SEEING THE IMAGINED: A LATENT FUNCTIONAL ALIGNMENT IN VISUAL IMAGERY DECODING FROM FMRI DATA

Fabrizio Spera,[1*] **Tommaso Boccato,**[2] **Michał Olak**[2], **Sara Cammarota**[1], **Matteo Ciferri**[1], **Michelangelo Tronti**[1], **Nicola Toschi**[1,3†], **Matteo Ferrante**[1,2†].

[1]Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy.
[2]Tether Evo.
[3]A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, USA.
[†]Equal contribution
[*]Correspondence: Fabrizio.Spera@uniroma2.it

ABSTRACT

Recent progress in visual brain decoding from fMRI has been enabled by large-scale datasets such as the Natural Scenes Dataset (NSD) and powerful diffusion-based generative models. While current pipelines are primarily optimized for perception, their performance under mental-imagery remains less well understood. In this work, we study how a state-of-the-art (SOTA) perception decoder (DynaDiff) can be adapted to reconstruct imagined content from the Imagery-NSD benchmark. We propose a latent functional alignment approach that maps imagery-evoked activity into the pretrained model's conditioning space, while keeping the remaining components frozen. To mitigate the limited amount of matched imagery–perception supervision, we further introduce a retrieval-based augmentation strategy that selects semantically related NSD perception trials. Across four subjects, latent functional alignment consistently improves high-level semantic reconstruction metrics relative to the frozen pretrained baseline and a voxel-space ridge alignment baseline, and enables above-chance decoding from multiple cortical regions. These results suggest that semantic structure learned from perception can be leveraged to stabilize and improve visual imagery decoding under out-of-distribution conditions.

1 INTRODUCTION

Decoding the human brain is a rapidly evolving research area that has seen significant progress in recent years. The intersection between neuroscience and modern artificial intelligence (AI) has enabled deeper insights into human visual processing, allowing the reconstruction and retrieval of visual stimuli from neural activity. These advances also pave the way for the development of brain computer interfaces (BCIs) for future clinical applications.
Vision has historically received considerable attention, due to relatively simple experimental settings and increasing availability of large and diverse datasets. These datasets rely on both invasive and non invasive techniques to acquire neural signals, including electrocorticography (ECoG) and implanted microelectrodes, as well as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG), respectively. Although invasive methods generally provide higher signal to noise ratios (SNR), non invasive techniques offer greater safety and reduced risks for patients. Notably, the release of the large scale Natural Scenes Dataset (NSD)(Allen et al., 2022) has demonstrated that competitive decoding performance can be achieved using fMRI data and AI based pipelines to reconstruct images.
Lot of research has been done in visual brain decoding and several recurring patterns emerge: the

task maps a noisy, high-dimensional spatiotemporal signal to a natural image, tipically focusing on regions of interest (ROI) within the visual cortex to reduce neural data dimensionality, based on the assumption that perceptually relevant information is primarily encoded in these areas. Neural activity is then projected into semantic embedding spaces aligned with text and image representations learned by large pretrained models(Radford et al., 2021), which can be used to condition diffusion-based generative models for photorealistic image reconstruction.
In this work, we extend a SOTA visual perception decoding model to mental-imagery fMRI, Dynadiff(Careil et al., 2025). Mental-imagery is probably one of the most fascinating and defining human cognitive processes, which enables the generation of internal visual representations, allowing individuals to mentally reconstruct events from the past or imagine future scenarios. Despite its central role in cognition, decoding mental-imagery from brain activity remains significantly more challenging than decoding visual perception-driven signals, due to its lower SNR signals compared to visual perception(Roy et al., 2023) and the absence of large scale datasets for this task.
Recent results on the newly released Imagery-NSD(Kneeland et al., 2025) indicate that moderately faithful reconstructions of imagined stimuli can be obtained at inference time using models trained on visual perception data. However, the performance of current open source SOTA models show inconsistent performance and appear to strongly depend on architectural design choices. This variability is likely attributable to the distribution shift and the increased inter-subject specific variability: although semantic concepts may be shared between individuals, their mental visualizations can differ substantially, limiting the ability of existing pipelines to produce coherent reconstructions.
To address this distribution shift between imagery and perception data, we propose a latent functional alignment strategy that encourages our model to emphasize shared semantic representations in imagery related neural activity. This alignment leverages the Imagery-NSD dataset while using vision based data, on which the model was originally trained, as a semantic reference through a novel neural data augmentation technique that allowed us to mitigate the paucity of visual data available for this task.
Finally, we conducted a systematic analysis of the contribution of multiple cortical areas to explore whether the proposed pipeline generalizes to additional cortical regions.
The results we obtained with our evaluation protocol significantly improve semantic reconstruction metrics, suggesting that our pipeline is able to significantly improve mental-imagery decoding.
Fig.1 shows our proposed latent functional alignment pipeline to decode imagined content from brain activity.

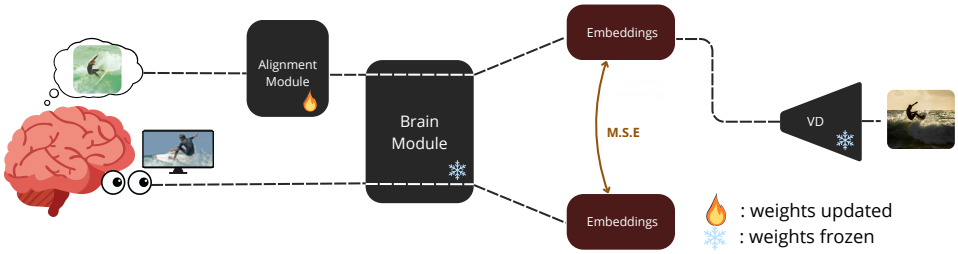

Figure 1: Overview of our pipeline. We use CLIP-Image embeddings predicted by the brain module of the pretrained model from vision trials as targets and train an alignment module to minimize the mean squared error (MSE) between embeddings predicted from imagery trials and the corresponding vision targets; all other components are kept frozen.

1.1 Related Work

Remarkable progress in visual brain decoding has been driven by both the widespread availability of open source generative diffusion models, such as Stable Diffusion (SD)(Rombach et al., 2022), Versatile Diffusion (VD)(Xu et al., 2024) and the emergence of large pretrained vision language models like CLIP(Radford et al., 2021). Equally important has been the release of large scale public datasets, most notably the NSD and the newly published Imagery-NSD, which extends the former by adding dedicated imagery runs, with an addition of runs dedicated to imagery trials.
Current SOTA open source visual brain decoding models generally project fMRI activity into high-dimensional, semantically structured CLIP-Image and CLIP-Text embedding spaces, which

are subsequently used to condition the denoising process of pretrained diffusion models. These approaches often employ multi stage training pipelines and rely on generalized linear model (GLM) approximations to collapse the temporal dimension of the fMRI signals. In this work, the models considered for quantitative and qualitative comparisons in this work include MindEye1(Scotti et al., 2023), BrainDiffuser(Ozcelik and VanRullen, 2023) and MindEye2(Scotti et al., 2024), from the results reported in(Kneeland et al., 2025).

MindEye1 adopts a dual pipeline approach that disentangles semantic and perceptual reconstruction. The high-level pipeline maps fMRI voxels to CLIP-Image embeddings for semantic image generation via diffusion models, while the low-level pipeline predicts SD Variational Auto Encoder (VAE)(Kingma and Welling, 2022)latents to recover coarse perceptual structure, guiding the diffusion process and improving low-level image fidelity.

