ALIGN, THEN ADAPT: FEW-SHOT FMRI VISUAL RECONSTRUCTION FOR NEW SUBJECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing visual images from fMRI data is a central but challenging problem in neuroscience. Despite recent progress, current methods fall short when data and computation are limited—precisely the conditions under which this task is most critical. We introduce Adapter Alignment, a staged, architecture-agnostic training procedure for adapting a pre-trained reconstruction model to a new subject with little data. Rather than letting a shared representation space emerge implicitly, lightweight subject-specific adapters are first explicitly aligned to a pre-trained reference subject on a small set of shared calibration images, and the full pipeline is then fine-tuned while preserving this alignment. The resulting representations not only place different subjects in a shared space but also align them semantically. Our results demonstrate significant gains in low-data regimes, improving reconstruction metrics by up to 8%, and pre-diffusion retrieval shows that these gains reflect additional image-specific information in the fMRI embeddings rather than the generative prior. We also propose a coverage-based algorithm to select representative calibration images. Using both techniques together, we surpass baseline models fine-tuned with one hour of data using only 24 minutes of subject-specific data. The procedure is effective across architectures, reference subjects, and datasets, including adapting 3T subjects to a 7T reference.

1 INTRODUCTION

Over the past several decades, the use of machine learning techniques has enabled the decoding and/or reconstruction of information represented in neural activation patterns. These approaches offer a key step toward characterizing and quantifying cognition (Naselaris et al., 2011). Beyond establishing the presence of information in brain activity, another important goal is to compare decoding or reconstruction performance across experimental manipulations, such as cued visual attention (Serences & Boynton, 2007; Kamitani & Tong, 2005; Scolari et al., 2012; Sprague & Serences, 2013; Itthipuripat et al., 2019; Sprague et al., 2018) and visual working memory (Serences et al., 2009; Harrison & Tong, 2009; Ester et al., 2013; Sprague et al., 2014; Christophel et al., 2012; Li et al., 2021). However, for these approaches to be practical in a cognitive neuroscience laboratory, it is typically necessary to acquire large amounts of data.

Recent efforts have begun to address the challenges associated with decoding and reconstructing brain activity. Early approaches, including self-supervised learning frameworks (Beliy et al., 2019; Gaziv et al., 2022; Mozafari et al., 2020; Shen et al., 2019; Seeliger et al., 2018; Ren et al., 2021) struggled to generate semantically accurate image reconstructions, often resulting in images with limited resemblance to the original stimuli. Significant advancements were achieved by incorporating multimodal data, such as text (Lin et al., 2022), and then by leveraging the generative power of latent diffusion models (LDMs) (Takagi & Nishimoto, 2023; Scotti et al., 2023; Lu et al., 2023; Xia et al., 2024). A more detailed literature review is presented in Appendix A. However, these state-of-the-art methods require extremely large datasets with dozens of hours of fMRI data per participant. Thus, there is a clear need to maximize the data efficiency of image reconstruction methods so that a model for a new participant can be trained using as little subject-specific data as possible.

Recent works by Scotti et al. (2024) and Liu et al. (2024) have reduced the data needed for new subjects by projecting fMRI signals from multiple subjects into a shared representation space. However, in these pipelines the shared space is treated only as a convergence point of subject-specific

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.

pathways and is not explicitly aligned. We show in our analysis that though subject representations in this space share a similar structure, manipulations such as a simple orthogonal transformation substantially increase cross-subject similarity. This motivates an explicit alignment strategy: instead of relying on the shared space to emerge implicitly, lightweight subject-specific neural network modules called adapters (Liu et al., 2024) can be trained to align each subject to a reference subject.

We propose Adapter Alignment (AA), a training procedure for low-data adaptation of new subjects in shared-space reconstruction pipelines. Its components are established (Scotti et al., 2024; Ferrante et al., 2024; Haxby et al., 2011); our contribution is their organization into a two-stage procedure, explicit adapter alignment to a pre-trained reference followed by end-to-end fine-tuning that retains the alignment loss, which outperforms alternative alignment strategies (Appendix G). AA substantially reduces data and training time while improving reconstruction quality. We also present a data-efficient image selection strategy that identifies the most representative training samples by covering key dimensions of variance in the common space. This method reduces the burden of data collection and further complements Adapter Alignment by making fine-tuning more efficient. Finally, AA is modular and composable: it can be integrated with existing reconstruction architectures (such as Scotti et al. (2023) and Scotti et al. (2024)) and complementary alignment techniques (such as Dai et al. (2025),Wang et al. (2024)), enabling practitioners to combine multiple strategies for improved performance. All the experiments were run on a single NVIDIA H200 GPU with 141GB of RAM.

