z_v^{(k)}\}_{k=1}^K$, fuse and project them via the HVF and projector to obtain z_c, and train the IP-Adapter jointly with the HVF and projector (UNet frozen) by minimizing $\mathcal{L}_{\text{prior}}$ in Eq. (7) under empty text prompts, yielding a stable, text-free fusion prior. (ii) *Brain–fusion alignment:* freeze the pretrained fusion prior (HVF, projector and IP-Adapter) and the UNet, and update only the brain side (i.e., the MBP module only) on paired (x_b, x_v) with the symmetric InfoNCE loss in Eq. (5) so that z_b lies in the fusion space of z_f. (iii) *Reconstruction:* given test brain signals x_b, compute $z_b = f_b(x_b)$, feed it to the projector to obtain z_c, and use z_c as the sole condition for the frozen IP-Adapter/UNet; a standard diffusion sampler(SDXL uses an Euler–ancestral sampler (Karras et al., 2022)) then produces $\hat{x}_v$, yielding stable and semantically faithful reconstructions.

4 EXPERIMENT

4.1 EXPERIMENTAL DETAILS

We train the contrastive stage on a single NVIDIA 5090 32GB GPU for 25 epochs with a global batch size of 1024. We use AdamW with a peak learning rate of 5×10^{-4} under a cosine decay schedule and a 10-step warmup from zero. Unless otherwise stated, retrieval uses a fixed encoder set comprising OpenAI CLIP RN50, LAION CLIP ViT-B/32 (Schuhmann et al., 2022), and an SDXL VAE; each backbone follows its canonical preprocessing. The VAE supports multiple input resolutions and defaults to 128×128. The temperature τ is initialized to 0.07. For generation, we swap RN50 for LAION CLIP ViT-H/14, freeze the pretrained HVF on the visual side, and train only the MBM module of the brain modality.

[1] https://huggingface.co/h94/IP-Adapter/resolve/main/sdxl_models/ip-adapter_sdxl_vit-h.safetensors

Table 1: Average Top-1 / Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-**EEG** and THINGS-**MEG**. All numbers are subject-wise averages; "–" indicates not reported.

Method	EEG				MEG			
	Intra-subject		Inter-subject		Intra-subject		Inter-subject	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
BraVL	5.8	17.5	1.8	7.0	–	–	–	–
NICE	16.1	43.6	6.2	21.4	12.8	36.0	–	–
NICE-SA	14.7	41.7	7.0	23.1	12.7	35.0	–	–
NICE-GA	15.6	42.8	5.9	21.6	14.3	42.3	–	–
MB2C	28.5	60.4	–	–	–	–	–	–
ATM	28.5	60.4	11.8	33.7	–	–	–	–
VE-SDN	37.2	69.9	–	–	–	–	–	–
CC-All	35.6	80.2	–	–	–	–	–	–
UBP	50.9	79.7	12.4	33.4	26.7	55.2	2.2	10.4
Ours	**76.1**	**94.2**	**20.0**	**44.1**	**33.7**	**60.5**	**5.4**	**15.2**

The contrastive stage is trained on THINGS-EEG (Grootswagers et al., 2019; Gifford et al., 2022b) and THINGS-MEG (Hebart et al., 2023). For THINGS-EEG (10 participants), the training split contains 1654 concepts with 10 images per concept and 4 repetitions per image; the test split contains 200 concepts with 1 image per concept and 80 repetitions per image. We follow prior work(Li et al., 2024; Wu et al., 2025) to select 17 occipito–parietal channels (O+P) and standard preprocessing (Song et al., 2023). For THINGS-MEG (4 participants, 271 channels), the training split consists of $1854\times12\times1$ (concepts × images × reps) and the test split $200\times1\times12$. To improve signal-to-noise ratio(SNR), repetitions for the same stimulus are averaged within subject in both datasets (training and test). Additional details are provided in the appendix.

For fusion-prior pretraining, we explore multiple prior configurations. Unless noted, training uses two NVIDIA 5090 32 GB GPUs, a fixed learning rate of 1×10^{-4}, SDXL-base as the diffusion backbone, and the largest feasible batch size of 12 per GPU. Each configuration is trained at 512×512 for 100k steps, about two epochs, and takes roughly 15 hours per prior configuration. Pretraining uses ImageNet-1k with about 1.3M images. For reconstruction at inference we use SDXL-Turbo with a 4-step sampler for fast evaluation.

