BRIDGING VISION AND IMAGERY: LATENT FUNCTIONAL ALIGNMENT FOR VISUAL IMAGERY DECODING FROM FMRI DATA

Anonymous authors
Paper under double-blind review

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 NSD-Imagery benchmark. We propose a latent functional alignment (LFA) approach that maps imagery-evoked activity to the pretrained model's semantically enriched conditioning space, by adding a simple alignment module, while keeping the original remaining components frozen. To mitigate the limited amount of matched imagery–perception supervision, we further introduce a neural retrieval-based augmentation strategy that selects semantically related NSD perception trials from the same participants. Across four subjects, LFA 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 at the intersection of neuroscience and artificial intelligence (AI), with recent advances enabling the reconstruction of visual stimuli from neural activity and opening new perspectives for brain-computer interfaces (BCIs). Vision has received particular attention due to relatively controlled experimental settings and the increasing availability of large-scale datasets. Neural signals can be acquired using invasive techniques, such as electrocorticography (ECoG) and implanted microelectrodes, or non-invasive methods, including functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG). Although invasive recordings generally provide higher signal-to-noise ratios (SNR), non-invasive techniques offer substantially lower risks. In particular, the release of the Natural Scenes Dataset (NSD) (Allen et al., 2022) demonstrated that competitive visual decoding can be achieved from fMRI using AI-based pipelines for image reconstruction.

Visual brain decoding typically maps noisy and high-dimensional brain data to natural images, often restricting the input to regions of interest (ROIs) in the visual cortex to reduce dimensionality. Neural activity can then be mapped into semantic embedding spaces learned by large pretrained models (Radford et al., 2021), which provide conditioning signals for diffusion-based generative models.

In this work, we extend the SOTA visual perception decoding model, DynaDiff (Careil et al., 2025), to mental imagery. Mental imagery is a fundamental cognitive process through which individuals generate internal visual representations of past or future events. However, decoding imagery from brain activity is substantially more challenging than decoding visual perception, due to its lower SNR (Roy et al., 2023) and the limited availability of large-scale imagery datasets.

Recent results on the newly released NSD-Imagery benchmark dataset (Kneeland et al., 2025), which extends previous imagery-related dataset (Shen et al., 2019; Lee & Kuhl, 2016; Senden et al., 2018), show that models trained exclusively on visual perception data can produce moderately faithful reconstructions of imagined stimuli at inference time. However, existing open-source SOTA

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

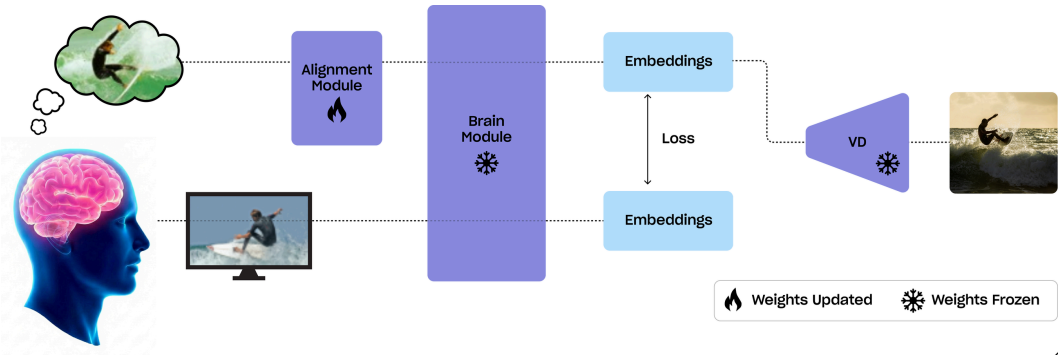


Figure 1: Overview of our pipeline. CLIP-Image embeddings predicted by the pretrained visual decoding model from vision trials are used as targets for the alignment module. The module minimizes the loss between embeddings predicted from imagery trials and their corresponding visual targets, while all other components remain frozen.

models exhibit inconsistent performance, likely due to the distribution shift between perception and imagery data and increased inter-subject variability. While semantic concepts may be shared across individuals, their internal visual representations can differ substantially, limiting the ability of existing pipelines to produce coherent reconstructions.

To address this distribution shift, we propose a strategy that encourages the model to exploit semantic representations shared between perception and imagery. We leverage the NSD-Imagery dataset together with the substantially larger NSD visual-perception dataset used to train the original model. Specifically, NSD visual-task neural data provide supervision targets for aligning the representations predicted from the imagery data, thereby leveraging the larger amount of available visual data to mitigate the limited size of the imagery dataset.

Finally, we systematically analyze the contribution of multiple cortical areas to assess whether the proposed alignment generalizes beyond the visual cortex. Our evaluation shows consistent improvements in semantic reconstruction metrics, demonstrating the effectiveness of the proposed approach for mental-imagery decoding. An overview of the proposed LFA pipeline is shown in Fig. 1.

1.1 Related Work

Remarkable progress in visual brain decoding from fMRI data has been enabled by the availability of large-scale neural datasets (Horikawa & Kamitani, 2017; Chang et al., 2019; Allen et al., 2022), pretrained vision-language models such as CLIP (Radford et al., 2021), and open-source diffusion models including Stable Diffusion (SD) (Rombach et al., 2022) and Versatile Diffusion (VD) (Xu et al., 2024).

Current SOTA visual brain decoding approaches typically map fMRI activity into high-dimensional CLIP-Image and/or CLIP-Text embedding spaces, which are subsequently used to condition pretrained diffusion models. Earlier approaches generally rely on multi-stage pipelines and linear models to reduce the temporal dimension of fMRI signals. Among these, MindEye1 (Scotti et al., 2023), BrainDiffuser (Ozcelik & VanRullen, 2023), and MindEye2 (Scotti et al., 2024) constitute relevant baselines for visual decoding and have also been evaluated in Kneeland et al. (2025).

MindEye1 employs separate semantic and perceptual pathways: fMRI activity is mapped to CLIP-Image embeddings for semantic reconstruction, while a second pathway predicts Stable Diffusion VAE (Kingma & Welling, 2022) latents to recover coarse perceptual structure and guide diffusion. BrainDiffuser similarly follows a two-stage framework, first predicting hierarchical VDVAE (Child, 2020) latents for coarse image reconstruction and then refining the result with Versatile Diffusion conditioned on CLIP-Image and CLIP-Text representations predicted from fMRI. MindEye2 extends this paradigm through multi-subject pretraining followed by target-subject fine-tuning, while jointly modeling semantic and perceptual information and incorporating predicted image captions for textual guidance during reconstruction refinement.

More recently, DynaDiff (Careil et al., 2025) introduced a single-stage framework that decodes images directly from fMRI time series. Its lightweight brain module applies a subject-specific projection, transformations specific to each fMRI time point, and temporal aggregation, mapping the

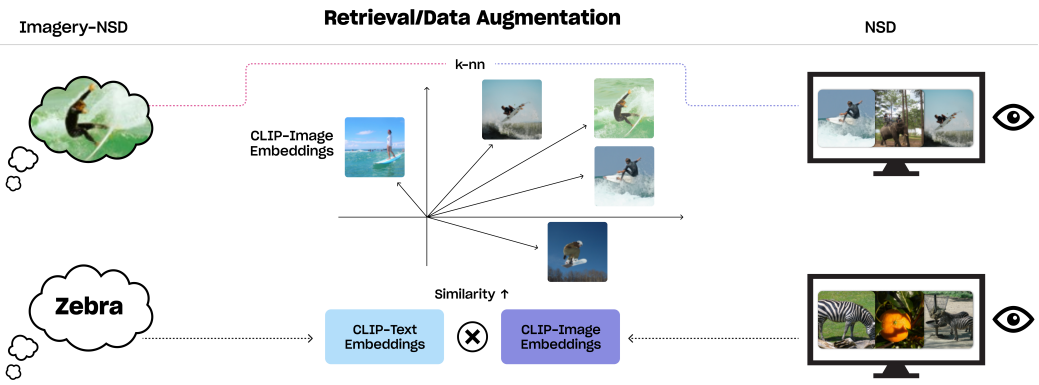


Figure 2: Illustration of the retrieval procedure. For complex stimuli, we compute CLIP-Image embeddings and retrieve nearest neighbors from NSD via k-nn in cosine space, using the associated NSD brain data as targets. For conceptual stimuli, we compute CLIP-Text embeddings for the target words and retrieve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

neural time series to embeddings with the dimensionality of the image-conditioning tokens of VD. These embeddings replace the CLIP-Image tokens in the cross-attention layers of the U-Net (Rombach et al., 2022), and the brain module is optimized jointly with a LoRA adapter (Hu et al., 2021). Despite its strong performance for visual perception, the ability of DynaDiff to generalize to mental imagery remains unexplored. We investigate this setting through LFA on NSD-Imagery.

2 METHODS

2.1 DATA AUGMENTATION

One of the main challenges of the NSD-Imagery dataset is its limited amount of data available to construct a robust and efficient pipeline using only imagery data: each subject of the NSD performed additional dedicated runs on vision and imagination tasks on different categories of stimuli; these stimuli are divided into three different blocks, each consisting of six stimuli, categorized as *simple*, *complex*, and *conceptual* stimuli, providing a total of 96 imagery trials per block, and 48 trials for the associated vision run; more details about the dataset can be found in Appendix A.1 and in Kneeland et al. (2025).

To overcome this difficulty, we retrieved additional fMRI data from NSD trials, focusing on complex and conceptual stimuli; in line with previous works, we focused only on subjects $1, 2, 5$ and 7 since they completed all 40 NSD sessions.

For each of the six complex stimuli we computed its CLIP-Image embedding with CLIP ViT-B/32 and performed a nearest-neighbour search, in cosine distance, over the image embeddings of the NSD training split of the same subject (about 27,000 trials of 9,000 images, each shown three times). Five of the six complex stimuli belong to the NSD *shared*1000 test-set, which is excluded from this split, and the sixth does not appear in NSD, so neither the target images nor their associated NSD trials can be retrieved. We retained the corresponding K nearest neural vision trials per stimulus and used them as visual targets. Reconstruction quality peaked at $K = 1000$ for subject 1, as shown in Appendix B; therefore we used $K = 1000$ for all four subjects.