Scope. AA is few-shot, not zero-shot: each new subject's adapter is fit to that subject's responses to a small calibration set shared with, or matched to, the reference. Our contributions are:

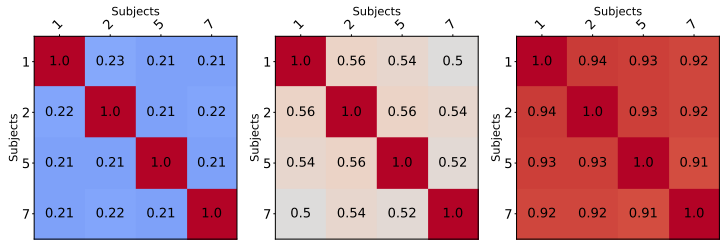
- We show that subject representations in shared fMRI decoding spaces exhibit consistent structure but remain misaligned, and that explicit alignment substantially improves cross-subject correspondence, including on held-out images (Section 2, Appendix K).
- We formally define Adapter Alignment and the image selection algorithm, and evaluate them through extensive experiments across subjects, datasets, and training regimes, showing consistently improved performance and faster convergence (up to 200x), particularly in limited-data settings (Sections 3.2, 4.1, 3.4 and 4.2). We also compare against standard selection baselines and evaluate pre-diffusion retrieval (Appendices F and L).
- We validate AA on NSD, THINGS (Hebart et al., 2023), and BOLD5000 (Chang et al., 2019) datasets, showing consistent low-data gains across architectures (Appendix W), data acquisition protocols (7T vs. 3T) (Section 5, Appendix S), training strategies (Section 4.3, Appendix G), and reference–target subject pairs (Appendix E).

2 A Common Visual Representation Space

At the core of our approach is the hypothesis that each individual's brain (or at least the tissue involved in processing static visual images) can be viewed as an instance of a shared 'common brain', with visual activity patterns that vary along similar dimensions in a shared representational space. The mapping of each individual's voxels into this space, however, must be learned from that individual's data. This idea is supported by prior neuroscience findings, including consistent category-selective tissue responses (e.g., faces, scenes) and continuous dimensions of high-level visual selectivity in ventral temporal cortex such as animate/inanimate, large/small (real-world size), and spiky/round (Huth et al., 2012; Long et al., 2018; Bao et al., 2020). Retinotopic organization provides another example, where adaptive models transform individual anatomical mappings from retinal to cortical coordinates into a shared reference space (Sprague & Serences, 2013; Wandell & Winawer, 2015). In our work, we analyze embeddings from multiple subjects and provide evidence that although fMRI signals do not naturally align in a shared space, they exhibit a similar underlying structure. We further explore structural and semantic properties of this space through eigenvector analysis, visual embedding comparisons, and alignment experiments, with additional results of semantic categories (e.g. animate / inanimate) provided in Appendix M.

Existing works, such as MindEye2 (Scotti et al., 2024), employ simple linear adapters to project subject-specific fMRI signals into a common representation space. A key indication of the existence of such a space is precisely this structural similarity across subjects. To investigate it, we use a pre-trained MindEye2 model to extract common-space embeddings for shared images across multiple subjects. Initial cross-subject cosine similarity is low, but as shown in Figure 1, applying an orthogonal

transformation with Procrustes analysis (Schönemann, 1966) reveals that the embeddings can be structurally aligned. However, this alignment is still incomplete, reflecting both the limitations of orthogonal transformations and the absence of an explicit alignment objective.



(a) Original alignment. (b) Post orthogonal transformation. (c) Post Adapter Alignment.

Figure 1: Cross-subject cosine similarity of fMRI embeddings in the common space extracted from a pre-trained MindEye2 model. (a) Cross-subject alignment is weak before any transformation. (b) Applying an orthogonal transformation substantially improves cross-subject similarity. (c) Explicit alignment through lightweight adapter modules achieves near-perfect alignment. Only Subjects 1, 2, 5, and 7 are shown for consistency throughout the manuscript.

As a preliminary experiment, we pre-trained a single-subject model and explicitly mapped subsequent subjects to this reference subject using lightweight subject-specific adapters. Even with a simple adapter design (a linear layer with a GELU), this approach yielded near-perfect alignment in the common space (Figure 1 (c)). This agreement is expected by construction on the aligned images; Appendix K shows that alignment also improves on held-out images, including within independently defined retinotopic and category-selective regions (functional, not anatomical, correspondence). The choice of adapter is not critical. Any transformation that goes beyond an orthogonal mapping can substantially improve alignment highlighting that it is the explicit alignment objective, rather than architectural complexity, that drives the improvement.