4.2 Quantitative Evaluation

We evaluate two tasks, brain–visual retrieval and brain–visual reconstruction. For retrieval, we report 200-way zero-shot top-1 and top-5 accuracy on THINGS-EEG and THINGS-MEG under both intra-subject and inter-subject protocols. For reconstruction, following prior work (Ozcelik & VanRullen, 2023; Benchetrit et al., 2023; Li et al., 2024), we measure low-level fidelity with PixCorr and SSIM and adopt the remaining semantic and feature-level metrics from these works, including AlexNet(2/5), Inception, CLIP and SwAV distance, where lower is better.

The retrieval baselines are BraVL (Du et al., 2023), NICE and its spatial variants (NICE-SA, NICE-GA) (Song et al., 2023), ATM (Li et al., 2024), VE-SDN (Chen et al., 2024), MB2C (Wei et al., 2024), UBP (Wu et al., 2025), and CognitionCapturer (C.C., All/Image/Depth/Text) (Zhang et al., 2025). For reconstruction, we compare with ATM (Li et al., 2024), CognitionCapturer (Zhang et al., 2025), and Brain Decoding (B.D.) (Benchetrit et al., 2023). When prior work reports single-subject results only (e.g., ATM on subj-8), we indicate this in the tables.

Compared to the strongest prior work (UBP), our model consistently improves 200-way zero-shot retrieval across all protocols (Top-1/Top-5): EEG intra 76.1/94.2 vs 50.9/79.7, EEG inter 20.0/44.1 vs 12.4/33.4, MEG intra 33.7/60.5 vs 26.7/55.2, and MEG inter 5.4/15.2 vs 2.2/10.4 (Tab. 1). Gains are largest in the inter-subject setting, indicating stronger cross-participant generalization.

Table 2 summarizes reconstruction. On MEG our model matches or exceeds prior work on both low-level similarity and semantic alignment while maintaining a competitive SwAV distance. On EEG it improves the commonly reported subj-8 case and delivers clear subject-averaged gains over ATM and C.C. The average EEG PixCorr increases from 0.150 with C.C.(All) to 0.186 with our model

Table 2: Quantitative assessments of the reconstruction quality for EEG and MEG.

Method	Dataset	Low-level		High-level				
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
MEG	B.D.	0.076	0.336	0.736	0.826	0.671	0.767	**0.584**
	ATM	0.104	**0.340**	0.613	0.672	0.619	0.603	0.651
	Ours	**0.137**	0.292	**0.737**	**0.836**	**0.721**	**0.775**	0.600
EEG	C.C.(All)	0.150	0.347	0.754	0.623	0.669	0.715	0.590
	C.C.(Image)	0.132	0.321	0.813	0.671	0.664	0.715	0.590
	C.C.(Depth)	0.104	**0.370**	0.796	0.638	0.565	0.579	0.686
	C.C.(Text)	0.102	0.288	0.727	0.582	0.586	0.598	0.673
	Ours	**0.195**	0.336	**0.843**	**0.905**	**0.756**	**0.808**	**0.554**
EEG (subj-8)	ATM	0.160	0.345	0.776	0.866	0.734	0.786	0.582
	Ours	**0.227**	**0.361**	**0.878**	**0.924**	**0.796**	**0.826**	**0.531**

while SSIM remains comparable, and semantic similarities improve across AlexNet, Inception, and CLIP with a lower SwAV distance than C.C. and B.D. Taken together, the metrics indicate that our approach raises both fidelity and semantic agreement and that the improvements persist beyond single-subject evaluation.