After increasing the number of visual targets, each visual target had to be paired with an imagery input. Therefore, for each stimulus, the imagery training trials were replicated until their number matched that of the retrieved visual training targets. To avoid exact duplicates, Gaussian noise with standard deviation 0.02σ was added to each copy beyond the original trials, where σ is the standard deviation of the training trials for that stimulus. This replication was applied only to the training and validation data. A visual representation of our retrieval procedure from NSD is shown in Fig. 2. For conceptual stimuli, we embedded each target word with the CLIP-Text-encoder from the same CLIP ViT-B/32 model and retrieved the $K = 1000$ NSD training vision trials whose image embeddings had the highest cosine similarity to it.

Since simple stimuli lie outside the NSD distribution, no alignment module was trained for them and imagery responses were decoded directly by the pretrained model.

2.2 LATENT FUNCTIONAL ALIGNMENT

A natural first approach is to learn, for each subject, a linear mapping from imagery to perception responses directly in voxel space, analogous to the mappings used for across-subject functional alignment (Ferrante et al., 2024; Wang et al., 2025; Haxby et al., 2020). However, this approach aims to predict, in a very noisy and high-dimensional space, fine details that may not capture the relationship between these two different cognitive processes. Therefore, our strategy relies on the hypothesis that mental images are typically less detailed and more variable than percepts, while their semantic content is more stable (Reddy et al., 2010; Pearson et al., 2015): an alignment performed in the embedding space that conditions the diffusion model could be more effective for this particular task, rather than capturing detailed differences in the voxel space. Following this hypothesis, imagery responses are passed through an alignment module f_ϕ placed before the frozen brain module g of DynaDiff. The module first applies a learned linear mixing of the $T = 6$ time points, shared across voxels and initialized to the identity, since stimulus information might not be uniformly distributed across time points; it then applies a multilayer perceptron with one hidden layer, with weights shared across time points, which maps the input voxels to the the number of *nsdgeneral* voxels V_{vis} expected by g. The alignment module therefore operates in voxel space, but is trained exclusively through a latent objective, while the original brain module and the diffusion model remain frozen.

Given an imagery input $\mathbf{x}$, paired with a perceptual response $\mathbf{v}$ to one of the K NSD vision trials retrieved for that stimulus, we compare $\mathbf{e} = g(f_\phi(\mathbf{x}))$ with $\mathbf{e}^{\text{vis}} = g(\mathbf{v})$, i.e. the embedding that the frozen brain module predicts from the perception response, minimizing

$$\mathcal{L} = \mathcal{L}_{\text{MSE}}\big(\mathbf{e}, \mathbf{e}^{\text{vis}}\big) + \lambda\, \mathcal{L}_{\text{SupCon}}\big(\mathbf{e}, \mathbf{e}^{\text{vis}}; \tau\big), \qquad \lambda = 0.1,\ \tau = 0.05. \tag{1}$$

The $\mathcal{L}_{MSE}$ term regresses each input towards the mean target of its stimulus; with weakly informative imagery inputs, predictions for different stimuli drift towards a common embedding. The supervised contrastive term (Khosla et al., 2020) counteracts this collapse: each imagery embedding acts as an anchor, the perception embeddings of the same stimulus in the batch act as positives and those of the other stimuli as negatives, so that each prediction must be closer to the targets of its own stimulus than to those of the others. More details on the training procedure can be found in Appendix A.2.

2.3 EVALUATION

In brain decoding there is no single metric that fully captures the quality of reconstructed images. We therefore evaluate complementary aspects of reconstruction using low-level, high-level, and semantic metrics, motivated by the hierarchical correspondence between deep neural networks (DNNs) and the human visual system (Kriegeskorte, 2015; Doerig et al., 2025). Low-level metrics, PixCorr and SSIM, measure pixel-wise correlation and structural similarity, respectively (Wang et al., 2004), providing an estimate of local visual fidelity but not of image meaning. High-level metrics are computed from intermediate or final representations of pretrained DNNs and assess correspondence at increasingly abstract levels beyond pixel-level similarity. Specifically, AlexNet-2 and AlexNet-5 use layers 2 and 5 of AlexNet (Krizhevsky et al., 2012), CLIP uses the output of the ViT-L/14 CLIP-Vision model (Radford et al., 2021), and Inception uses the last pooling layer of InceptionV3 (Szegedy et al., 2015). EffNet-B and SwAV are distance-based metrics derived from EfficientNet-B1 (Tan & Le, 2020) and SwAV-ResNet50 (Caron et al., 2021). For the identification-based metrics, each reconstructed image is compared with the correct target and with 11 other distractors, following the evaluation protocol of Kneeland et al. (2025). The 11 distractors are selected from the remaining simple and complex stimuli. Performance is the proportion of comparisons in which the correct target is more similar to the reconstruction than the distractor; therefore, chance level is 50%. Since our approach aims to improve the semantic correspondence between imagery-based reconstructions and target stimuli rather than their low-level appearance, high-level metrics provide the most informative evaluation of our LFA. We consider two baselines: first, we use the pretrained DynaDiff model, providing a reference for the generalization of the original visual decoding pipeline to imagery data. Second, as a baseline for functional alignment, we use cross-validated Ridge regression to predict NSD-retrieved visual-task fMRI activity from imagery-task fMRI activity in the voxel space.

3 RESULTS

In this section, we report the results of our LFA pipeline between mental imagery and vision. As shown in Fig. 3 for subject 1, our reconstructions are semantically coherent with the corresponding targets. The reconstructions shown were selected by visual inspection as the best among 10 different partitions, in order to compare them with Kneeland et al. (2025). For quantitative evaluation, no selection is performed: one reconstruction is generated for each of the 10 random partitions with a fixed sampling seed, and metrics are averaged across partitions, reported as mean ± standard deviation (SD) for single subjects and mean ± standard error of the mean (SEM) across subjects. Qualitatively, compared to the reconstructions of other open-source models (Ozcelik & VanRullen, 2023; Scotti et al., 2023; 2024) reported in Kneeland et al. (2025), our results appear more photorealistic and semantically coherent. Although fine details may be lost, the semantic content is largely preserved, which is consistent with the hypothesis that mental images retain their conceptual content more reliably than their low-level characteristics. Quantitative results and qualitative reconstructions for all subjects and stimulus classes, from the visual cortex, are reported in Appendix D and Appendix E.

3.1 VISUAL BRAIN REGION

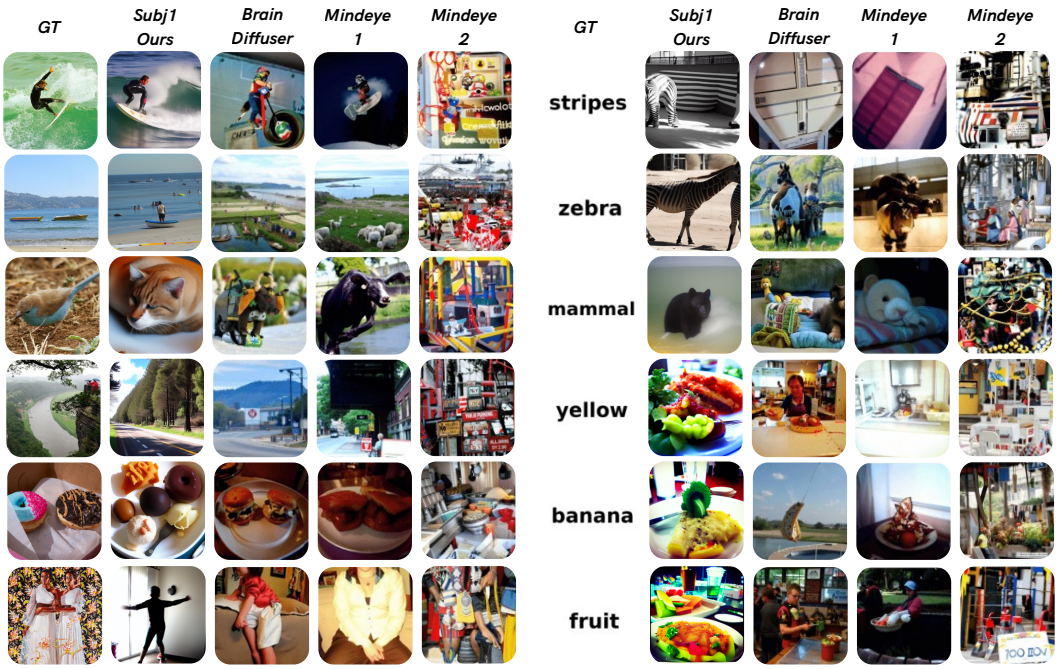


Figure 3: Qualitative results of the *best* visually inspected reconstruction for subject 1 of imagined complex (left) and conceptual stimuli (right). Our results show a strong semantic consistency between the ground truth and reconstructed images, visually improving the performance obtained by other SOTA models.

We report in Tab. 1 the metrics averaged across the four subjects, computed only for complex stimuli, since no alignment is applied to simple stimuli, whose reconstructions are therefore identical to the baseline. For each subject and partition, 20% of the imagery trials of each stimulus were held out and averaged into a single test input, in order to increase its SNR, yielding six test reconstructions per stimulus set. With respect to the DynaDiff baseline, LFA improves all high-level and distance metrics, as well as PixCorr. Fig. 4 shows AlexNet-5, CLIP and Inception scores before and after LFA for each subject; significance was assessed with a one-sided Wilcoxon signed-rank test over the 10 partitions (Tab. 6 in Appendix C). From Tab. 1, functional alignment also obtains strong high-level scores, but these can be inflated by representations that capture shared semantic structure across stimuli. We therefore investigate stimulus specificity separately, computing the similarity

Table 1: Quantitative reconstruction metrics for complex stimuli averaged across four subjects; uncertainties are obtained with SEM. Arrows indicate whether higher (↑) or lower (↓) values correspond to better performance.