We next develop a training procedure that explicitly enforces such alignment and compare it with standard baseline fine-tuning (further analysis in Appendix K).

3 Methods

We present two methods that leverage the presence of a shared space for fMRI-to-image reconstruction. Our framework introduces Adapter Alignment (AA) as a lightweight strategy for subject-specific alignment, integrates this into a reconstruction pipeline, and augments it with a greedy image selection algorithm to further improve data efficiency. Together, these components form the basis for the experiments presented in the following section.

3.1 Dataset and Split Strategy

We primarily use the Natural Scenes Dataset (NSD) (Allen et al., 2022), which provides high-SNR 7T fMRI recordings of subjects viewing natural images. We also evaluate transfer to THINGS-fMRI (Hebart et al., 2023), which contains 3T data collected under different stimulus and task protocols, and to BOLD5K (Chang et al., 2019).

For NSD, we adopt a modified split strategy compared to the standard setup. Traditionally, images shared across subjects are reserved for testing, while training consists of subject-specific unique images. However, ideal Adapter Alignment requires a one-to-one correspondence between subjects. We therefore include the shared images in the training set and move an equal number of previously unique training images into the test set. As a result, each subject is evaluated on a distinct set of unseen images. Test splits are constructed to maintain a roughly balanced category distribution. To ensure fairness, all baselines are re-trained and evaluated on this modified split, and results are averaged across three independently generated splits.

A matched rerun on the standard split shows that this test set is harder (Appendix B); baselines use their published, untuned hyperparameters, which AAMax inherits. Similar to prior work, we focus

on Subjects 1, 2, 5, and 7, as they have complete 40-hour fMRI datasets available in NSD, enabling better reconstruction performance.

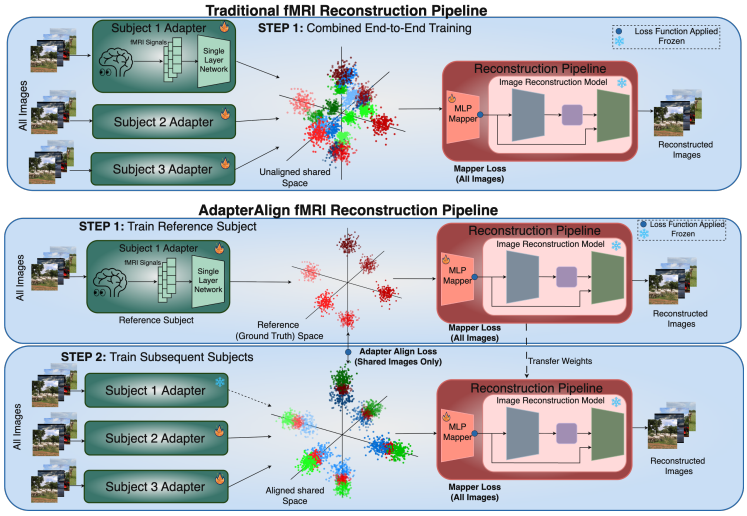


Figure 2: Training procedures for fMRI-Image Reconstruction. **Top:** Traditional Approach to reconstruction. All subjects are trained simultaneously and loss functions are applied at the MLP Mapper to align the output to pre-trained CLIP space. **Bottom:** Proposed AdapterAlign training pipeline. First, we train a single reference subject and obtain its embeddings in the shared space. Then we train subsequent subjects to minimize loss at both the shared space and the mapper. The same mapper weights are transferred over in step 2.

3.2 Adapter Alignment

Current approaches for training multi-subject pipelines with a common representation space typically follow an end-to-end strategy, where fMRI inputs are processed by the model and the loss is computed at the final layer to minimize the discrepancy between the ground-truth and predicted embeddings. Throughout the paper, we refer to this fine-tuning strategy as the baseline. In this setup, the common space emerges organically during training without explicit constraints. However, this approach has two key drawbacks: (1) it requires extended training to form a satisfactory common space and achieve convergence, and (2) the subject-specific embeddings in the common space are not strongly aligned, making it harder for the residual MLP to map multiple subject inputs to a unified output (Figure 2).