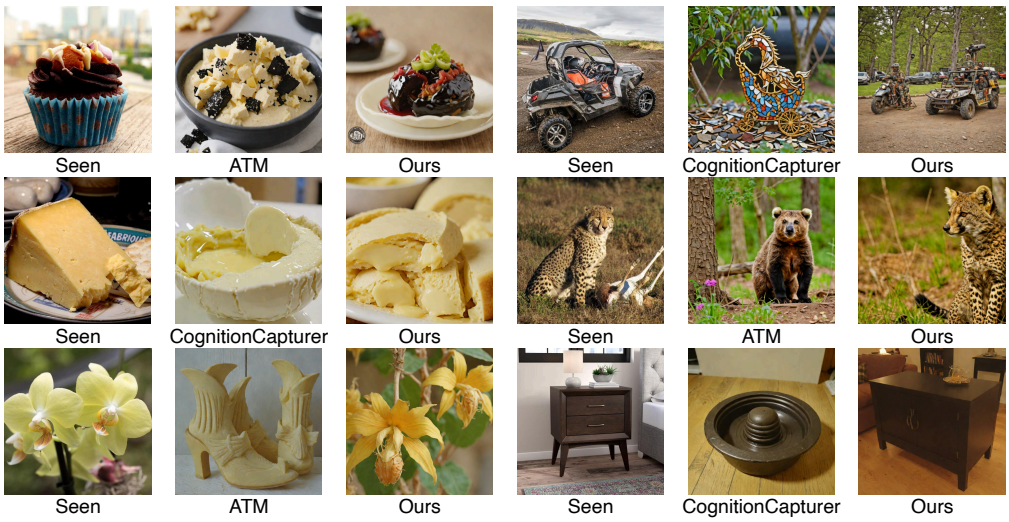


Figure 2: **Qualitative comparison of brain-to-image reconstructions.** Each triplet shows the ground-truth stimulus (left), baseline (middle), and our reconstruction (right). All examples use EEG recordings from subject 8.

4.3 Visual Comparison

In Fig. 3, we show the top-5 retrieved images on the Hard-Case set for our method and the UBP baseline, with our method performing better. We further provide qualitative comparisons with previous brain decoding approaches. As shown in Fig. 2, our method reconstructs images with clearer object contours and more faithful color distribution compared to CognitionCapturer (Zhang et al., 2025) and ATM (Li et al., 2024). In particular, our reconstructions preserve fine-grained structural details while capturing semantically consistent attributes that are often missing in the baselines. Moreover, the overall perceptual quality aligns more closely with the ground-truth stimuli, demonstrating the effectiveness of our framework in bridging brain signals and visual representations.

4.4 Ablation Studies

We study how the composition of visual encoders affects both retrieval and reconstruction. Across various settings, fusing complementary semantics with pixel-level cues consistently outperforms

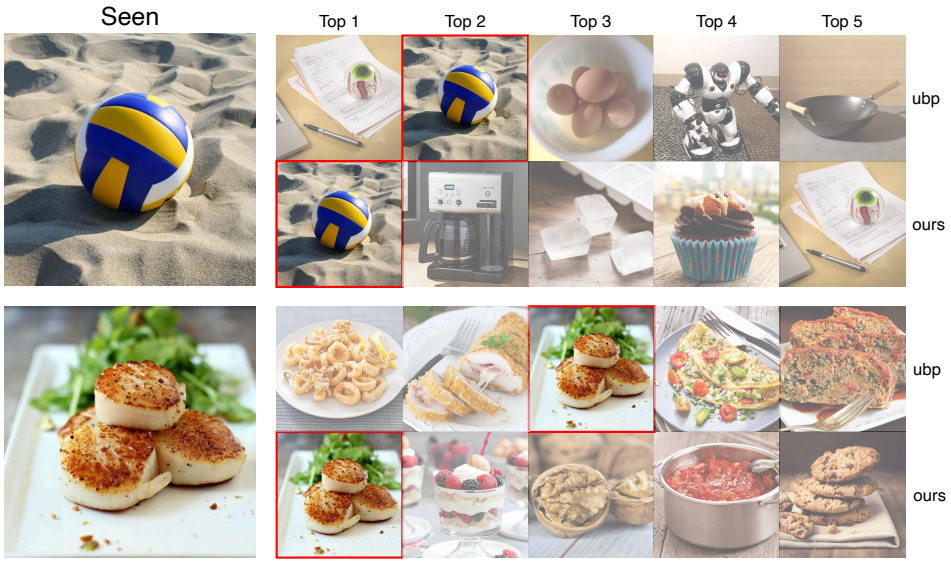


Figure 3: **Hard-case retrieval comparison.** The top-5 retrieved images on the hard-case set from our method and the UBP baseline.

Table 3: Ablation on EEG retrieval: average top-1/top-5 accuracy (%) for 200-way zero-shot; we compare single encoders, pairwise, and triple combinations, with the VAE input fixed at 128×128.

Setting	**Configuration**	**Intra-subject**		**Inter-subject**	
		Top-1	Top-5	Top-1	Top-5
Individual module	B32	52.2	83.3	13.3	33.9
	RN50	48.1	80.4	12.7	31.7
	VAE	44.3	75.2	10.2	23.9
Pairwise combination	RN50 + B32	56.9	86.1	14.4	36.8
	RN50 + VAE	65.8	90.4	17.4	37.3
	B32 + VAE	73.6	94.3	19.1	41.2
Triple combination	RN50 + B32 + VAE	**75.8**	**94.5**	**20.0**	**44.1**

single-encoder baselines. Stacking multiple semantic encoders (e.g., RN50 + B32) brings smaller, saturating gains, whereas adding a VAE latent (fixed at 128×128) provides the largest improvements—suggesting that localized pixel-level features complement CLIP-style semantics that dominate most pipelines.