Method	Low-Level		High-Level				Distance	
	PixCorr ↑	SSIM ↑	AlexNet-2 ↑	AlexNet-5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
NSD-Imagery Trials – Complex stimuli – Subject-averaged – Visual Cortex								
DynaDiff Baseline	-0.004 ± 0.010	0.241 ± 0.003	$76.86 \pm 1.64\%$	$55.15 \pm 2.21\%$	$67.01 \pm 2.41\%$	$53.33 \pm 2.52\%$	0.961 ± 0.004	0.622 ± 0.003
Functional Alignment	-0.006 ± 0.010	0.189 ± 0.011	$77.05 \pm 1.21\%$	$\mathbf{68.90 \pm 3.25\%}$	$72.12 \pm 0.86\%$	$64.36 \pm 1.18\%$	0.916 ± 0.005	$\mathbf{0.581 \pm 0.009}$
LFA	$\mathbf{0.035 \pm 0.007}$	$\mathbf{0.249 \pm 0.008}$	$\mathbf{80.98 \pm 4.23\%}$	$67.20 \pm 5.86\%$	$\mathbf{78.83 \pm 5.06\%}$	$\mathbf{67.42 \pm 5.65\%}$	$\mathbf{0.908 \pm 0.020}$	0.583 ± 0.017

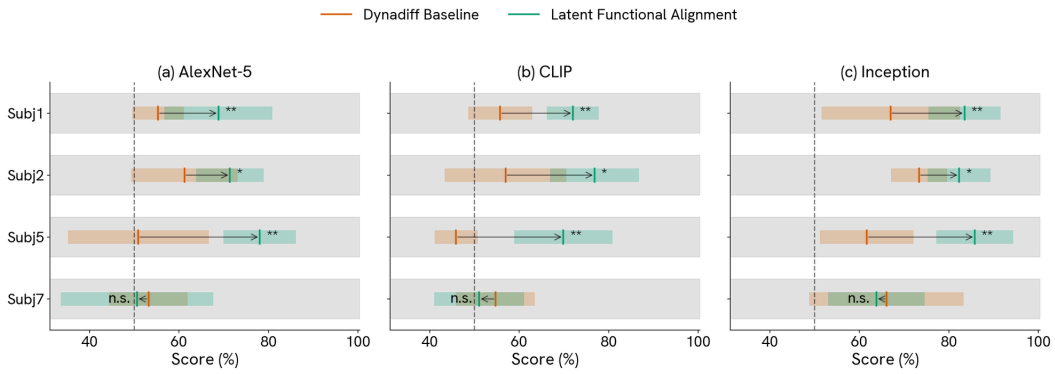


Figure 4: Per-subject AlexNet-5, CLIP and Inception scores for complex stimuli, with the corresponding significance levels. The gray vertical dashed line denotes chance level for each metric. DynaDiff baseline performs close to chance, whereas LFA improves performance across subjects, with the exception of subject 7, which shows no significant improvement. $*$ ($p < 0.05$), $**$ ($p < 0.005$).

between CLIP-Image embeddings of the generated images and the CLIP-Text embeddings from the COCO captions (Lin et al., 2015) of each of the 5 stimuli from the NSD *shared*1000 set, while the remaining caption was hand-written. Results are reported in Fig. 5. For conceptual stimuli, visual

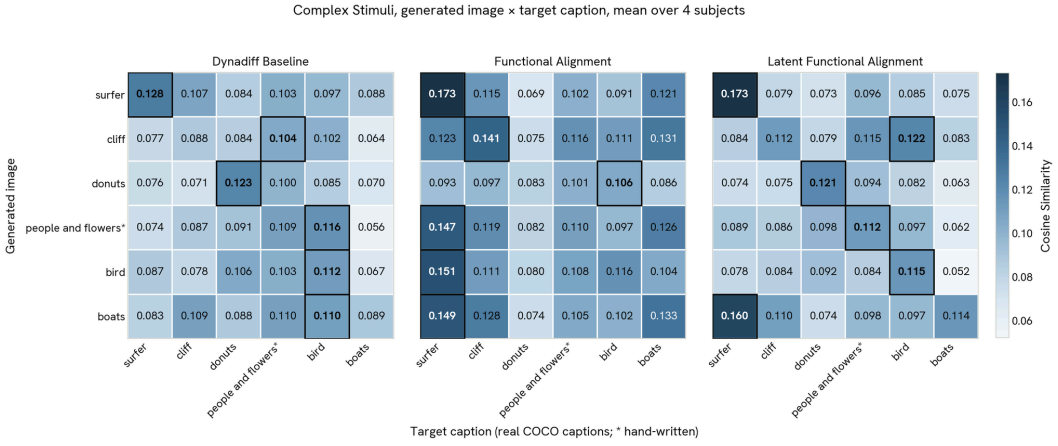


Figure 5: Subject-averaged heatmaps for the three different methods reporting cosine similarity between CLIP-Text embeddings and CLIP-Image embeddings from reconstructions of complex stimuli. Functional alignment, despite the performance reported in Tab. 1, shows substantial overlapping, indicating a collapse in the generation process, especially for stimuli related to the presence of water, with only two terms along the diagonal indicating maximum correspondence. In contrast, LFA improves over the DynaDiff baseline along the diagonal, with four row maxima on the diagonal against three for the baseline, while preserving the separation among different concepts.

targets were retrieved from NSD as described in Section 2.1. Since no ground-truth image exists,

we generated a caption for each reconstruction with BLIP (Li et al., 2022) and computed the cosine similarity between the MPNet (Song et al., 2020) sentence embeddings of the captions and of the target words. Results are shown in Tab. 2.

Table 2: Cosine similarity between BLIP-generated captions and target words for conceptual stimuli in MPNet sentence embeddings space. Values are reported as mean ± standard deviation across the ten partitions for each subject and as mean ± SEM across subjects.

Method	Subj1	Subj2	Subj5	Subj7	All Subjects (Mean ± SEM)
NSD-Imagery Trials – Conceptual Stimuli – Text Similarity – Visual Cortex					
DynaDiff Baseline	0.109 ± 0.050	0.119 ± 0.048	0.082 ± 0.036	0.086 ± 0.028	0.099 ± 0.009
Functional Alignment	0.184 ± 0.040	$\mathbf{0.223 \pm 0.044}$	0.183 ± 0.034	$\mathbf{0.163 \pm 0.053}$	0.188 ± 0.012
LFA	$\mathbf{0.187 \pm 0.027}$	0.156 ± 0.041	$\mathbf{0.190 \pm 0.042}$	0.152 ± 0.035	0.171 ± 0.010

As for complex stimuli, functional alignment reaches a high similarity with the target words, even slightly higher than LFA. However, Tab. 3 shows that this similarity is not specific to the target: since the six words form two semantic groups (*banana*, *yellow*, *fruit*; *zebra*, *stripes*, *mammal*), a generic reconstruction can match several words at once.

Table 3: Difference between the mean cosine similarity of diagonal and non-diagonal terms, weighted by the inverse number of corresponding terms, in the text-similarity matrix for the DynaDiff baseline, functional alignment, and LFA on conceptual stimuli. LFA achieves the highest specificity.

Method	Subj1	Subj2	Subj5	Subj7	All Subjects (Mean ± SEM)
NSD-Imagery Conceptual Stimuli - Specificity - Visual Cortex					
DynaDiff Baseline	0.039 ± 0.043	0.050 ± 0.031	0.007 ± 0.033	0.014 ± 0.030	0.028 ± 0.010
Functional Alignment	0.035 ± 0.038	0.074 ± 0.045	0.038 ± 0.033	0.012 ± 0.052	0.040 ± 0.013
LFA	$\mathbf{0.073 \pm 0.026}$	$\mathbf{0.075 \pm 0.031}$	$\mathbf{0.079 \pm 0.040}$	$\mathbf{0.040 \pm 0.047}$	$\mathbf{0.067 \pm 0.009}$

3.2 Role of non-visual brain regions

The mechanisms and the cortical regions involved in mental imagery are still debated (Abraham & Bubić, 2015), and studies on the role of the visual cortex have yielded mixed findings across studies and individuals (Xie et al., 2020; Dijkstra, 2024; Bridge et al., 2012; Milton, 2024; Dijkstra et al., 2017; Ganis et al., 2004). Here we compare quantitatively the contribution of the main cortical areas, for both complex and conceptual stimuli.

We selected each ROI from the HCP−MMP1 atlas in the functional space of each subject, with the same resolution and preprocessing used for the *nsdgeneral* ROI (Appendix A.1), giving, as inputs, the voxels indicated in Tab. 4 in Appendix A.2. We then repeated the analysis on the union of each ROI with *nsdgeneral*, to measure the interplay between each region and posterior cortex. Note that no unaligned baseline exists for non-visual regions, since the pretrained brain module only accepts *nsdgeneral* voxels; every value reported here is obtained with LFA from the corresponding ROI, and the two baselines shown for reference are DynaDiff and LFA, both from visual cortex, with the values reported in Section 3.1. The results are averaged across the four subjects. For complex stimuli, as depicted in Fig. 6, every ROI scores above the DynaDiff baseline, being significant in all three metrics only for the parietal and temporal cortices, as reported in Appendix C. Adding the *nsdgeneral* ROI to each region brings all of them close to the values obtained from the LFA visual cortex baseline; the only region whose performance is not improved by this addition is the parietal cortex for CLIP. Hence, for complex stimuli, no non-visual ROI yields a consistent gain over visual cortex alone, in line with previous neuroscientific work (Stokes et al., 2009; Reddy et al., 2010; Naselaris et al., 2014).