BrainDiffuser employs a two stage reconstruction framework. In the first stage, fMRI activity is linearly mapped to hierarchical Very Deep VAE (VDVAE)(Child, 2020) latents to produce a coarse low resolution image. In the second stage, a pretrained VD model refines the output by conditioning on CLIP-Image and CLIP-Text embeddings predicted from fMRI.

MindEye2 extends these approaches by pretraining on multi-subject data and fine tuning on a target subject, unifying semantic and perceptual processing. It also predicts image captions to provide textual guidance during final reconstruction refinement.

Recent advances have further lightened the existing pipelines, culminating with the advent of the recent single stage visual brain decoding model DynaDiff(Careil et al., 2025). This model exploits the temporal dynamics of the fMRI data of the NSD participants through a dedicated brain module that processes the neural time series, consisting of a lightweight architecture that applies subject specific projections, timestep dependent transformations through a one dimensional convolutional layer and temporal aggregation to produce embeddings matching the conditioning dimensionality expected by the VD. These embeddings replace CLIP-Image tokens in the cross attention layers of the U-Net(Ronneberger et al., 2015) in the diffusion model. The brain module and the Low-Rank adapter (LoRA)(Hu et al., 2021) for the diffusion model are trained jointly using a standard diffusion loss. For further details on the model, we refer the reader to(Careil et al., 2025).

Although DynaDiff demonstrates strong performance on visual perception tasks, its generalization to other brain processes, such as mental-imagery, remains unexplored. To investigate this, we utilized the Imagery-NSD dataset, which provides data related to imagination tasks from the same subjects of the NSD.

2 METHODS

2.1 DATASETS AND PREPROCESSING

The NSD provides high resolution fMRI measurements from eight participants, each exposed to 9000–10000 distinct natural color images over 30–40 scanning sessions at 7T. The stimuli were drawn from the Microsoft Common Objects in Context (COCO) dataset(Lin et al., 2015), with a shared subset of 1000 images and subject specific unique images. In line with previous research, we focused only on subjects $1, 2, 5$ and 7 since they performed a higher number of runs.

The Imagery-NSD dataset includes 12 additional runs per NSD participant divided into different stimulus and task classes. The dataset contains 18 stimuli in total, divided into six simple stimuli (single or double bars at various orientations), six complex stimuli (same distribution of the NSD training set with one novel image), and six conceptual stimuli (target words). Each stimulus is repeated eight times per run, with 48 trials per run and it has been associated with a particular cue letter that the participants memorized in practice sessions before scanning sessions.

The runs consist of vision tasks for each stimulus type (3 runs) with the presentation of both stimulus and a cue letter, that may or may not match the corresponding stimulus on the screen, and imagination tasks, repeated twice per stimulus (6 runs); in these runs, only cue letters were presented on the screen to the participants, who were instructed to use the cue letter to mentally visualize the associated stimulus. Three additional runs referred to attention tasks were not related to imagination and we excluded them from our study. We used the data from the same subjects we considered in the NSD.

Visual task runs from both the NSD and Imagery-NSD datasets correspond to each subject's nsdgeneral ROI, manually drawn on fsaverage and encompassing voxels responsive to the posterior

cortex (Allen et al., 2022). Additional ROIs used exclusively for the imagination task were defined by parcellating the cortex according to the HCP_MM1 atlas (Glasser et al., 2016). Since the DynaDiff model captures the temporal dynamics of the neural signals, we worked directly with the time series data; keeping the same prescriptions used for the pretrained model we selected data from the functional space of the subjects with an isotropic resolution of 1.8 mm; additionally, a cosine-drift linear model to each voxel in the time series was subtracted from the raw signal; finally, each voxel time series is z-score normalized after each run, in order to not compromise the specificity of the data, linked to the type of task performed, each with its own SNR, as suggested by the authors of Imagery-NSD. Since the time resolution (TR) of acquisition of the Imagery-NSD is equal to 1.0 seconds, but the pretrained model expects data with TRs of 1.3 seconds, we performed per-voxel linear interpolation in order to resample them as expected from the model.

2.2 Data Augmentation

One of the main challenges of the Imagery-NSD dataset is the limited amount of data available to construct a robust and efficient pipeline using only imagery data.
As reported in the Imagery-NSD benchmark, approximately 50% of vision task trials exhibit a mismatch between the displayed image and the corresponding cue letter. We therefore discarded these trials from our analysis. Consequently, half of the 48 total vision trials per run remained usable, making any alignment procedure relying only on the Imagery-NSD data-limited and challenging to implement. To increase the amount of visual data available, we retrieved additional fMRI data from NSD trials, focusing on complex and conceptual stimuli.
For each of the six unique complex stimuli presented to the subjects, we computed the corresponding CLIP-Image embeddings and used a k-nearest neighbors (k-nn) algorithm, based on the relative cosine distance, to identify the nearest CLIP-Image embeddings among the images shown during the NSD runs, as these originate from the same distribution; in order to preserve the semantic coherence with the corresponding ground truth image, we choose a value of k equal to 180 for each stimulus. Once the closest matches were identified, we selected the associated NSD neural data. This strategy substantially increased the amount of usable visual data, since the NSD contains about 27000 visual trials per-subject, far more than the Imagery-NSD ones, enabling us to identify perception-side counterparts that are semantically closer to the imagined stimuli. Once we increased the number of visual data, we augmented the original imagery data by adding a small amount of noise from a standardized gaussian distribution with a variance equal to 0.002, matching the number of the new visual data. This procedure has been applied exclusively in the training phase. A visual representation of our retrieval procedure from NSD is shown in Fig.2.
For conceptual stimuli, we applied a similar procedure using CLIP-Text embeddings derived from the six target words, computing their similarity to the CLIP-Image embeddings of the NSD to select the 180 nearest neural data for each target word.
For simple stimuli we could not rely on this retrieval procedure since their distribution falls completely outside the NSD distribution.

2.3 Latent Functional Alignment

As first attempt, we tried to perform a functional alignment from imagination to visual perception with a Ridge regression for each subject(Ferrante et al., 2024; Wang et al., 2025; Haxby et al., 2020). Please note that we are referring to this procedure as "functional alignment" here, but we are aligning imagery and perception neural representation from fMRI, within each subject, using the same objectives commonly used for across-subject functional alignment. However, this approach generalized poorly and was unable to smooth out the differences between the two different neural inputs for the visual decoding pipeline. Probably, this was due to the lack of available data and the intrinsic differences between imagination and visual data, both in terms of SNR and activations.
To overcome this problem, since imagination and vision runs are available for each stimulus type in Imagery-NSD, we sought to establish a latent functional alignment between the neural responses evoked during perception and those elicited during imagery, as shown in fig.1. The underlying idea is pretty simple: when we imagine something, the mental image could be very blurry or with very little detail, and this vary across people as well, but the concept is usually very clear. So the question is: **why instead of looking for a detail match in a very high dimensional and noisy space we do not look for this alignment on a more semantic space?**