On the 200-way EEG retrieval (Tab. 3), single encoders form reasonable baselines (e.g., B32: 52.2/83.3 Top-1/Top-5 intra; 13.3/33.9 inter), but semantic stacking alone is modest (RN50+B32: 56.9/86.1 intra; 14.4/36.8 inter). In contrast, pairing a semantic encoder with the VAE yields large jumps (B32+VAE: 73.6/94.3 intra; 19.1/41.2 inter; RN50+VAE: 65.8/90.4 intra; 17.4/37.3 inter). The triple combination (RN50+B32+VAE) offers a small, consistent further boost to 75.8/94.5 (intra) and 20.0/44.1 (inter). Gains persist on the harder inter-subject split with attenuated absolute scores, indicating that low-level structure stabilizes cross-subject variability more effectively than semantics alone.

For reconstruction (Tab. 4), multiscale conditioning improves both pixel-level and recognition metrics. Relative to H14 alone, adding B32 increases SSIM (0.327 $\rightarrow$ 0.340) and strengthens recognition (AlexNet(5): 0.872 $\rightarrow$ 0.908; CLIP: 0.773 $\rightarrow$ 0.814) while lowering SwAV$\downarrow$ (0.574 $\rightarrow$ 0.547). Incorporating the VAE reaches the best PixCorr (0.195) and the strongest AlexNet(2/5) (0.843/0.905) with CLIP competitive (0.808); SSIM remains close to the two-CLIP setting (0.336 vs. 0.340), and Inception shows a mild trade-off (0.756 vs. 0.783). Visually (Fig. 4), B32 enhances

global layout and semantics, VAE sharpens edges and textures, and the combined H14+B32+VAE setting offers the best perceptual balance.

Table 4: Ablation study on the effect of different fusion priors for Brain-to-Image reconstruction.

Prior Setting	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
H14	0.174	0.327	0.825	0.872	0.733	0.773	0.574
H14 + B32	0.187	0.340	0.836	0.908	0.783	0.814	0.547
H14 + VAE	0.173	0.312	0.789	0.838	0.672	0.721	0.611
H14 + B32 + VAE	0.195	0.336	0.843	0.905	0.756	0.808	0.554

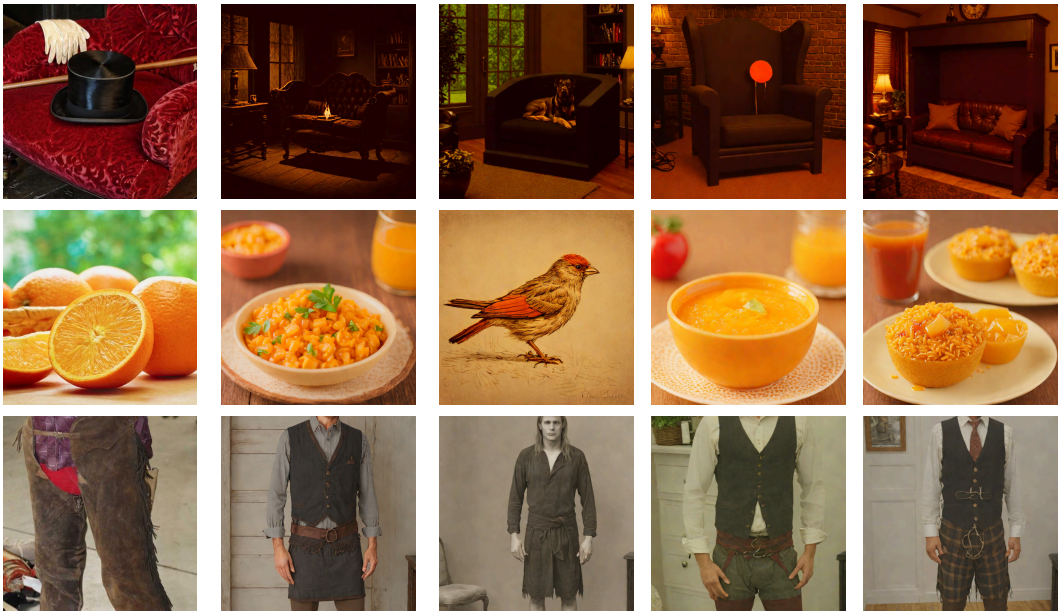


Figure 4: **Ablative study.** Each row shows the ground-truth stimulus and reconstructions produced with different fused configuration: H14, H14+B32, H14+VAE, and H14+B32+VAE. All examples use EEG recordings from subject 8.