We investigated possible contributions distributed across the cortex also for conceptual stimuli. Fig. 7 reports caption-to-word similarity in MPNet space and specificity described in Section 3.1. Each ROI on its own lies closer to the LFA visual-cortex value than to the DynaDiff baseline, and the temporal cortex shows the smallest gap. The contribution of *nsdgeneral* ROI increases performance

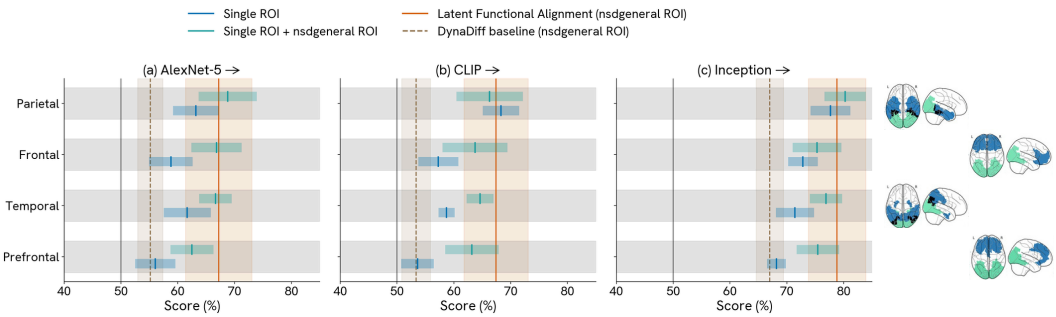


Figure 6: Bar plot averaged across the four subjects for AlexNet-5, CLIP and Inception for complex stimuli reconstructions. Results are obtained from the single ROI voxels (blue bars) and their combination with the *nsdgeneral* ROI (greenish bars). The comparison is performed both on DynaDiff Baseline (vertical brown dashed line) and LFA (vertical orange line) from the *nsdgeneral* ROI, taking the values reported in Tab. 1. 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.

for all regions, bringing them to values comparable to the ones of LFA visual cortex. Additional details on statistical reliability of our analysis can be found in Appendix C.

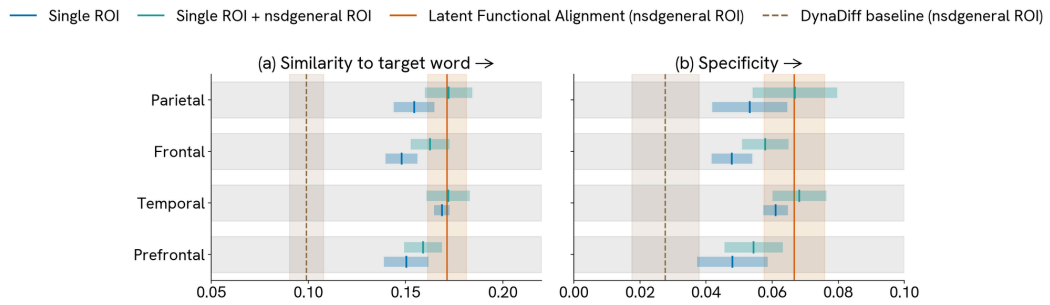


Figure 7: Panel (a): cosine similarity between BLIP-generated captions and target words for conceptual stimuli in MPNet sentence embeddings space for different ROIs. Both values related to single areas and with the addition of the *nsdgeneral* ROI are above DynaDiff baseline (vertical brown dashed line). Panel (b): difference between the mean of the diagonal and non-diagonal terms, weighted by the inverse of their respective number of terms, for each ROI and its combination with the *nsdgeneral* ROI. Adding the *nsdgeneral* ROI improves performance for all other ROIs, bringing them close to LFA (vertical orange line).

4 Discussion

In this work, we investigated brain decoding from visual imagery by analyzing the contribution of distinct cortical regions. Compared to visual perception, imagery is a more challenging decoding target, as it is driven by internally generated working-memory and top-down processes rather than by direct sensory input (Dentico et al., 2014; Ganis et al., 2004).

As an initial step, we focused on visual cortical regions, leveraging a pretrained perception-model (Careil et al., 2025), following the procedure of Kneeland et al. (2025) and previous research, attributing to visual cortex a major role in visual imagery (Dijkstra et al., 2017; Ganis et al., 2004). The proposed LFA framework extracts semantically meaningful representations from imagery-related signals: averaged across four participants, it improves every high-level and distance metric over the unaligned model, and the improvement is significant in three subjects out of four for both stimulus classes, although not the same three: in Appendix C, subject 7 shows no improvement for complex stimuli as reported in Tab. 6, but a significant gain for conceptual stimuli, while subject 2 shows

the opposite pattern in Tab. 7. These findings are consistent with the weak and variable nature of imagery signals and with the inter-subject variability reported in previous work (Roy et al., 2023; Dijkstra et al., 2019).

The comparison with functional alignment also carries a methodological message. That approach was competitive on the standard reconstruction metrics, while producing reconstructions that were not specific to the imagined stimulus. Only measuring the match to the correct target with respect to the others revealed the difference. Therefore, metrics designed for perceptual decoding should not be used alone to assess imagery reconstructions, as they may overlook potential discrepancies with human judgments (Sinha & Russell, 2011).

Subsequently, we extended the analysis to the parietal, temporal, frontal, and prefrontal regions. The posterior cortex emerged as the primary contributor to decoding performance, consistent with whole-brain analyses in Wang et al. (2024). For complex stimuli, individual non-visual ROIs generally yielded lower scores than LFA from the visual cortex, the only exception being the parietal cortex on CLIP. All regions improved over the DynaDiff baseline, significantly on all three metrics only for the parietal and temporal cortices. Adding the *nsdgeneral* ROI improved every single region, except the parietal cortex on CLIP, bringing all of them to values comparable to those obtained from the LFA from visual cortex, without a consistent gain. One plausible explanation is that a single additional ROI does not contribute sufficiently complementary information, an effect likely amplified by the limited amount of imagery data available to learn cross-regional interactions.

For conceptual stimuli the picture differs: every single region was closer to the values obtained from the LFA visual cortex than to the DynaDiff baseline, with the temporal lobes the closest; combined with *nsdgeneral*, all regions reach values comparable to the LFA visual cortex. This suggests that the representation supporting conceptual imagery is less confined to posterior cortex than the one supporting the imagery of specific images, although our data-driven approach cannot establish this directly and more experiments are needed.

More broadly, these results are compatible with the idea that, although internally generated representations differ from perceptual ones, they remain semantically related to them. Overall, our findings are consistent with the view of visual imagery as a process that involves higher-order regions in addition to posterior cortex, particularly for conceptual content.

4.1 Limitations and Conclusion

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 neuroscientific conclusions that can be drawn about the functional role of individual ROIs; the lack of improvement for subject 7 on complex stimuli, and the non-significant improvement for subject 2 on conceptual stimuli, are the clearest illustration. Second, the alignment module is trained and evaluated on the same six stimuli per set, with held-out trials, so our results demonstrate generalization across trials and not to unseen stimuli; the correlation analysis in Appendix A.1 further characterizes this setting, showing substantial variability across imagery trials. Third, as noted in Section 3.2, non-visual ROI results cannot be compared with an unaligned counterpart of the same region. Fourth, the parietal and temporal ROIs partially overlap with the *nsdgeneral* ROI, so they are not purely non-visual; the frontal and prefrontal ROIs, which do not overlap, provide the cleanest evidence, for conceptual stimuli, that imagery-related information extends beyond posterior cortex.

Finally, while retrieval-based augmentation improved performance substantially, relying on perception data may bias the alignment towards perceptual rather than purely imagery-driven features, and for conceptual stimuli the retrieved target sets of semantically related words partially overlap, which limits the separation the model can learn.

Despite these limitations, our results show that imagery-related information can be decoded beyond posterior cortex, particularly for conceptual stimuli, while confirming the primary role of posterior cortex for visual imagery, supporting multi-ROI decoding as a promising direction for BCIs and semantic reconstruction from neural signals. Future work will explore more sophisticated ROI fusion, finer parcellations, imagery-specific evaluation metrics, and validation 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 creativity and inner experience. Such knowledge could also pave the way for future clinical applications, particularly for patients who are unable to communicate or see but retain the ability to internally evoke previously experienced scenes and concepts.

AI USE STATEMENT

In this work, we used generative AI tools for polishing text in English. We have not used generative AI tools for generating concepts and procedure ideas. We have reviewed all AI-assisted work. Part of the code was LLM-generated and was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The datasets are publicly accessible at: `https://natural-scenes-dataset.s3.amazonaws.com/index.html`. The present study makes use of the publicly available fMRI datasets Allen et al. (2022); Kneeland et al. (2025), which were collected under informed consent and institutional ethical approval. The authors did not acquire new data. All data handling adheres to the licenses and usage terms specified by the dataset providers, and no attempt was made to re-identify individuals or to infer private attributes unrelated to the stated research purposes. All the data are acquired from healthy subjects who do not suffer from aphantasia, or related clinical conditions.

However, results and research efforts in decoding human mind raise broader conceptual and ethical questions: as neural data become more accessible and AI models more powerful, the information that can be decoded from brain activity continues to expand. Decoding imagination, one of the most private aspects of human experience, offers new opportunities for human–AI co-creation and the externalization of memories, dreams, or creative intent, while raising concerns about privacy, neural rights, model bias, and interpretability. Key open questions include how to safeguard individuals against misuse and how to disentangle neural noise, subjective variability, and algorithmic bias in decoded representations.

REPRODUCIBILITY STATEMENT

To reproduce our results, we recommend following the pipeline in the following order. First, preprocess the NSD-Imagery dataset as described in Appendix A.1, to ensure compatibility with the pretrained perception model. Second, apply the data augmentation pipeline described in Section 2.1 to obtain the data required by the LFA method, whose training procedure is detailed in Appendix A.2 and Section 2.2. In order to include the other ROIs use the same parcellation of Section 3.2, from HCP−MMP1 atlas.

Finally, follow the evaluation protocol described in Section 2.3 to reproduce the results reported in Section 3.

The code to reproduce results is publicly available at: `https://github.com/anonymousiclr27/brain_decoding_imagery/tree/main`.

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`.

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.