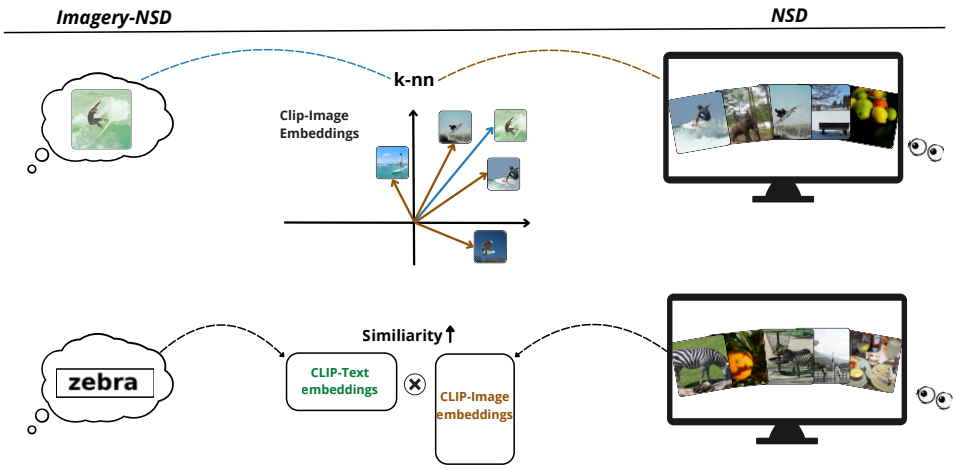

Figure 2: Illustration of the retrieval procedure. For complex stimuli, we compute CLIP-Image embeddings and retrieve nearest neighbors from NSD via k-nn in cosine space, using the associated NSD brain data as targets. For conceptual stimuli, we compute CLIP-Text embeddings for the target words and retrieve NSD images via CLIP-Text/CLIP-Image similarity.

To perform this alignment, we minimized the MSE between the output of the brain module for imagination and the corresponding visual target data, thereby enforcing similarity between the CLIP-Image embeddings expected from the visual decoder.
To reduce the latent space distance between the two different processes, we introduced an alignment module preceding the brain module for imagery data. This module applies voxel wise nonlinear transformations, projecting them into a higher-level feature space independently at each time step (no temporal mixing) of the fMRI signal.
During the training of the alignment module, we used an AdamW optimizer with a learning rate of 10^{-5} and used 20% of the augmented training data for the validation set; the batch size for both is equal to 36; the other modules of the original architecture were kept frozen. The number of epochs was different depending on the class of stimuli and chosen via early stopping on the validation set; in particular, for simple stimuli, we trained the alignment module just for a maximum of three epochs only when we used data not from the visual cortex , since we had only the vision data from the visual runs of the Imagery-NSD as target for them; while for data from the visual cortex, we used the output of DynaDiff. For complex and conceptual stimuli we empirically evaluated that exceeding eight epochs would have entailed overfitting. More in detail, the alignment modules for the individual ROIs consist of a multilayer perceptron with a single hidden layer, layer normalization and a non linear GELU activation function. Its output dimensionality matches the voxel count expected by the original brain module, which corresponds to the number of voxels, V_{vis}, in the nsdgeneral ROI for each subject.
Formally, we indicate the input time series data as $\mathbf{X}$, such that $\mathbf{X} \in \mathbb{R}^{B\times V\times T}$, where B is the batch dimension, V is the number of input voxels that depends both on the subject and the ROI considered as reported in table 1, while T is the number of time points. The first transformation can be represented as a multidimensional function that maps the original voxel data from a dimensional space V, to a vector $\mathbf{h} = \mathbf{f}(\mathbf{X}; \boldsymbol{\theta})$ of dimension L. Next, we performed standard layer normalization, rescaling $\mathbf{h}$ with its mean value $\boldsymbol{\mu}$ and variance σ^2, obtaining $\hat{\mathbf{h}}$. Then we applied the non linear transformation $\mathbf{z} = GELU(\hat{\mathbf{h}})$ and projected $\mathbf{z}$ into a V_{vis} dimensional space, as expected from the pretrained model, by a function $\mathbf{g}(\mathbf{z}; \boldsymbol{\omega})$. Then we used the resulting tensor as input for the rest of the DynaDiff backbone architecture modules, which we collectively indicate as a function that maps time-series neural data from the visual cortex to the corresponding space of the generated images of 3 channels and size 512×512 pixel: $\mathcal{D} : \mathbb{R}^{B\times V_{vis}\times T} \rightarrow \mathbb{R}^{B\times 3\times 512\times 512}$.

Table 1: Number of voxels in different ROIs for each subject.

Subject	nsdgeneral	prefrontal cortex	frontal cortex	temporal lobes	parietal cortex
1	15724	14467	10553	12166	14200
2	14278	15587	11360	12487	12494
5	13039	11791	9792	10811	11408
7	12682	11738	8760	10240	10430

2.3.1 EVALUATION

In brain decoding literature, there is no single metric capable of fully describing the quality of a reconstructed image; for this reason different metrics are used, such that each of them would be sensitive to a complementary aspect of the reconstruction. Therefore, to evaluate the quality of reconstructed images for both simple and complex stimuli, we computed metrics related to low-level details and high-level and semantic coherence, which have become a benchmark for quantifying the quality of results in visual brain decoding, since many studies have highlighted how hierarchical and multi-level deep neural networks (DNNs) seem to share common mechanisms with the visual process in human brain(Kriegeskorte, 2015; Doerig et al., 2025).
Low-level metrics PixCorr and SSIM measure, respectively, the pixel-level correlation and the structural similarity index metric(Wang et al., 2004). They assess local visual fidelity and are useful for measuring the retrieval of basic visual information, but they do not capture the meaning of the image. The high-level metrics AlexNet(2), AlexNet(5), CLIP and Inception are recovered from specific layers of different DNNs, which allow us to estimate the semantic and conceptual quality of reconstructions, going beyond pixel-wise comparison and capturing the content of the image even when visual details differ. Evaluation is performed using a pairwise identification protocol: for each reconstructed image, the similarity between its feature embedding and that of the correct target image is compared against the similarities with the feature embeddings of all other target images in the test set. Performance is defined as the proportion of pairwise comparisons in which the correct target is ranked higher than the incorrect one. Since each comparison is binary, the chance-level performance is 50%. Alex(2) and Alex(5) refer to the layers 2 and 5 of AlexNet(Hinton et al., 2012), CLIP to the output layer of the ViT-L/14 CLIP-Vision model(Radford et al., 2021) and Inception to the last pooling layer of InceptionV3(Szegedy et al., 2015). Lastly, EffNet-B and SwAV are distance metrics derived from EfficientNet-B13(Tan and Le, 2020) and SwAV-ResNet50(Caron et al., 2021) models. Since our pipeline is based on enhancing the semantical correspondence between reconstructed images from imagery tasks and real target stimuli, rather than low-level details, the most informative metrics for our analysis are the high-level ones, since they provide a better estimate of the correspondence of the meaning of the decoded content to a given target image.
As baselines, we used both the pretrained visual decoding model, as done in(Kneeland et al., 2025), in order to have a term of comparison on the generalization capability of our functional latent alignment approach with respect to the original decoding pipeline for visual stimuli and then a Ridge regression predicting fMRI voxels activity from the imagery data to the visual ones in the input space before the brain module, with values of α, the penalty parameter, ranging from $[10^2, 10^6]$, as baseline for functional alignment. The code to reproduce the results is available at the following link: https://github.com/Hortomuso/LatentFunctionalAlignment.