The fused configuration, which integrates complementary semantic encoders with a VAE latent, produces a well-balanced system across retrieval and reconstruction. It strengthens cross-subject robustness and preserves fine-grained local structure while maintaining high-level semantic fidelity, reducing typical errors such as oversmoothing and category drift. While effective on average, we do not claim this three-way fusion to be optimal; alternative encoder sets, latent resolutions, or fusion depths may further improve the trade-offs. More qualitative examples are provided in the supplementary materials.

5 Conclusion

In this paper, we present a brain-to-image framework that unifies retrieval and generation through contrastive learning and pretrained vision priors. By integrating multi-level fusion of CLIP and VAE features, our method achieves precise brain-image alignment, while the use of strong diffusion backbones enables high-fidelity image reconstruction. Extensive experiments and ablation studies demonstrate the effectiveness of our design choices, showing clear gains from multi-stream feature fusion, moderate-resolution VAEs, and robust pretrained priors. Our results highlight a scalable and generalizable approach that advances both retrieval accuracy and generative quality in brain visual decoding.

REFERENCES

Pranav Aggarwal, Hareesh Ravi, Naveen Marri, Sachin Kelkar, Fengbin Chen, Vinh Khuc, Midhun Harikumar, Ritiz Tambi, Sudharshan Reddy Kakumanu, Purvak Lapsiya, et al. Controlled and conditional text to image generation with diffusion prior. *arXiv preprint arXiv:2302.11710*, 2023.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*, 2023.

Gary G Blasdel and Jennifer S Lund. Termination of afferent axons in macaque striate cortex. *Journal of Neuroscience*, 3(7):1389–1413, 1983.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Hongzhou Chen, Lianghua He, Yihang Liu, and Longzhen Yang. Visual neural decoding via improved visual-eeg semantic consistency. *arXiv preprint arXiv:2408.06788*, 2024.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):27755, 2016a.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Similarity-based fusion of meg and fmri reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, 26(8):3563–3579, 2016b.

Fernando Lopes da Silva. Eeg and meg: relevance to neuroscience. *Neuron*, 80(5):1112–1128, 2013.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Honghao Fu, Hao Wang, Jing Jih Chin, and Zhiqi Shen. Brainvis: Exploring the bridge between brain and visual signals via image reconstruction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022a. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119754. URL https://www.sciencedirect.com/science/article/pii/S1053811922008758.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022b.

Tijl Grootswagers, Amanda K Robinson, and Thomas A Carlson. The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188:668–679, 2019.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

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

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=RxkcroC8qP.

Haoyu Li, Hao Wu, and Badong Chen. Neuraldiffuser: Neuroscience-inspired diffusion guidance for fmri visual reconstruction. *IEEE Transactions on Image Processing*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Fang Liu, Pei Yang, Yezhi Shu, Niqi Liu, Jenny Sheng, Junwen Luo, Xiaoan Wang, and Yong-Jin Liu. Emotion recognition from few-channel eeg signals by integrating deep feature aggregation and transfer learning. *IEEE Transactions on Affective Computing*, 15(3):1315–1330, 2023a.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36:44860–44879, 2023b.

Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding. *arXiv preprint arXiv:2302.12971*, 2023c.

Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *nature*, 412(6843):150–157, 2001.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022.

William H Merigan and JH Maunsell. How parallel are the primate visual pathways? *Annual review of neuroscience*, 1993.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9879–9889, 2020.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12475–12486, 2021.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14453–14463, 2023a.

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

T Tsumoto, OD Creutzfeldt, and CR Legendy. Functional organization of the corticofugal system from visual cortex to lateral geniculate nucleus in the cat: With an appendix on geniculo-cortical mono-synaptic connections. *Experimental Brain Research*, 32(3):345–364, 1978.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Jia Wang, Jie Hu, Xiaoqi Ma, Hanghang Ma, Xiaoming Wei, and Enhua Wu. Image editing with diffusion models: A survey. *arXiv preprint arXiv:2504.13226*, 2025.