Nadine Chang, John A. Pyles, Austin Marcus, Abhinav Gupta, Michael J. Tarr, and Elissa M. Aminoff. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, 6(1):49, May 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0052-3. URL https://doi.org/10.1038/s41597-019-0052-3.

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

Daniela Dentico, Bing Leung Cheung, Jui-Yang Chang, Jeffrey Guokas, Melanie Boly, Giulio Tononi, and Barry Van Veen. Reversal of cortical information flow during visual imagery as compared to visual perception. *NeuroImage*, 100:237–243, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2014.05.081. URL https://www.sciencedirect.com/science/article/pii/S1053811914004662.

N. Dijkstra, M. Hinne, S. E. Bosch, and M. A. J. van Gerven. Between-subject variability in the influence of mental imagery on conscious perception. *Scientific Reports*, 9(1):15658, October 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-52072-1. URL https://doi.org/10.1038/s41598-019-52072-1.

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 L Thompson, and Stephen M Kosslyn. Brain areas underlying visual mental imagery and visual perception: an fmri study. *Cognitive Brain Research*, 20(2):226–241, 2004. ISSN 0926-6410. doi: https://doi.org/10.1016/j.cogbrainres.2004.02.012. URL https://www.sciencedirect.com/science/article/pii/S0926641004000709.

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.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):15037, May 2017. ISSN 2041-1723. doi: 10.1038/ncomms15037. URL https://doi.org/10.1038/ncomms15037.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

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.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Hongmi Lee and Brice A. Kuhl. Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. *The Journal of Neuroscience*, 36:6069 – 6082, 2016. URL https://api.semanticscholar.org/CorpusID:8432024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

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.

Thomas Naselaris, Cheryl Olman, Dustin Stansbury, Kamil Ugurbil, and Jack Gallant. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*, 105, 10 2014. doi: 10.1016/j.neuroimage.2014.10.018.

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.

Leila Reddy, Naotsugu Tsuchiya, and Thomas Serre. Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*, 50(2):818–825, 2010. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2009.11.084. URL https://www.sciencedirect.com/science/article/pii/S1053811909012701.

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.

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.

Mario Senden, Thomas Emmerling, H. J. van Hoof, Martin Frost, and Rainer Goebel. Reconstructing imagined letters from early visual cortex reveals tight topographic correspondence between visual mental imagery and perception, 08 2018.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15:e1006633, 01 2019. doi: 10.1371/journal.pcbi.1006633.

Pawan Sinha and Richard Russell. A perceptually based comparison of image similarity metrics. *Perception*, 40:1269–1281, 01 2011. doi: 10.1068/p7063.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020. URL https://arxiv.org/abs/2004.09297.

Mark G. Stokes, Russell Thompson, Rhodri Cusack, and John Duncan. Top-down activation of shape-specific population codes in visual cortex during mental imagery. *The Journal of Neuroscience*, 29:1565 – 1572, 2009. URL https://api.semanticscholar.org/CorpusID:6562071.

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 Appendix

A.1 Dataset 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. All participants provided written informed consent prior to the 7T fMRI sessions, including explicit authorization for anonymized data sharing with the research community; the study protocols were approved by the Institutional Review Board (IRB) at the University of Minnesota in accordance with ethical standards for human subjects research.

The NSD-Imagery 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 (five of them taken from the NSD *shared*1000 test 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.

In our analyses, for visual-task runs we analyze each subject's activity within the *nsdgeneral* ROI, manually drawn on fsaverage and encompassing visually responsive voxels in posterior cortex (Allen et al., 2022). Additional ROIs used exclusively for the imagination task were defined by parcellating the cortex according to the HCP$-$MMP1 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 and linear detrending; finally, each voxel time series is z-scored within 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 Kneeland et al. (2025). Since the repetition time (TR) of acquisition of the NSD-Imagery is equal to 1.0 seconds, but the pretrained model expects data with TRs of 1.33 seconds, we performed per-voxel linear interpolation in order to resample them as expected from the model. As reported in the NSD-Imagery 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 NSD-Imagery data-limited and challenging to implement.

Before conducting experiments, we verified results on vision runs from NSD-Imagery for all the 4 subjects, reporting in Tab. 10 in Appendix D averaged-subjects metrics with SEM and qualitative results in Fig. 10 in Appendix E. As expected, reconstruction quality improves when using visual perception data rather than visual imagery data, further validating the preprocessing procedure.

We performed additional analyses of the correlations between brain activity patterns across trials for both the NSD and NSD-Imagery datasets, considering both visual perception and visual imagery, respectively, across all four subjects. Fig. 8 reports the Pearson correlation coefficients estimated from the neural data, revealing differences both across modalities and between datasets. Each trial is flattened into a single vector and standardized, and every pair of trials is compared by Pearson correlation; pairs are then split according to whether the two trials correspond to the same stimulus or not. Correlations are averaged in Fisher's z within a *block*, defined as one subject and one stimulus set, which gives 12 blocks for the three lower rows (690 same-stimulus and 3681 different-stimulus pairs per block for imagery–imagery); for the NSD row a block is one subject and one retrieved

image pool, and the six pools are averaged within subject first, so that the subject remains the unit of analysis in every row. Confidence intervals are computed across blocks rather than across pairs, since pairs share trials and are therefore not independent.

The first and second rows report the correlations for the visual perception task in the NSD and NSD-Imagery datasets, respectively. Pattern reliability on NSD-Imagery is lower than on NSD, possibly due to differences in the experimental settings of the two datasets and to the smaller number of vision-task trials. Different performance between the two dataset are also reported in Kneeland et al. (2025), The third row shows the correlation between visual imagery and visual perception responses in the NSD-Imagery runs; although relatively small, a measurable correlation is present, providing further support for aligning visual imagery representations with those learned from visual perception data. Finally, the last row reports the Pearson correlations between imagery trials, in which same and different stimulus-pairs are statistically indistinguishable; the NSD-Imagery perception row has a higher mean but a wide interval, reflecting the small number of perceptual trials per stimulus. This pattern is consistent with the strong variability and near-chance performance observed with the DynaDiff baseline, while also supporting the generalization of our approach across trials, given the very weak correlation between distinct imagery trials of the same stimulus.

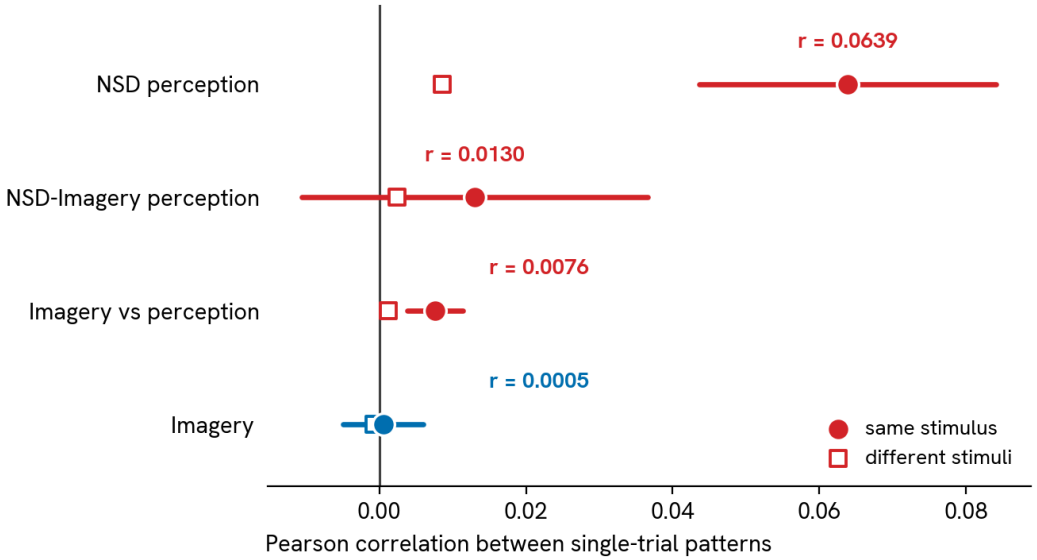


Figure 8: Pearson correlation between single trial fMRI patterns (voxels × time samples). Filled circles: pairs of trials of the same stimulus, mean and 95% confidence interval across blocks; open squares: pairs of different stimuli. The lower three rows average four subjects on the three different stimuli sets; the top row averages NSD perceptual repetitions over the four subjects. Two imagery trials of the same stimulus are uncorrelated, $r = 0.0005$, whereas two perceptual repetitions of the same NSD image reach $r = 0.0639$.

A.2 Details on training procedure

During the training of the alignment module we used an AdamW optimizer with a learning rate of 10^{-3} and weight decay 10^{-2}, with a batch size of 36; all the other modules of the original architecture were kept frozen. Training ran for at most 30 epochs with early stopping after 8 epochs without improvement, and the retained model was the one maximising the difference of the mean of the diagonal and the mean of the off-diagonal entries of the cosine-similarity matrix between the per-stimulus centroids of predicted and target embeddings.
The imagery data were split by trial, reserving 20% for test and 15% of the remainder for validation; the augmented targets were split independently, 90% for training and 10% for validation, keeping the groups of three repetitions of the same NSD image such that no repetition is shared between the two sets. The alignment module, both for the visual cortex and for the individual ROIs, consists of the temporal mixing stage followed by a multilayer perceptron with a single hidden layer of 8000

units, layer normalization, *GELU* activation and dropout with probability 0.25. 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 Tab. 4, while T is the number of time points, with $T = 6$.

The first transformation is a temporal mixing stage, introduced because the subsequent modules process the T time points as independent observations sharing the same weights, whereas information about the imagined content might not be uniformly distributed across them. Let $\mathbf{W} \in \mathbb{R}^{T\times T}$ be a learnable matrix; the input is remapped as:

$$\tilde{X}_{(b,v,t)} = \sum_{s=1}^{T} W_{t,s} X_{b,v,s}, \qquad \tilde{\mathbf{X}} \in \mathbb{R}^{B\times V\times T},$$

so that every output time point is a learned linear combination of the T input ones, with the same weights for every voxel. The operation is implemented as a one-dimensional convolution with T input channels, T output channels and kernel size 1 along the voxel axis, without bias: a wider kernel would not be meaningful, since voxels that are adjacent in the flattened vector are not adjacent in the brain. The matrix is initialized to the identity, $\mathbf{W}^{(0)} = \mathbf{I}_T$, so that at step 0 the model coincides exactly with the one without temporal mixing.