3 RESULTS

In this section we report the results of our latent functional alignment pipeline between mental-imagery and vision, quantifying the effect of the proposed alignment on decoding performance of visual decoding pipelines from imagination data. As shown in Fig. 3 for subject 1, our best reconstructions are semantically coherent with the corresponding targets. They were not necessarily selected from the same generation process of the VD, in order to visually compare our most visually faithful reconstructions with the ones obtained in(Kneeland et al., 2025). Comparing our results with the ones obtained with the open source models(Ozcelik and VanRullen, 2023)(Scotti et al., 2023)(Scotti et al., 2024), reported in(Kneeland et al., 2025), we observe that our pipeline generates more photorealistic reconstructions.
Indeed, even if little details could be missed in the reconstructions, the semantic content is correctly

preserved, which is consistent with the nature of mental-imagery, where the general concept has greater importance than lower levels characteristics. We reported in the appendix the reconstructions for all the subjects for each class of stimuli from the visual cortex.

3.1 VISUAL BRAIN REGION

Figure 3: Qualitative results of the *top* reconstruction for subject 1 of imagined complex and conceptual stimuli. Our results show a strong semantical consistency between the ground truth and reconstructed images, improving the performance obtained by other SOTA models.

We report in table 2 the metrics averaged between the 4 subjects: the values for each subject were obtained using 20 remaining trials in the test set after the initial splitting, of which 10 were referred to simple stimuli and 10 to complex stimuli. To quantify variability, we performed 10 different reconstructions on the same images of the test set, using different random seeds for VD; then we used the standard error mean (SEM) on all 4 subjects to obtain averaged values between subjects, with the corresponding standard deviations.

Table 2: Quantitative reconstruction results for simple and complex stimuli, averaged across all subjects. All metrics are obtained with SEM. Arrows indicate whether higher (↑) or lower (↓) values correspond to better performance.

Method	**Low-Level**		**High-Level**				**Distance**	
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
NSD-Imagery Mental-Imagery Trials – All Subjects – Visual Cortex								
Dynadiff Baseline	**0.0295**	**0.3431**	51.03%	50.21%	51.39%	48.94%	0.9890	0.6210
Functional Alignment	−0.0003	0.3408	46.59%	43.55%	44.33%	43.02%	1.0002	0.6476
Latent Functional Alignment	0.0013	0.3376	**52.02%**	**58.71%**	**54.94%**	**59.13%**	**0.9616**	**0.5964**

The results show an improvement for the high-level and distance metrics, demonstrating that our latent functional alignment effectively increases the decoding performance for the imagery task, with respect to both the original model and the functional alignment; in particular, in this last case, we can see a drop for the decoding performance, suggesting that a simple linear Ridge regression could not be enough complex to capture the relationship between imagery and vision; the corresponding standard deviations are reported in appendix in table 4. Figure 4 illustrates the increase in both CLIP and Alex(5) after our alignment for subjects 1,2 and 5, as confirmed by Wilcoxon signed-rank

test, whose results are reported in the appendix in table 5. These scores show that it is possible

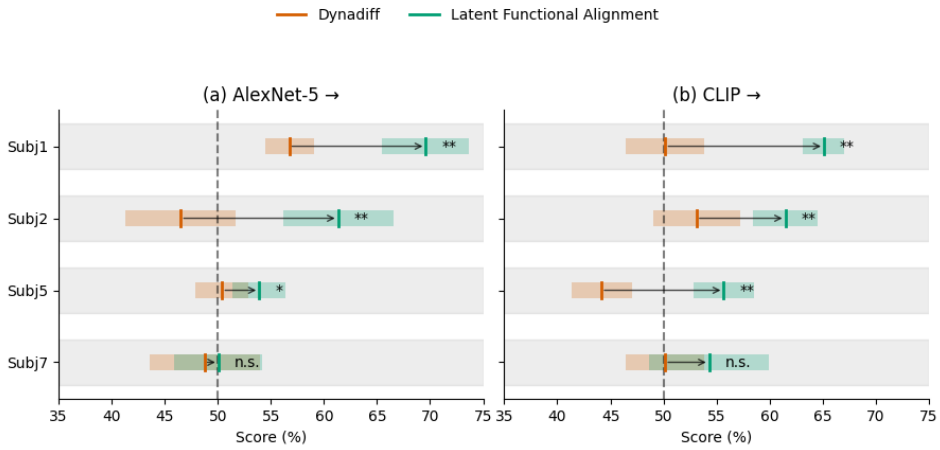

Figure 4: Per-subject Alex(5) and CLIP scores, with relative significances. The gray dashed line denotes chance level for each metric. The Dynadiff baseline performs close to chance, whereas latent functional alignment improves performance across subjects, in particular we obtained significant improvements for subjects 1,2 and 5, while subject 7 shows more variability in the results.

to reconstruct coherent images also for imagination data that keep the semantic meaning of the corresponding target image, using pretrained visual models.
Also for conceptual stimuli, we conducted a similar analysis to retrieve additional vision runs from the NSD to use as target for the alignment; the corresponding values we obtained for the CLIP-Image and CLIP-Text similarity, for each subject, are reported in Table 3.

Table 3: Similarity between CLIP-Image and CLIP-Text embeddings for conceptual stimuli in visual cortex. Values are mean ± standard deviation across ten different reconstruction trials for single subject while SEM for the averaged values.

Method	Subj1	Subj2	Subj5	Subj7	All Subjects (Mean ± SEM)
NSD-Imagery Conceptual Stimuli – CLIP-Image/text Similarity-Visual Cortex					
Dynadiff Baseline	0.1878 ± 0.0169	0.1956 ± 0.0240	0.1947 ± 0.0288	0.1895 ± 0.0273	0.1919 ± 0.0020
Functional Alignment	0.1907 ± 0.0409	0.1885 ± 0.0294	0.1966 ± 0.0347	0.1833 ± 0.0298	0.1898 ± 0.0028
Latent Functional Alignment	$\mathbf{0.2196 \pm 0.0369}$	$\mathbf{0.2381 \pm 0.0443}$	$\mathbf{0.2114 \pm 0.0395}$	$\mathbf{0.2045 \pm 0.0322}$	0.2184 ± 0.0072

3.2 Role of non-visual brain regions

From a neuroscientific point of view, the mechanisms and ROIs involved during mental-imagery have always been highly debated and investigated (Pearson et al., 2015)(Abraham and Bubić, 2015). Most studies on the role of the visual cortex in this mental process remain active and have yielded mixed findings depending on subject-specific factors and individual neuroanatomy(Xie et al., 2020; Dijkstra, 2024; Bridge et al., 2012; Milton, 2024; Dijkstra et al., 2017; Ganis et al., 2004).
Our goal in this section is to quantitatively compare the contribution of the main cortical areas of the brain, in order to infer which ROIs could play a major role for this tasks.
Our procedure consists in selecting the specific ROIs of the cerebral cortex from the HCP_MM1 atlas, in the functional space of the subjects, using the same resolution and preprocessing as for the data of the visual cortex from the ndsgeneral ROI.
First, we evaluated the performance of the single areas of the brain, as in the previous section in particular for the high-level features CLIP and Alex(5), then we added to each area the voxels related to the nsdgeneral ROI used in the previous section, in order to measure how much the contribution of the posterior cortex could be important in mental-imagery.
To have a global view of the results, we averaged the results across all the 4 subjects, which are reported in Fig.5. While for conceptual stimuli we obtain the averaged results, reported in Fig.6.