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

Zehan Wang, Yang Zhao, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 36:22099–22114, 2023.

Yayun Wei, Lei Cao, Hao Li, and Yilin Dong. Mb2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8992–9000, 2024.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2246–2257, 2025.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4563–4567. IEEE, 2022.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9653–9663, 2022.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

Xiran Xu, Bo Wang, Boda Xiao, Yadong Niu, Yiwen Wang, Xihong Wu, and Jing Chen. Beware of overestimated decoding performance arising from temporal autocorrelations in electroencephalogram signals. *arXiv preprint arXiv:2405.17024*, 2024.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1179–1189, 2023.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14486–14493, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8552–8562, 2022.

A LLM USAGE STATEMENT

We used LLM for grammar checking and language polishing to improve readability.

B DATASETS DETAILS

THINGS-EEG THINGS-EEG (Gifford et al., 2022a) is a large-scale dataset of electroencephalography (EEG) recordings from 10 participants. Signals are acquired with a 64-channel EASYCAP arranged according to the international 10–10 system. The training split spans 1,654 object concepts, each represented by 10 images; every image is shown four times to each participant ($1{,}654 \times 10 \times 4$). The test split covers 200 concepts with a single image per concept, repeated 80 times ($200 \times 1 \times 80$). Preprocessing follows Wu et al. (2025): raw EEG is filtered to 0.1–100 Hz, yielding 63 channels at 1,000 Hz; trials are segmented from 0–1,000 ms post-stimulus with baseline correction using the prior 200 ms average. Data is then downsampled to 250 Hz, and 17 posterior channels over occipital and parietal sites associated with visual processing are retained. To improve signal-to-noise ratio, repetitions are averaged, producing 16,540 training samples and 200 test samples per participant.

THINGS-MEG THINGS-MEG (Hebart et al., 2023) is a large-scale dataset of magnetoencephalography (MEG) recordings from 4 participants. Signals are acquired with 271 channels. Each trial presents an image for 500 ms, followed by a blank screen of 1000 ± 200 ms. The training split spans 1,854 object concepts, each represented by 12 images; every image is shown once to each participant ($1{,}854 \times 12 \times 1$). The test split covers 200 concepts with a single image per concept, repeated 12 times ($200 \times 1 \times 12$). To construct the zero-shot task, 200 test concepts are discarded from the training set. Preprocessing follows Wu et al. (2025): raw MEG is filtered to 0.1–100 Hz; trials are segmented from 0–1,000 ms post-stimulus with baseline correction. Data is then downsampled to 200 Hz. To improve signal-to-noise ratio, repetitions are averaged, producing 19,848 training samples and 200 test samples per participant.

C RESULTS DETAILS

Per-Subject retrieval on THINGS-EEG and THINGS-MEG We report 200-way zero-shot Top-1/Top-5 accuracy per subject for THINGS-EEG and THINGS-MEG. For each subject, we evaluate individual encoders (RN50, B32, VAE), pairwise stacks (RN50+B32, RN50+VAE, B32+VAE), and the triple stack (RN50+B32+VAE) with the VAE input fixed at 128×128.

Table 5: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG.

Method	Sub1		Sub2		Sub3		Sub4		Sub5		Sub6		Sub7		Sub8		Sub9		Sub10		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
BraVL	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE	13.2	39.5	13.5	40.3	14.5	42.7	20.6	52.7	10.1	31.5	16.5	44.0	17.0	42.1	22.9	56.1	15.4	41.6	17.4	45.8	16.1	43.6
NICE-SA	13.3	40.2	12.1	36.1	15.3	39.6	15.9	49.0	9.8	34.4	14.2	42.4	17.9	43.6	18.2	50.2	14.4	38.7	16.0	42.8	14.7	41.7
NICE-GA	15.2	40.1	13.9	40.1	14.7	42.7	17.6	48.9	9.0	29.7	16.4	44.4	14.9	43.1	20.3	52.1	14.1	39.7	19.6	46.7	15.6	42.8
MB2C	23.7	56.3	22.7	50.5	26.3	60.2	34.8	67.0	21.3	53.0	31.0	62.3	25.0	54.8	39.0	69.3	27.5	59.3	33.2	70.8	28.5	60.4
ATM-S	25.6	60.4	22.0	54.5	25.0	62.4	31.4	60.9	12.9	43.0	21.3	51.1	30.5	61.5	38.8	72.0	34.4	51.5	29.1	63.5	28.5	60.4
VE-SDN	32.6	63.7	34.4	69.9	38.7	73.5	39.8	72.0	29.4	58.6	34.5	68.8	34.5	68.3	49.3	79.8	39.0	69.6	39.8	75.3	37.2	69.9
CognitionCapturer-All	31.4	79.7	31.4	77.8	38.2	85.7	40.4	85.8	24.4	66.3	34.8	78.8	34.7	81.0	48.1	88.6	31.4	79.4	35.6	79.3	35.6	80.2
UBP	41.2	70.5	51.2	80.9	51.2	82.0	51.1	76.9	42.2	72.8	57.5	83.5	49.0	79.9	58.6	85.8	45.1	76.2	61.5	88.2	50.9	79.7
Ours	64.3	88.8	76.3	95.3	74.0	95.0	67	91.8	68.0	91.5	81.5	96.3	76.8	96.8	84.8	98.5	76.8	95.8	87.3	99.3	75.7	94.6