The second transformation can be represented as a multidimensional function that maps the mixed voxel data from a dimensional space V to a vector $\mathbf{h}$ of dimension $L = 8000$, applied identically to each of the T time points. Next, we performed standard layer normalization, rescaling $\mathbf{h}$ with its mean value and variance, obtaining $\hat{\mathbf{h}}$. Then we applied the non linear transformation $\mathbf{z} = \mathrm{GELU}(\hat{\mathbf{h}})$, followed by dropout with probability $p = 0.25$, and projected the result into a V_{vis} dimensional space, as expected from the pretrained model.

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: $\mathcal{D} : \mathbb{R}^{B\times V_{vis}\times T} \to \mathbb{R}^{B\times 3\times 512\times 512}$.

Table 4: Number of voxels in different ROIs for each subject.

Subject	nsdgeneral	prefrontal cortex	frontal cortex	temporal lobes	parietal cortex
1	15724	14969	10553	12166	14200
2	14278	15995	11360	12487	12494
5	13039	12293	9792	10811	11408
7	12682	12107	8760	10240	10430

B Ablation Studies

We conducted ablation experiments to assess the individual and combined contributions of our two main components: data augmentation and LFA. All experiments were conducted on Subject 1 and evaluated on complex stimuli. We first evaluated the effect of varying the number of selected NSD samples, K, using our proposed LFA approach. Since each image was presented three times to each subject in the NSD, there are three neural data for each image.

We considered the three high-level semantic metrics AlexNet-5, CLIP and Inception. Fig. 9 shows the performance profile of each metric across the evaluated values of K, together with their mean profile (black line). For each metric m and value of K, we measured the relative change with respect to the DynaDiff baseline, $\Delta_{K,m}$, and normalized it by the maximum variation observed for that metric:

$$s_{K,m} = \frac{\Delta_{K,m}}{\max_{K'}(\Delta_{K',m})}. \tag{2}$$

Here, $s_{K,m}$ denotes the normalized score of metric m at the selected value of K. Based on the mean profile across the three semantic metrics, we selected $K = 1000$ as the reference value for our augmentation strategy. This value corresponds to the maximum of the mean profile, after which

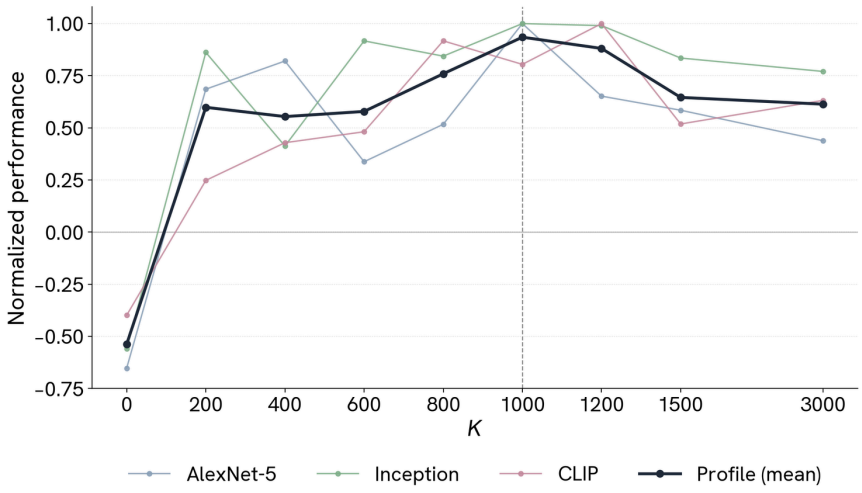


Figure 9: Performance profiles of the three high-level semantic metrics obtained with LFA relative to the DynaDiff baseline, together with their mean (black line). Values represent the normalized variation in performance across the evaluated values of K.

performance starts to decrease. After fixing K, we ran two further analyses in Tab. 5. In panel (a), the functional alignment fitted on the retrieved targets (second row) improves over the frozen decoder (first row), and LFA on the same targets gives the best results (fourth row). Without retrieval, LFA is numerically below the frozen decoder on all three metrics (third row), consistent with $K = 0$ in Fig. 9: with such few targets the module likely overfits, so retrieval-based augmentation is necessary rather than merely beneficial. Panel (b) isolates the role of non-linearity. The linear model replaces the hidden layer, the GELU and the dropout with a single linear map from the input voxels to V_{vis}, still trained with the latent objective of equation 1. Adding the temporal mixing keeps the model linear, as a kernel-size-one convolution is itself a linear operator; it accounts for most of the gain on Inception, while the non-linearity accounts for most of the gain on AlexNet-5 and CLIP. Notably, with the same linear function class, the latent objective does not outperform the voxel-space ridge on AlexNet-5 and CLIP: its advantage on these metrics emerges only in combination with the non-linearity.

Table 5: Ablation study of augmentation and alignment module architecture on Subject 1 for complex stimuli with $K = 1000$ and 10 partitions. Values are reported as mean $\pm$ standard deviation across partitions. Panel (a) compares alignment methods and supervision targets; panel (b) investigates the role of non-linearity in the alignment module. Arrows indicate whether higher (↑) or lower (↓) values correspond to better performance.

(a) Augmentation × LFA

Augmentation	**LFA**	**High-Level**		
		AlexNet-5 ↑	Incep ↑	CLIP ↑
No	No	$55.30 \pm 5.77\%$	$66.97 \pm 15.43\%$	$55.76 \pm 7.17\%$
Yes	No	$67.73 \pm 11.52\%$	$71.06 \pm 8.31\%$	$64.24 \pm 7.08\%$
No	Yes	$46.52 \pm 12.80\%$	$57.73 \pm 12.65\%$	$47.73 \pm 13.37\%$
Yes	Yes	$\mathbf{68.79 \pm 12.08\%}$	$\mathbf{83.48 \pm 8.03\%}$	$\mathbf{71.97 \pm 5.77\%}$

(b) Linearity vs Non-Linearity

Augmentation	**Architecture**	**High-Level**		
		AlexNet-5 ↑	Incep ↑	CLIP ↑
Yes	Linear	$54.09 \pm 12.35\%$	$75.15 \pm 8.02\%$	$60.00 \pm 11.70\%$
Yes	Linear + Conv	$60.15 \pm 10.00\%$	$81.67 \pm 6.52\%$	$61.52 \pm 10.26\%$
Yes	Non-Linear + Conv	$\mathbf{68.79 \pm 12.08\%}$	$\mathbf{83.48 \pm 8.03\%}$	$\mathbf{71.97 \pm 5.77\%}$

C STATISTICAL TESTS

Table 6: Subject-wise statistical comparison, for complex stimuli, between DynaDiff Baseline and LFA for the AlexNet-5, CLIP and Inception metrics, from visual cortex. 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. Since the interval refers to the median while the test is rank-based, the two may disagree. Bold marks statistically significant gains ($p < 0.05$).

Metric	Subj	DynaDiff	LFA	Median Δ	95% CI	W	p
AlexNet-5	1	55.30 ± 5.77	$\mathbf{68.79 \pm 12.08}$	$+15.91$	$[+4.55, +22.73]$	52	0.004
	2	61.21 ± 11.91	$\mathbf{71.36 \pm 7.57}$	$+11.36$	$[-3.03, +22.73]$	45	0.042
	5	50.91 ± 15.73	$\mathbf{78.03 \pm 8.09}$	$+25.76$	$[+19.70, +37.12]$	55	< 0.001
	7	53.18 ± 8.70	50.61 ± 17.01	$+0.00$	$[-18.18, +12.12]$	24	0.624
CLIP	1	55.76 ± 7.17	$\mathbf{71.97 \pm 5.77}$	$+18.18$	$[+7.58, +21.21]$	55	< 0.001
	2	56.97 ± 13.56	$\mathbf{76.82 \pm 9.92}$	$+19.70$	$[-3.03, +43.94]$	49	0.014
	5	45.91 ± 4.79	$\mathbf{69.85 \pm 10.96}$	$+23.48$	$[+15.15, +28.79]$	55	< 0.001
	7	54.70 ± 8.79	51.06 ± 10.00	-2.27	$[-13.64, +6.06]$	17	0.862
Inception	1	66.97 ± 15.43	$\mathbf{83.48 \pm 8.03}$	$+15.15$	$[+7.58, +24.24]$	52	0.005
	2	73.33 ± 6.23	$\mathbf{82.27 \pm 7.00}$	$+13.64$	$[+1.52, +18.18]$	46	0.030
	5	61.67 ± 10.45	$\mathbf{85.76 \pm 8.61}$	$+23.48$	$[+16.67, +34.85]$	55	< 0.001
	7	66.06 ± 17.25	63.79 ± 10.77	-5.30	$[-18.18, +13.64]$	25	0.615

Table 7: Subject-wise statistical comparison between DynaDiff Baseline and LFA on conceptual stimuli. 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. Since the interval refers to the median while the test is rank-based, the two may disagree. Bold marks statistically significant gains ($p < 0.05$). Baseline and LFA values are mean $\pm$ SD.

Subj	Baseline	LFA	Median Δ	95% CI	W	p
1	0.109 ± 0.050	$\mathbf{0.187 \pm 0.027}$	$+0.087$	$[+0.043, +0.123]$	54	0.002
2	0.119 ± 0.048	0.156 ± 0.041	$+0.068$	$[-0.010, +0.082]$	40	0.116
5	0.082 ± 0.036	$\mathbf{0.190 \pm 0.042}$	$+0.082$	$[+0.064, +0.148]$	55	< 0.001
7	0.086 ± 0.028	$\mathbf{0.152 \pm 0.035}$	$+0.072$	$[+0.035, +0.088]$	53	0.003

Table 8: Per-ROI gains of LFA over the DynaDiff baseline for complex stimuli across subjects. For each condition we report the metric improvement Δ and the paired p-value. Bold marks statistically significant gains ($p < 0.05$). Significance was assessed with a one-sided Wilcoxon signed-rank test over the 10 partitions, with metrics averaged across subjects. Arrows indicate whether higher (↑) values correspond to better performance.