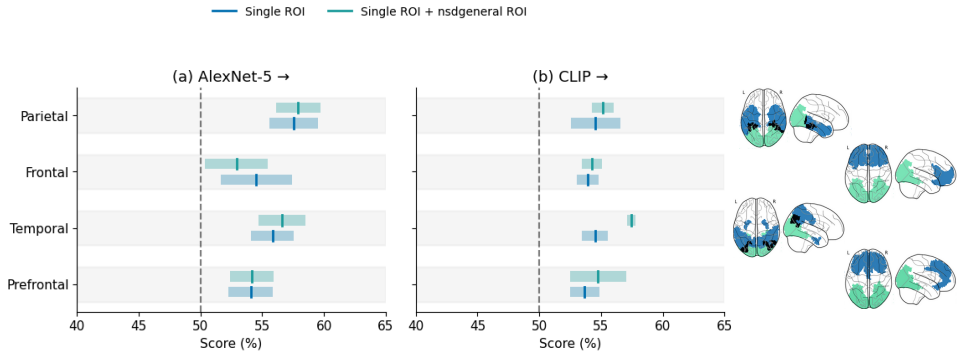

Figure 5: Bar plot of the averaged value across the four subjects for Alex(5) and CLIP for complex and simple stimuli reconstructions. These results are from both the contribution of single ROIs and with the addition on the nsdgeneral ROI. On the right side there are the corresponding regions activated; the teal color refers to the posterior cortex, while blue to the other ROIs; the black points indicate the overlap between the posterior cortex and the considered ROI.

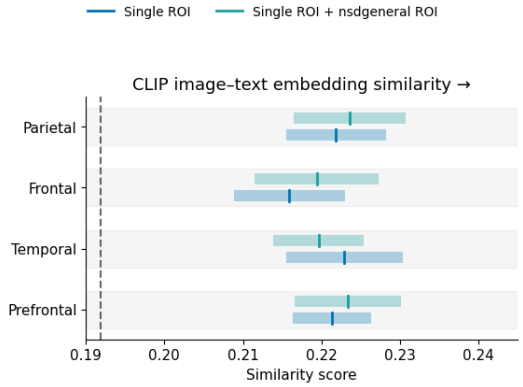

Figure 6: Bar plot of the averaged value across the 4 subjects for the similarity score between image reconstructions and the corresponding target text embeddings for conceptual stimuli. These results are from both the single ROIs and with the addition on the nsdgeneral ROI; the vertical dashed line indicates the best baseline from table 3.

4 Discussions

In this work, we investigated the problem of brain decoding from visual imagery by systematically analyzing the contribution of distinct cortical regions. Compared to visual perception, imagery constitutes a substantially more challenging decoding target, as it is driven by internally generated working-memory and top-down processes rather than by direct sensory input. As an initial step, we focused exclusively on visual cortical regions, leveraging a pretrained model specific to this ROI to enable direct comparison with the previous study of Kneeland et al. (2025) and to reflect the well-established role of visual cortex in mental-imagery (Dijkstra et al., 2017; Ganis et al., 2004). Results obtained both at the single-subject level and when averaged across the four participants we considered demonstrate that the proposed latent functional alignment framework extracts semantically meaningful representations from imagery-related neural signals. Importantly, all high-level evaluation metrics were consistently above chance, indicating that despite the weaker and more variable nature of imagery signals, a shared latent semantic structure can be robustly recovered.
We subsequently extended the analysis to parietal, temporal, frontal, and prefrontal regions to assess the contribution of non-visual areas to imagery decoding. Within this framework, posterior cortex emerged as the primary contributor to decoding performance, as observed in (Wang et al., 2024), while reconstructions based on individual non-visual ROIs generally yielded lower scores. Nevertheless, all regions consistently performed above chance, with values comparable to those

reported in previous imagery decoding studies. This observation provides evidence that imagery-related semantic information is not restricted to early visual areas but is distributed across multiple cortical systems, in line with prior findings (Wang et al., 2024). Combining individual ROIs with the posterior cortex resulted in only marginal performance improvements. One plausible explanation is that, given a fixed visual alignment target, the addition of a single ROI may not contribute sufficiently complementary information to substantially enhance decoding. This effect is likely amplified by the limited amount of available imagery data, which limits the ability of the model to learn more complex cross-regional interactions. A notable exception to this trend is the temporal cortex. Across experiments, this region consistently benefited from integration with additional ROIs, leading to more stable and, in some cases, improved reconstruction metrics. This result aligns with the established involvement of temporal regions in semantic and conceptual processing (Berger and Ehrsson, 2014; Jing et al., 2025), which is particularly relevant for visual imagery, where semantic content plays a central role in the shaping of sensory-like representations. In contrast, frontal and prefrontal regions exhibited more variable and generally limited contributions to decoding performance. This variability suggests that their involvement in imagery decoding may depend strongly on model architecture or fusion strategy. Rather than directly encoding visual or semantic information, these regions may primarily support higher-order cognitive functions such as control, goal maintenance, or imagery initiation, which could not be optimally captured by the current evaluation metrics. More broadly, our results suggest that the proposed alignment framework can capture signatures of top-down generative and working-memory processes underlying image-related neural activity, originating in higher-level cortical regions and propagating toward visual areas. Although these internally generated representations differ fundamentally from perceptual ones, they remain semantically related. The model's ability to exploit such signals to achieve reliable decoding performance above chance level underscores its sensitivity to non-sensory neural representations.

Overall, these findings support the view of visual imagery as a distributed process engaging both sensory and higher-order cognitive regions. The consistent decoding performance observed across all ROIs provides converging evidence that semantic information during imagery is widely represented across the cortex, rather than being restricted to early visual areas.

4.1 Limitations and Conclusion

Despite these promising results, several limitations must be acknowledged. First, the limited availability of imagery data, combined with the strong inter-subject variability inherent to mental-imagery, restricts the extent to which precise neuroscientific conclusions about the specific functional role of individual ROIs can be drawn. Second, the evaluation metrics used in this study were originally designed for perception-based decoding and may not fully capture the qualitative nature of imagery-related representations.

Moreover, while data augmentation strategies improved performance, the reliance on visual data derived from perception runs may have partially influenced the learned representations, potentially biasing the alignment toward perceptual rather than purely imagery-driven features.

Beyond methodological considerations, these results raise broader conceptual and ethical questions. As access to neural data increases and AI models become more powerful, the amount of information that can be decoded from brain activity continues to grow. Decoding imagination, which is arguably one of the most private and subjective aspects of human experience, opens unprecedented opportunities, from new forms of human–AI co-creation to technologies that could externalize memories, dreams, or creative intent. At the same time, it calls for serious reflection on issues of privacy, neural rights, model bias, and interpretability. How can safeguards be designed to protect individuals from misuse of such technologies? How can we disentangle neural noise, subjective variability, and algorithmic bias in decoded representations? These remain open and pressing questions.

Despite these limitations, the present study provides evidence that visual imagery engages a broad and distributed cortical network and demonstrates that multi-ROI decoding strategies constitute a promising direction for advancing BCIs and semantic reconstruction from neural signals. Future work will focus on developing more sophisticated ROI fusion mechanisms, including more specific areas of the cortical region rather than extended regions, designing evaluation metrics tailored specifically to imagery, and validating the generalizability of these findings across larger cohorts and more diverse semantic domains. Ultimately, understanding how imagination is encoded in the brain may not only improve decoding performance, but also deepen our understanding of the neural foundations of human creativity and inner experience.

REFERENCES

Anna Abraham and Andreja Bubić. Semantic memory as the root of imagination. *Frontiers in Psychology*, 6:325, 03 2015. doi: 10.3389/fpsyg.2015.00325.

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL https://doi.org/10.1038/s41593-021-00962-x.