Table 6: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG across different configurations.

Configuration	Sub1		Sub2		Sub3		Sub4		Sub5		Sub6		Sub7		Sub8		Sub9		Sub10		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
B32	39.3	75.3	48.8	79.3	53.3	84.5	54.8	87.0	42.8	75.0	57.8	84.3	47.0	81.0	62.3	89.3	44.3	80.0	61.8	94.3	51.2	83.0
RN50	40.0	69.5	48.5	79.5	48.5	85.0	45.8	82.0	41.5	74.0	55.8	83.0	48.5	77.5	55.5	88.0	41.8	77.0	55.3	89.5	48.1	80.5
VAE	38.8	71.5	41.3	72.5	43.3	75.3	33.5	65.3	39.8	70.5	50.5	81.8	44.3	75.0	55.0	86.3	42.8	73.3	52.0	83.5	44.1	75.5
RN50+B32	47.3	79.0	55.8	81.0	56.3	87.5	59.8	88.8	46.8	81.0	63.0	87.5	53.0	85.0	65.0	91.3	50.0	86.5	68.3	94.5	56.5	86.2
RN50+VAE	60.8	86.0	62.8	92.0	59.8	91.0	53.3	87.0	58.0	84.0	73.0	94.0	62.5	88.8	77.0	97.3	68.3	90.8	77.0	96.5	65.2	90.7
B32+VAE	63.0	88.3	70.3	94.0	73.5	94.3	64.3	92.3	70.5	91.0	78.0	96.0	73.3	93.3	84.8	97.5	75.0	96.0	84.8	98.8	73.7	94.1
RN50+B32+VAE	64.3	88.8	76.3	95.3	74.0	95.0	67	91.8	68.0	91.5	81.5	96.3	76.8	96.8	84.8	98.5	76.8	95.8	87.3	99.3	75.7	94.6

Table 7: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-MEG

Method	Sub1		Sub2		Sub3		Sub4		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
NICE	9.6	27.8	18.5	47.8	14.2	41.6	9.0	26.6	12.8	36.0
NICE-SA	9.8	27.8	18.6	46.4	10.5	38.4	11.7	27.2	12.7	35.0
NICE-GA	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.3	42.3
UBP	15.0	38.0	46.0	80.5	27.3	59.0	18.5	43.5	26.7	55.2
Ours	32.6	63.7	34.4	69.9	38.7	73.5	39.8	72.0	37.2	69.9

Table 8: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG across different configurations.

Configuration	Sub1		Sub2		Sub3		Sub4		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
B32	9.8	31.5	52.8	81.3	31.8	67.0	19.5	47.5	28.4	56.8
RN50	12.0	37.5	50.3	83.0	29.8	65.8	19.0	44.5	27.8	57.7
VAE	12.0	37.5	50.3	83.0	29.8	65.8	19.0	44.5	27.8	57.7
RN50+B32	9.8	31.5	52.0	83.0	32.5	67.8	18.8	47.8	28.3	57.5
RN50+VAE	9.0	22.8	48.0	85.3	26.5	61.3	11.5	30.3	23.8	49.9
B32+VAE	12.0	33.5	64.5	91.3	39.0	76.8	17.3	43.8	33.2	61.3
RN50+B32+VAE	14.0	31.8	63.8	91.8	41.0	78.3	17.0	41.0	33.9	60.7

Per-Subject reconstruction metrics We further report reconstruction metrics per subject. For each subject, we compute low-level measures (PixCorr, SSIM) and high-level perceptual similarity (AlexNet(2/5), Inception, CLIP) with SwAV↓ as a diversity/consistency proxy. Results are shown for the single target (H14 only), semantic pair (H14+B32, H14+VAE) and the full multiscale stack (H14+B32+VAE). The last row gives subject-wise means.