ROI	**High-Level metrics (Complex stimuli)**					
	AlexNet-5 ↑		Incep ↑		CLIP ↑	
	Δ	p	Δ	p	Δ	p
Single ROI						
Parietal	$\mathbf{+8.03}$	0.005	$\mathbf{+10.68}$	0.001	$\mathbf{+14.96}$	0.001
Frontal	$+3.64$	0.116	$\mathbf{+5.83}$	0.032	$+3.94$	0.080
Temporal	$\mathbf{+6.52}$	0.002	$\mathbf{+4.43}$	0.042	$\mathbf{+5.38}$	0.005
Prefrontal	$+0.91$	0.385	$+1.17$	0.216	$+0.27$	0.278
ROI ∪ nsdgeneral						
Parietal	$\mathbf{+13.64}$	0.001	$\mathbf{+13.26}$	0.001	$\mathbf{+12.95}$	0.001
Frontal	$\mathbf{+11.67}$	0.001	$\mathbf{+8.30}$	0.003	$\mathbf{+10.38}$	0.003
Temporal	$\mathbf{+11.48}$	0.001	$\mathbf{+9.89}$	0.001	$\mathbf{+11.29}$	0.001
Prefrontal	$\mathbf{+7.35}$	0.010	$\mathbf{+8.45}$	0.002	$\mathbf{+9.85}$	0.003

Table 9: Per-ROI similarity gains of LFA over the DynaDiff Baseline for conceptual stimuli across subjects. For each ROI we report the improvement Δ and the paired p-value, both for the ROI alone and for the ROI combined with *nsdgeneral*. Significance was assessed with a one-sided Wilcoxon signed-rank test over the 10 partitions, with metrics averaged across subjects. Bold marks statistically significant gains ($p < 0.05$). Arrows indicate whether higher (↑) values correspond to better performance.

	Single ROI		**ROI ∪ nsdgeneral**	
	Δ Similarity ↑	p	Δ Similarity ↑	p
Conceptual stimuli				
Parietal	$\mathbf{+0.055}$	0.001	$\mathbf{+0.073}$	0.001
Frontal	$\mathbf{+0.049}$	0.001	$\mathbf{+0.064}$	0.001
Temporal	$\mathbf{+0.070}$	0.001	$\mathbf{+0.073}$	0.001
Prefrontal	$\mathbf{+0.051}$	0.001	$\mathbf{+0.060}$	0.001
nsdgeneral (LFA)	$\Delta = \mathbf{+0.072}$, $p = 0.001$			

D Quantitative Results

Table 10: Reconstruction of NSD-Imagery *vision* trials, decoded by the pretrained model without any alignment. All trials of a stimulus are averaged into a single input, so one reconstruction per stimulus is produced. Values are mean ± SEM across the four subjects. Arrows indicate whether higher (↑) or lower (↓) values correspond to better performance.

Stimuli	**Low-Level**		**High-Level**				**Distance**	
	PixCorr ↑	SSIM ↑	AlexNet-2 ↑	AlexNet-5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
NSD-Imagery-Vision task – Averaged-subjects – Visual Cortex								
Simple stimuli	0.094 ± 0.021	0.548 ± 0.027	$66.67 \pm 6.28\%$	$81.82 \pm 3.76\%$	$50.00 \pm 5.64\%$	$62.50 \pm 5.91\%$	0.986 ± 0.013	0.561 ± 0.011
Complex stimuli	0.137 ± 0.011	0.284 ± 0.005	$91.67 \pm 0.76\%$	$87.12 \pm 3.06\%$	$80.68 \pm 5.51\%$	$87.88 \pm 3.86\%$	0.874 ± 0.013	0.511 ± 0.003

Table 11: Per-subject results (mean ± SD over the 10 data partitions) and average across subjects (mean ± SEM across the four subjects) for imagined simple and complex stimuli reconstruction from visual cortex. Since no alignment is applied to simple stimuli, all differences originate from complex stimuli. Arrows indicate whether higher (↑) or lower (↓) values correspond to better performance.

Method	Subj	Low-Level		High-Level				Distance	
		PixCorr ↑	SSIM ↑	AlexNet-2 ↑	AlexNet-5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Simple and Complex stimuli – Imagination-task – Visual Cortex									
DynaDiff Baseline	1	0.019 ± 0.029	0.373 ± 0.023	$64.62 \pm 8.38\%$	$62.35 \pm 5.22\%$	$57.42 \pm 7.98\%$	$59.09 \pm 4.82\%$	0.976 ± 0.017	0.602 ± 0.016
	2	-0.011 ± 0.031	0.360 ± 0.016	$58.18 \pm 7.20\%$	$62.58 \pm 8.92\%$	$62.12 \pm 4.32\%$	$57.35 \pm 5.88\%$	0.978 ± 0.022	0.600 ± 0.013
	5	-0.018 ± 0.040	0.354 ± 0.038	$54.32 \pm 7.84\%$	$53.71 \pm 12.47\%$	$52.58 \pm 5.78\%$	$54.70 \pm 4.18\%$	0.977 ± 0.013	0.608 ± 0.017
	7	0.002 ± 0.025	0.361 ± 0.026	$53.71 \pm 7.41\%$	$60.30 \pm 5.47\%$	$52.65 \pm 9.96\%$	$59.92 \pm 4.73\%$	0.988 ± 0.017	0.606 ± 0.016
	Mean	-0.002 ± 0.008	0.362 ± 0.004	$57.71 \pm 2.51\%$	$59.73 \pm 2.07\%$	$56.19 \pm 2.28\%$	$57.77 \pm 1.15\%$	0.979 ± 0.003	0.604 ± 0.002
LFA	1	0.034 ± 0.030	0.382 ± 0.028	$65.45 \pm 4.90\%$	$69.09 \pm 7.81\%$	$65.68 \pm 6.09\%$	$67.20 \pm 5.48\%$	0.949 ± 0.024	0.580 ± 0.024
	2	0.019 ± 0.038	0.367 ± 0.017	$63.03 \pm 5.10\%$	$67.65 \pm 3.44\%$	$66.59 \pm 6.84\%$	$67.27 \pm 6.56\%$	0.938 ± 0.023	0.570 ± 0.019
	5	0.004 ± 0.034	0.358 ± 0.041	$58.71 \pm 7.29\%$	$67.27 \pm 8.08\%$	$64.62 \pm 4.86\%$	$66.67 \pm 7.40\%$	0.940 ± 0.023	0.582 ± 0.016
	7	0.014 ± 0.029	0.358 ± 0.037	$51.89 \pm 4.38\%$	$59.02 \pm 9.46\%$	$51.52 \pm 8.44\%$	$58.11 \pm 5.50\%$	0.985 ± 0.015	0.607 ± 0.022
	Mean	$\mathbf{0.018 \pm 0.006}$	$\mathbf{0.366 \pm 0.006}$	$\mathbf{59.77 \pm 2.97\%}$	$\mathbf{65.76 \pm 2.28\%}$	$\mathbf{62.10 \pm 3.55\%}$	$\mathbf{64.81 \pm 2.24\%}$	$\mathbf{0.953 \pm 0.011}$	$\mathbf{0.585 \pm 0.008}$

Table 12: Per-subject results (mean ± SD over the 10 data partitions) and average across subjects (mean ± SEM across the four subjects) for imagined simple stimuli reconstructions. No alignment is applied to simple stimuli, so the DynaDiff baseline and LFA coincide by construction. Arrows indicate whether higher (↑) or lower (↓) values correspond to better performance.

Subj	Low-Level		High-Level				Distance	
	PixCorr ↑	SSIM ↑	AlexNet-2 ↑	AlexNet-5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Simple stimuli – Imagination-task – Visual Cortex								
1	0.020 ± 0.023	0.498 ± 0.053	$49.55 \pm 14.73\%$	$69.39 \pm 6.05\%$	$47.88 \pm 10.06\%$	$62.42 \pm 6.99\%$	0.988 ± 0.023	0.581 ± 0.015
2	-0.007 ± 0.036	0.476 ± 0.024	$36.97 \pm 10.69\%$	$63.94 \pm 8.26\%$	$50.91 \pm 7.77\%$	$57.73 \pm 5.07\%$	1.001 ± 0.018	0.586 ± 0.021
5	-0.012 ± 0.040	0.474 ± 0.079	$33.18 \pm 11.46\%$	$56.52 \pm 16.00\%$	$43.48 \pm 9.23\%$	$63.48 \pm 8.61\%$	0.997 ± 0.013	0.598 ± 0.026
7	-0.001 ± 0.024	0.485 ± 0.059	$34.55 \pm 8.92\%$	$67.42 \pm 9.46\%$	$39.24 \pm 10.87\%$	$65.15 \pm 4.74\%$	1.005 ± 0.016	0.582 ± 0.029
Mean	$\mathbf{0.000 \pm 0.007}$	$\mathbf{0.483 \pm 0.005}$	$\mathbf{38.56 \pm 3.74\%}$	$\mathbf{64.32 \pm 2.83\%}$	$\mathbf{45.38 \pm 2.55\%}$	$\mathbf{62.20 \pm 1.59\%}$	$\mathbf{0.998 \pm 0.004}$	$\mathbf{0.587 \pm 0.004}$

Table 13: Per-subject results (mean ± SD over the 10 data partitions) and average across subjects (mean ± SEM across the four subjects) for imagined complex stimuli reconstructions. Arrows indicate whether higher (↑) or lower (↓) values correspond to better performance.