Christopher Berger and H.Henrik Ehrsson. The fusion of mental imagery and sensation in the temporal association cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 34:13684–13692, 10 2014. doi: 10.1523/JNEUROSCI.0943-14.2014.

Holly Bridge, Stephen Harrold, Emily A. Holmes, Mark Stokes, and Christopher Kennard. Vivid visual mental imagery in the absence of the primary visual cortex. *Journal of Neurology*, 259(6): 1062–1070, June 2012. ISSN 1432-1459. doi: 10.1007/s00415-011-6299-z.

Marlène Careil, Yohann Benchetrit, and Jean-Rémi King. Dynadiff: Single-stage decoding of images from continuously evolving fmri, 2025. URL https://arxiv.org/abs/2505.14556.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021. URL https://arxiv.org/abs/2006.09882.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images, 11 2020.

Nadine Dijkstra. Uncovering the role of the early visual cortex in visual mental imagery. *Vision*, 8: 29, 05 2024. doi: 10.3390/vision8020029.

Nadine Dijkstra, Sander Bosch, and Marcel van Gerven. Vividness of visual imagery depends on the neural overlap with perception in visual areas. *The Journal of Neuroscience*, 37:3022–16, 01 2017. doi: 10.1523/JNEUROSCI.3022-16.2016.

Adrien Doerig, Tim Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7:1220–1234, 08 2025. doi: 10.1038/s42256-025-01072-0.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Through their eyes: Multi-subject brain decoding with simple alignment techniques. *Imaging Neuroscience*, 2:imag–2–00170, May 2024. ISSN 2837-6056. doi: 10.1162/imag_a_00170. URL https://doi.org/10.1162/imag_a_00170. _eprint: https://direct.mit.edu/imag/article-pdf/doi/10.1162/imag_a_00170/2370814/imag_a_00170.pdf.

Giorgio Ganis, William Thompson, and Stephen Kosslyn. Brain areas underlying visual mental imagery and visual perception: an fmri study. *Brain research. Cognitive brain research*, 20: 226–41, 08 2004. doi: 10.1016/j.cogbrainres.2004.02.012.

Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, August 2016. ISSN 1476-4687. doi: 10.1038/nature18933. URL https://doi.org/10.1038/nature18933.

James Haxby, Jyothi Swaroop Guntupalli, Samuel Nastase, and Ma Feilong. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife Sciences*, 9:e56601, 06 2020. doi: 10.7554/eLife.56601.

Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Yoesoep Rachmad. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pages 1097–1105, 01 2012. doi: 10.1145/3065386.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Gu Jing, Xueyang Wang, Cheng Liu, Lin Yang, Jiaxin Fan, Jiangzhou Sun, Yoed Kenett, and Jiang Qiu. Cognitive and neural mechanisms of mental imagery supporting creative cognition. *Communications Biology*, 8, 09 2025. doi: 10.1038/s42003-025-08513-x.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Reese Kneeland, Paul S. Scotti, Ghislain St-Yves, Jesse Breedlove, Kendrick Kay, and Thomas Naselaris. Nsd-imagery: A benchmark dataset for extending fmri vision decoding methods to mental imagery, 2025. URL https://arxiv.org/abs/2506.06898.

Nikolaus Kriegeskorte. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1 (Volume 1, 2015):417–446, 2015. ISSN 2374-4650. doi: https://doi.org/10.1146/annurev-vision-082114-035447. URL https://www.annualreviews.org/content/journals/10.1146/annurev-vision-082114-035447. Type: Journal Article.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Fraser Milton. Mental imagery: The role of primary visual cortex in aphantasia. *Current Biology*, 34(21):R1088–R1090, 2024. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2024.09.076. URL https://www.sciencedirect.com/science/article/pii/S0960982224013447.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, September 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-42891-8. URL https://doi.org/10.1038/s41598-023-42891-8.

Joel Pearson, Thomas Naselaris, Emily Holmes, and Stephen Kosslyn. Mental imagery: Functional mechanisms and clinical applications. *Trends in Cognitive Sciences*, 19:590–602, 10 2015. doi: 10.1016/j.tics.2015.08.003.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Tiasha Roy, Jesse Breedlove, Ghislain St-Yves, Kendrick Kay, and Thomas Naselaris. Comparison of signal to noise in vision and imagery for qualitatively different kinds of stimuli. *Journal of Vision*, 23:5961, 08 2023. doi: 10.1167/jov.23.9.5961.

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors, 2023. URL https://arxiv.org/abs/2305.18274.

Paul S. Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data, 2024. URL https://arxiv.org/abs/2403.11207.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. URL https://arxiv.org/abs/1512.00567.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL https://arxiv.org/abs/1905.11946.

Haibao Wang, Jun Ho, Fan Cheng, Shuntaro Aoki, Yusuke Muraki, Misato Tanaka, Jong-Yun Park, and Yukiyasu Kamitani. Inter-individual and inter-site neural code conversion without shared stimuli. *Nature Computational Science*, 5:534–546, 07 2025. doi: 10.1038/s43588-025-00826-5.

Yanchen Wang, Adam Turnbull, Tiange Xiang, Yunlong Xu, Sa Zhou, Adnan Masoud, Shekoofeh Azizi, Feng Lin, and Ehsan Adeli. Decoding visual experience and mapping semantics through whole-brain analysis using fmri foundation models, 11 2024.

Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13:600 – 612, 05 2004. doi: 10.1109/TIP.2003.819861.

Siying Xie, Daniel Kaiser, and Radoslaw Cichy. Visual imagery and perception share neural representations in the alpha frequency band. *Current Biology*, 30:3062, 08 2020. doi: 10.1016/j.cub.2020.07.023.

Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model, 2024. URL https://arxiv.org/abs/2211.08332.

A TABLES AND QUALITATIVE RECONSTRUCTIONS

Table 4: Mean-subjects metrics with the corresponding standard deviations from table 2, obtained with SEM.

Method	**Low-Level**		**High-Level**				**Distance**	
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
NSD-Imagery Mental-Imagery Trials - All Subjects - Visual Cortex								
Dynadiff Baseline	$\mathbf{0.0295 \pm 0.0130}$	$\mathbf{0.3431 \pm 0.0044}$	$(51.03 \pm 1.60)\%$	$(50.21 \pm 1.82)\%$	$(51.39 \pm 1.72)\%$	$(48.94 \pm 1.88)\%$	0.9890 ± 0.0033	0.6210 ± 0.0025
Functional Alignment	-0.0003 ± 0.0072	0.3408 ± 0.0044	$(46.59 \pm 2.24)\%$	$(43.55 \pm 2.73)\%$	$(44.33 \pm 2.01)\%$	$(43.02 \pm 1.52)\%$	1.0002 ± 0.0037	0.6476 ± 0.0025
Latent Functional Alignment	0.0013 ± 0.0087	0.3376 ± 0.0103	$(\mathbf{52.02 \pm 2.92})\%$	$(\mathbf{58.71 \pm 4.31})\%$	$(\mathbf{54.94 \pm 2.09})\%$	$(\mathbf{59.13 \pm 2.53})\%$	$\mathbf{0.9616 \pm 0.0061}$	$\mathbf{0.5964 \pm 0.0079}$

Table 5: Subject-wise statistical comparison between Dynadiff and Latent Functional Alignment for CLIP and AlexNet-5 metrics. Significance was assessed using a one-sided Wilcoxon signed-rank test on paired samples. Effect sizes are reported as the median of paired differences (Median Δ) with 95% bootstrap confidence intervals.