Reconstruction from different subjects As shown in Fig. 5, for the same visual stimulus, we reconstruct images from EEG recorded from different subjects.

Figure 5: **Cross-subject EEG reconstructions.**

Table 9: Reconstruction metrics across subjects using the H14 setting (higher ↑ is better, lower ↓ is better).

	Low-level		High-level				
Subject	Pixcorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
1	0.179	0.305	0.828	0.870	0.719	0.732	0.588
2	0.174	0.331	0.826	0.868	0.712	0.769	0.588
3	0.177	0.317	0.832	0.872	0.703	0.802	0.574
4	0.167	0.326	0.803	0.863	0.752	0.778	0.573
5	0.163	0.315	0.804	0.846	0.676	0.745	0.593
6	0.181	0.316	0.838	0.874	0.715	0.763	0.588
7	0.155	0.328	0.811	0.874	0.736	0.775	0.569
8	0.193	0.349	0.852	0.906	0.781	0.795	0.550
9	0.163	0.330	0.820	0.872	0.765	0.762	0.561
10	0.192	0.350	0.837	0.870	0.773	0.810	0.556
Ave	0.174	0.327	0.825	0.871	0.733	0.773	0.574

Table 10: Reconstruction metrics across subjects using the H14+B32 setting (higher ↑ is better, lower ↓ is better).

	Low-level		High-level				
Subject	Pixcorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
1	0.193	0.317	0.816	0.886	0.761	0.771	0.568
2	0.190	0.346	0.845	0.919	0.789	0.821	0.551
3	0.191	0.330	0.834	0.903	0.758	0.827	0.559
4	0.183	0.334	0.825	0.905	0.805	0.840	0.535
5	0.176	0.326	0.825	0.903	0.720	0.795	0.561
6	0.191	0.326	0.833	0.907	0.791	0.817	0.552
7	0.166	0.337	0.831	0.910	0.765	0.794	0.556
8	0.207	0.365	0.861	0.918	0.815	0.827	0.528
9	0.183	0.348	0.838	0.904	0.792	0.797	0.535
10	0.190	0.365	0.848	0.921	0.830	0.854	0.521
Ave	0.187	0.339	0.836	0.908	0.783	0.814	0.547

Table 11: Reconstruction metrics across subjects using the H14+VAE setting (higher ↑ is better, lower ↓ is better).

	Low-level		High-level				
Subject	Pixcorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
1	0.156	0.301	0.755	0.762	0.653	0.646	0.658
2	0.175	0.323	0.801	0.852	0.643	0.743	0.608
3	0.171	0.290	0.793	0.853	0.651	0.730	0.621
4	0.167	0.307	0.798	0.851	0.707	0.761	0.589
5	0.174	0.295	0.783	0.847	0.670	0.723	0.611
6	0.174	0.304	0.806	0.845	0.669	0.720	0.612
7	0.154	0.315	0.764	0.828	0.655	0.711	0.622
8	0.196	0.335	0.817	0.859	0.691	0.734	0.590
9	0.166	0.315	0.765	0.827	0.678	0.701	0.605
10	0.194	0.337	0.810	0.856	0.703	0.746	0.594
Ave	0.173	0.312	0.789	0.838	0.672	0.721	0.611

Table 12: Reconstruction metrics across subjects using the H14+B32+VAE setting (higher ↑ is better, lower ↓ is better).

	Low-level		High-level				
Subject	Pixcorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
1	0.193	0.332	0.835	0.883	0.727	0.757	0.578
2	0.188	0.341	0.846	0.901	0.769	0.807	0.559
3	0.196	0.324	0.834	0.900	0.755	0.826	0.566
4	0.187	0.320	0.821	0.903	0.774	0.824	0.553
5	0.179	0.317	0.831	0.893	0.705	0.797	0.565
6	0.211	0.329	0.852	0.920	0.758	0.811	0.561
7	0.179	0.336	0.833	0.914	0.754	0.805	0.554
8	0.219	0.356	0.876	0.926	0.788	0.827	0.531
9	0.198	0.342	0.842	0.889	0.757	0.783	0.546
10	0.203	0.358	0.858	0.922	0.777	0.846	0.530
Ave	0.195	0.336	0.843	0.905	0.756	0.808	0.554