Method	Subj	Low-Level		High-Level				Distance	
		PixCorr ↑	SSIM ↑	AlexNet-2 ↑	AlexNet-5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Complex stimuli – Imagination-task – Visual Cortex									
DynaDiff Baseline	1	0.018 ± 0.051	0.248 ± 0.017	$79.70 \pm 5.90\%$	$55.30 \pm 5.77\%$	$66.97 \pm 15.43\%$	$55.76 \pm 7.17\%$	0.963 ± 0.032	0.623 ± 0.024
	2	-0.014 ± 0.060	0.245 ± 0.026	$79.39 \pm 5.45\%$	$61.21 \pm 11.91\%$	$73.33 \pm 6.23\%$	$56.97 \pm 13.56\%$	0.954 ± 0.033	0.615 ± 0.017
	5	-0.025 ± 0.074	0.233 ± 0.027	$75.45 \pm 8.32\%$	$50.91 \pm 15.73\%$	$61.67 \pm 10.45\%$	$45.91 \pm 4.79\%$	0.956 ± 0.026	0.618 ± 0.020
	7	0.006 ± 0.033	0.237 ± 0.024	$72.88 \pm 9.83\%$	$53.18 \pm 8.70\%$	$66.06 \pm 17.25\%$	$54.70 \pm 8.79\%$	0.971 ± 0.022	0.631 ± 0.023
	Mean	-0.004 ± 0.010	0.241 ± 0.003	$76.86 \pm 1.64\%$	$55.15 \pm 2.21\%$	$67.01 \pm 2.41\%$	$53.33 \pm 2.52\%$	0.961 ± 0.004	0.622 ± 0.003
LFA	1	0.048 ± 0.059	0.265 ± 0.020	$81.36 \pm 9.20\%$	$68.79 \pm 12.08\%$	$83.48 \pm 8.03\%$	$71.97 \pm 5.77\%$	0.909 ± 0.037	0.580 ± 0.043
	2	0.045 ± 0.062	0.258 ± 0.018	$89.09 \pm 5.38\%$	$71.36 \pm 7.57\%$	$82.27 \pm 7.00\%$	$76.82 \pm 9.92\%$	0.875 ± 0.050	0.555 ± 0.038
	5	0.021 ± 0.054	0.241 ± 0.020	$84.24 \pm 8.21\%$	$78.03 \pm 8.09\%$	$85.76 \pm 8.61\%$	$69.85 \pm 10.96\%$	0.884 ± 0.046	0.566 ± 0.040
	7	0.028 ± 0.047	0.232 ± 0.030	$69.24 \pm 6.43\%$	$50.61 \pm 17.01\%$	$63.79 \pm 10.77\%$	$51.06 \pm 10.00\%$	0.964 ± 0.026	0.633 ± 0.022
	Mean	$\mathbf{0.035 \pm 0.007}$	$\mathbf{0.249 \pm 0.008}$	$\mathbf{80.98 \pm 4.23\%}$	$\mathbf{67.20 \pm 5.86\%}$	$\mathbf{78.83 \pm 5.06\%}$	$\mathbf{67.42 \pm 5.65\%}$	$\mathbf{0.908 \pm 0.020}$	$\mathbf{0.583 \pm 0.017}$

Table 14: Comparison between Brain Diffuser, MindEye1, MindEye2 performance, reported in Kneeland et al. (2025), and DynaDiff Baseline and our LFA method on simple and complex stimuli reconstructions (12 stimuli). Uncertainties were not reported for other SOTA models in (Kneeland et al., 2025) and we omitted them from our results, however they are reported in Tab. 11.

Method	Low-Level		High-Level				Distance	
	PixCorr ↑	SSIM ↑	AlexNet-2 ↑	AlexNet-5 ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Simple and Complex stimuli – Imagination-task – Visual Cortex								
MindEye1	**0.086**	0.349	59.56%	61.00%	52.03%	54.72%	0.948	**0.564**
Brain Diffuser	0.064	0.401	52.14%	58.35%	52.73%	54.07%	**0.935**	0.585
MindEye2	0.036	**0.414**	47.60%	55.38%	46.02%	50.78%	0.966	0.591
DynaDiff Baseline	−0.002	0.362	57.71%	59.73%	56.19%	57.77%	0.979	0.604
LFA	0.018	0.366	**59.77%**	**65.76%**	**62.10%**	**64.81%**	0.953	0.585

E QUALITATIVE RESULTS



Figure 10: Qualitative results of the reconstruction for all subjects of viewed simple, complex and conceptual stimuli, from left to right respectively.

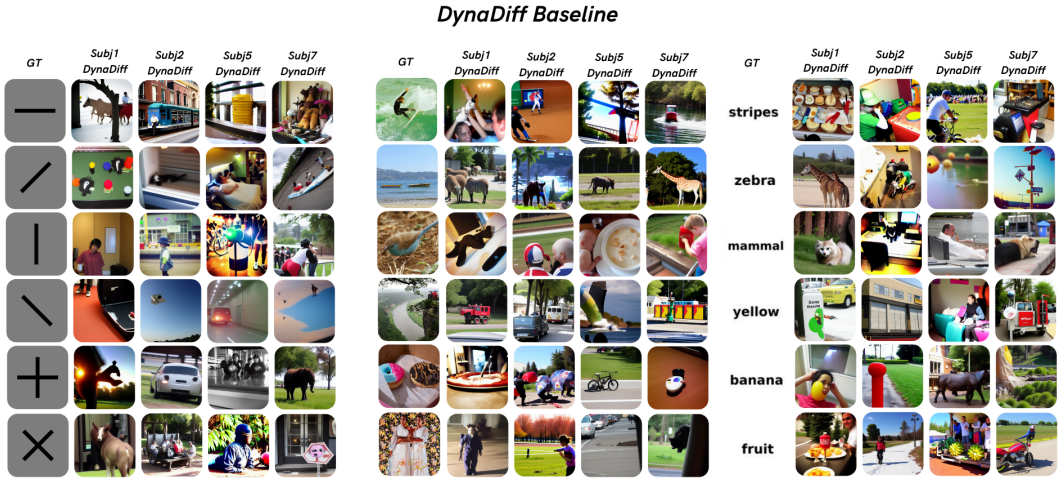


Figure 11: Qualitative results of the reconstruction for all the subject of imagined simple, complex and conceptual stimuli, from left to right respectively.

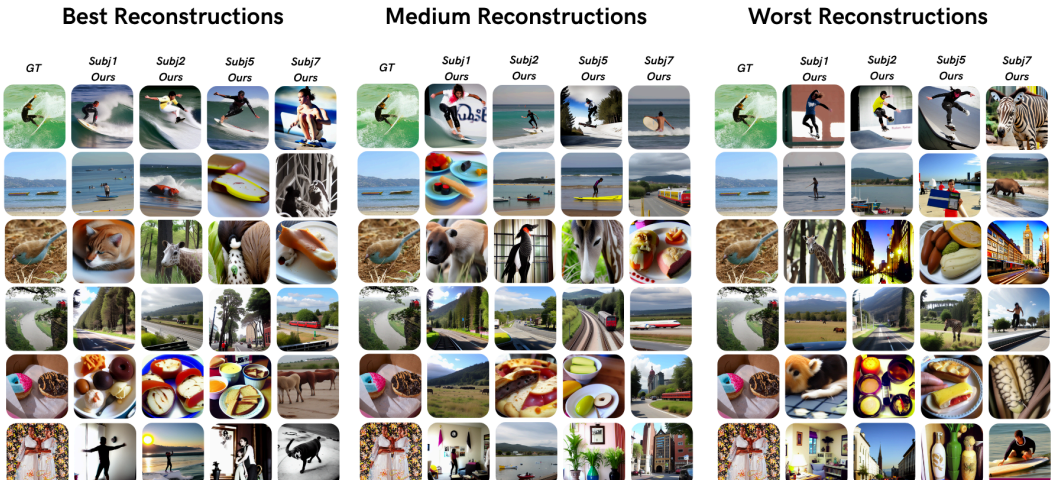


Figure 12: Qualitative results of the best, medium and worst reconstruction obtained with our LFA for all the subjects of imagined complex stimuli, from left to right respectively.

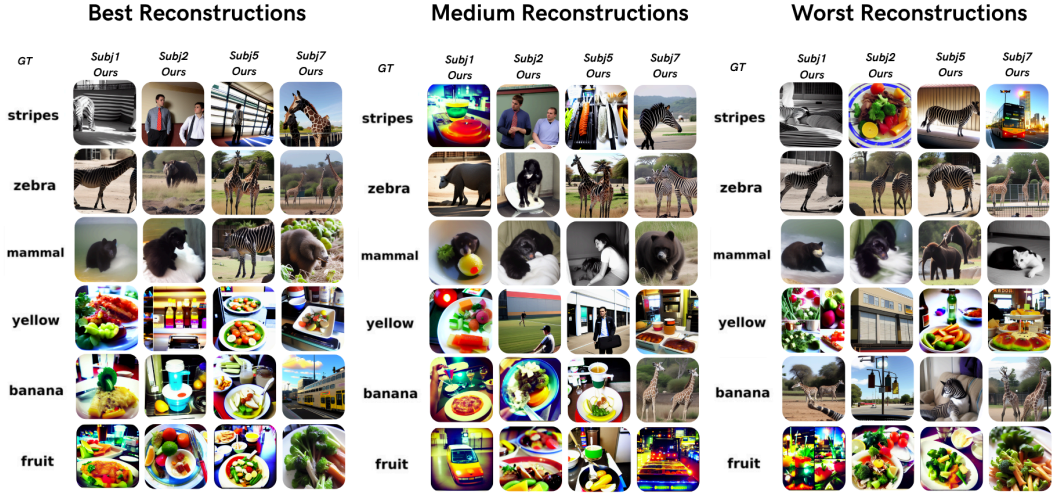


Figure 13: Qualitative results of the best, medium and worst reconstruction obtained with our LFA for all the subjects of imagined conceptual stimuli, from left to right respectively.