Subject	Median $\Delta \uparrow$	95% CI	W	p-value	Sig.
		CLIP			
1	$+\mathbf{0.1474}$	$[0.1316,\ 0.1674]$	55.0	0.0010	**
2	$+\mathbf{0.0513}$	$[0.0132,\ 0.1289]$	52.0	0.0049	**
5	$+\mathbf{0.1144}$	$[0.0711,\ 0.1605]$	55.0	0.0010	**
7	-0.0158	$[-0.0447,\ 0.0000]$	9.0	0.9512	n.s.
		AlexNet-5			
1	$+\mathbf{0.1329}$	$[0.0973,\ 0.1539]$	55.0	0.0010	**
2	$+\mathbf{0.1831}$	$[0.1289,\ 0.2395]$	55.0	0.0010	**
5	$+0.0474$	$[-0.0026,\ 0.0580]$	48.0	0.0186	*
7	-0.0053	$[-0.0540,\ 0.0526]$	23.5	0.4707	n.s.

Table 6: Per-subject table of low-level, high-level and distance metrics for complex and simple stimuli.

Method	**Low-Level**		**High-Level**				**Distance**	
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
NSD-Imagery Mental-Imagery Trials - All the Subjects - Visual Cortex								
Dynadiff Baseline - Sub1	$\mathbf{0.0608 \pm 0.0100}$	0.3505 ± 0.0085	$(54.13 \pm 3.77)\%$	$(55.13 \pm 4.18)\%$	$(53.79 \pm 4.99)\%$	$(48.32 \pm 3.36)\%$	0.9819 ± 0.0075	0.6154 ± 0.0038
Latent Functional Alignment - Sub1	0.0270 ± 0.0092	0.3245 ± 0.0094	$(\mathbf{57.97 \pm 2.53})\%$	$(\mathbf{69.58 \pm 4.10})\%$	$(\mathbf{57.74 \pm 5.85})\%$	$(\mathbf{65.11 \pm 1.95})\%$	0.9567 ± 0.0108	$\mathbf{0.5840 \pm 0.0079}$
Dynadiff Baseline - Sub2	0.0128 ± 0.0123	0.3308 ± 0.0060	$(48.71 \pm 2.17)\%$	$(46.53 \pm 3.43)\%$	$(50.13 \pm 6.44)\%$	$(53.13 \pm 4.09)\%$	0.9971 ± 0.0078	0.6245 ± 0.0066
Latent Functional Alignment - Sub2	-0.0043 ± 0.0131	0.3215 ± 0.0044	$(55.16 \pm 4.26)\%$	$(61.32 \pm 5.19)\%$	$(56.00 \pm 6.49)\%$	$(61.47 \pm 3.03)\%$	0.9640 ± 0.0065	0.5921 ± 0.0083
Dynadiff Baseline - Sub5	0.0042 ± 0.0151	0.3427 ± 0.0033	$(53.39 \pm 2.86)\%$	$(50.42 \pm 2.10)\%$	$(47.11 \pm 4.29)\%$	$(44.16 \pm 2.84)\%$	0.9911 ± 0.0090	0.6181 ± 0.0058
Latent Functional Alignment - Sub5	-0.0114 ± 0.0080	$\mathbf{0.3668 \pm 0.0091}$	$(50.26 \pm 3.04)\%$	$(53.89 \pm 2.49)\%$	$(57.24 \pm 2.42)\%$	$(55.66 \pm 2.83)\%$	$\mathbf{0.9486 \pm 0.0084}$	0.5900 ± 0.0066
Dynadiff Baseline - Sub7	0.0402 ± 0.0096	0.3484 ± 0.0052	$(47.87 \pm 4.05)\%$	$(48.76 \pm 5.19)\%$	$(54.53 \pm 6.41)\%$	$(50.16 \pm 3.70)\%$	0.9859 ± 0.0108	0.6260 ± 0.0087
Latent Functional Alignment - Sub7	-0.0060 ± 0.0140	0.3376 ± 0.0093	$(44.68 \pm 3.48)\%$	$(50.05 \pm 4.13)\%$	$(48.76 \pm 5.72)\%$	$(54.29 \pm 5.64)\%$	0.9772 ± 0.0065	0.6197 ± 0.0062

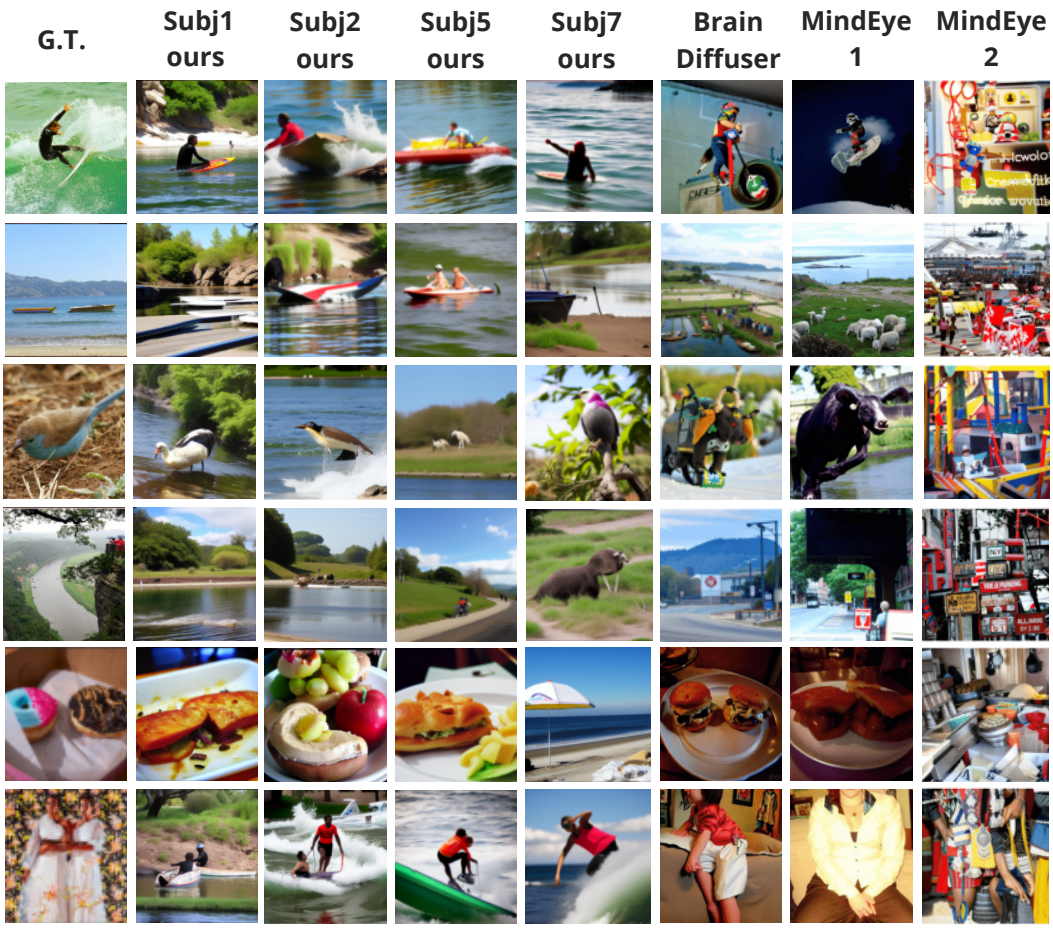

Figure 7: Qualitative results of the *top* reconstruction for all the subjects of imagined complex stimuli. Our results show a strong semantical consistency between the ground truth and reconstructed images, improving the performance obtained by other SOTA models.

Figure 8: Qualitative results of the *top* reconstruction for all the subjects of imagined conceptual stimuli. Our results show a strong semantical consistency between the ground truth and reconstructed images, improving the performance obtained by other SOTA models.

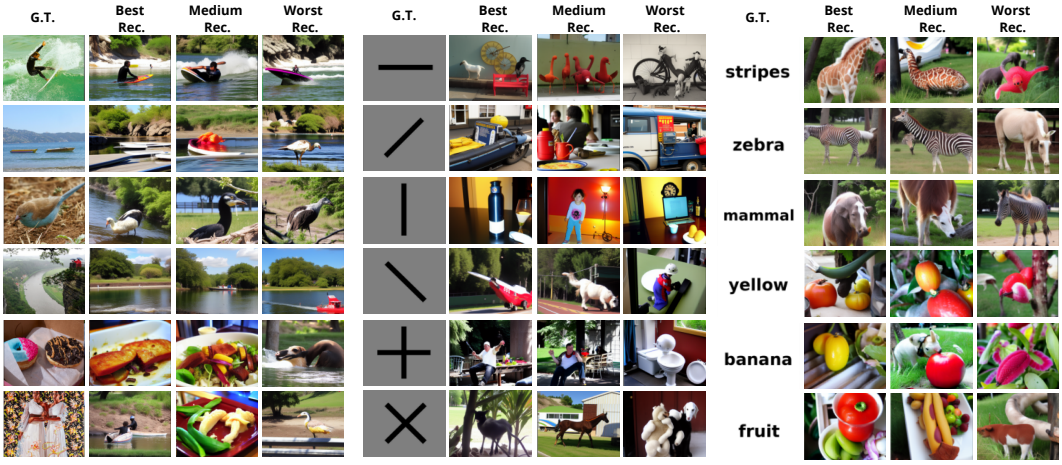

Figure 9: Best, medium and worst qualitative results of reconstruction for subject 1 for imagined complex, simple and conceptual stimuli from visual cortex.

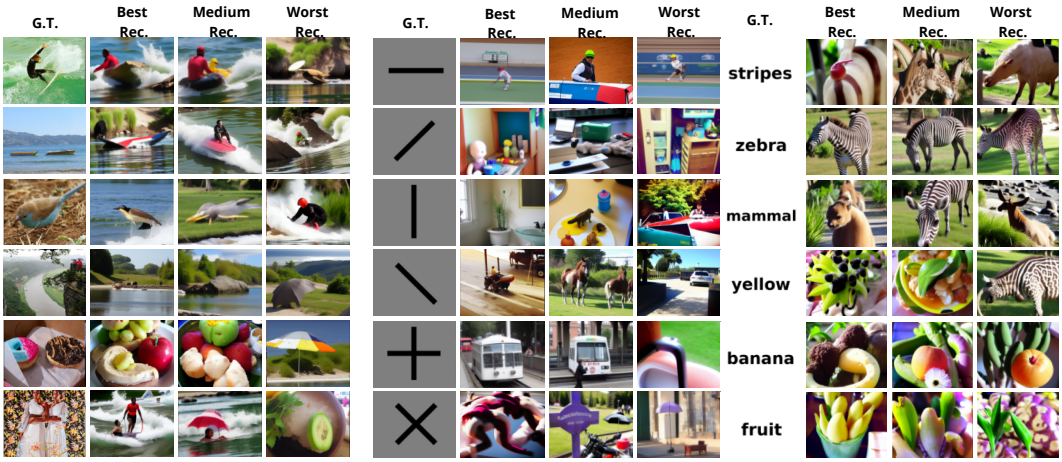

Figure 10: Best, medium and worst qualitative results of reconstruction for subject 2 for imagined complex, simple and conceptual stimuli from visual cortex.

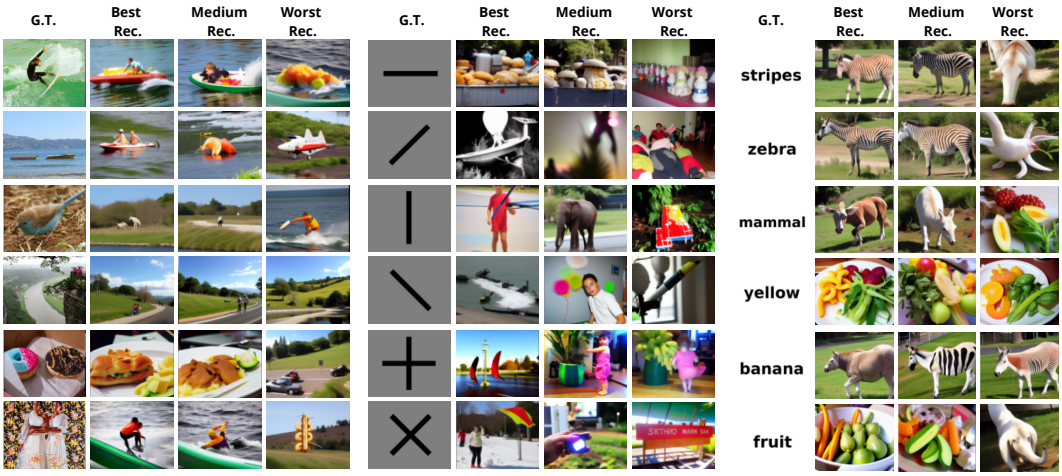

Figure 11: Best, medium and worst qualitative results of reconstruction for subject 5 for imagined complex, simple and conceptual stimuli from visual cortex.

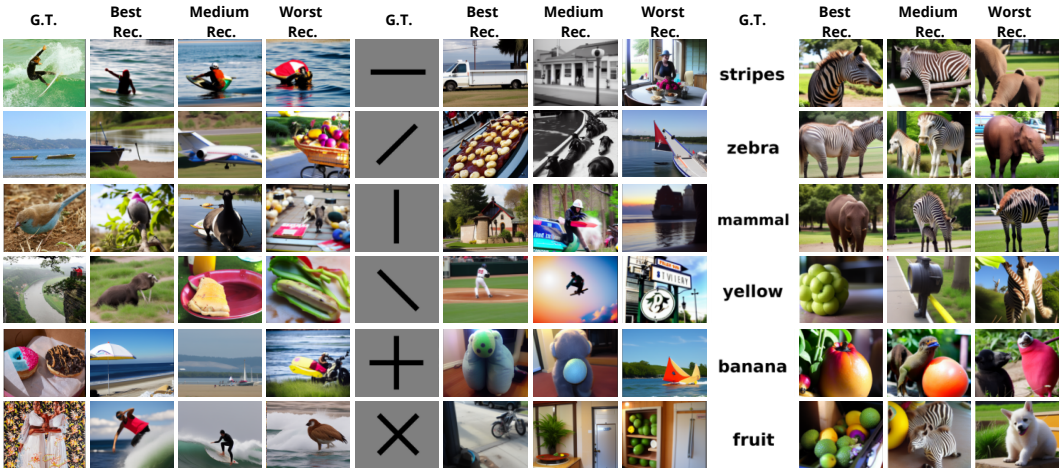

Figure 12: Best, medium and worst qualitative results of reconstruction for subject 7 for imagined complex, simple and conceptual stimuli from visual cortex.