To address these challenges, we propose Adapter Alignment (AA) training. In AA training, the pipeline is first pre-trained end-to-end on a single reference subject, chosen based on reconstruction performance and data availability (Row 2 of Figure 2). Among NSD subjects, Subject 1 consistently exhibited the best reconstruction performance, making it the ideal candidate to construct the reference space. Other reference choices are evaluated in Appendix E. After pre-training, the output embeddings for the subject-specific adapter, particularly those corresponding to the shared images, are extracted.

When fine-tuning a new subject, we adopt a two-step process. The first is **Adapter Alignment**, where we train the new subject's adapter using the shared images, with the objective of aligning its output to the embeddings of Subject 1 for the same images. The adapter is allowed to overfit on these shared images until the loss is minimized. The second step is **End-to-End Fine-tuning**, where we unfreeze the entire pipeline and resume training, applying the usual losses at the final output (Row 3 of Figure 2). Additionally, we introduce an MSE loss at the adapter level specifically for the common images, which allows the adapter to restructure itself to account for subject-specific discrepancies and adapt to the unique images encountered by the new subject. We refer to this process of overfitting the adapter with AA and fine-tuning end-to-end "AAMax". We present additional experiments in Appendix K to show that this two step process improves alignment in the shared space at all stages.

Our training method is effective across both high-data and low-data regimes. With full training data, AAMax achieves performance comparable to traditional end-to-end training while converging

more quickly. In limited data settings, AAMax shows a clear advantage, producing higher-quality reconstructions and greater accuracy with substantially fewer training samples.

3.3 Reconstruction Pipeline

In general, fMRI-to-image reconstruction pipelines follow a common structure: brain signals are mapped into an intermediate embedding space (semantic or virtual), which is then decoded into an image using a pre-trained generative model such as a diffusion decoder. We discuss multiple previous approaches and their pipelines in Appendix A. Since Adapter Alignment (AA) is robust to architectural choices, we adopt the pipeline used in MindEye1 for our main experiments and refer readers to MindEye1 (Scotti et al., 2023) and MindEye2 (Scotti et al., 2024) for architectural details. We also validate our method on MindEye2's architecture (Appendix W and C) and across varying component choices and shared space dimensionalities (Appendix J).

In summary, flattened fMRI voxels from the "nsdgeneral" region are passed through a subject-specific adapter (a single linear layer with GELU non-linearity) into a shared representation space. This shared representation is then processed by a residual MLP. The resulting output is mapped into the CLIP (Radford et al., 2021) space via a pre-trained diffusion prior (Ramesh et al., 2022). This serves as a conditioning signal for a diffusion model to produce the final reconstruction.

We detail the training losses used in this pipeline for both the high-level (semantic) and low-level (visual) branches below, with full equations provided in Appendices P and Q. The overall loss function for the high-level pipeline is:

$$\mathcal{L}_{\text{new}} = \begin{cases} \lambda_1 \mathcal{L}_{\text{Prior}} + \lambda_2 \mathcal{L}_{\text{CLIP}}, & \text{unique images} \\ \lambda_1 \mathcal{L}_{\text{Prior}} + \lambda_2 \mathcal{L}_{\text{CLIP}} & \\ \quad + \lambda_3 \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{adapter,new}}, \mathbf{z}_{\text{adapter,ref}}), & \text{common images} \end{cases}$$

where $\mathbf{z}_{\text{adapter, new}}$ and $\mathbf{z}_{\text{adapter, ref}}$ are the adapter-level embeddings for the new and reference subjects, respectively, on the shared common images.

The low-level pipeline also follows a similar structure to the high-level pipeline but maps the fMRI data to latent embeddings derived from a custom autoencoder. This autoencoder is trained on downscaled images from the NSD dataset (64x64x3) and reduces them to a latent space of size 1024. The low-level pipeline learns to map to this latent space; passing these embeddings through a pre-trained decoder produces blurry reconstructions. The loss function for this stage is:

$$\mathcal{L}_{\text{new}} = \begin{cases} \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{fMRI,new}}, \mathbf{z}_{\text{AE}}), & \text{unique images} \\ \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{fMRI,new}}, \mathbf{z}_{\text{AE}}) & \\ \quad + \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{adapter,new}}, \mathbf{z}_{\text{adapter,ref}}), & \text{common images} \end{cases}$$

where $\mathbf{z}_{\text{fMRI, new}}$ is the new subject's output embedding, $\mathbf{z}_{\text{AE}}$ is the image's AE latent, and the adapter-level embeddings are defined as above. After mapping brain signals to the CLIP space, the mappings are fed as conditioning to a frozen Diffusion Model (Xu et al., 2023) to generate final reconstructions.

3.4 Greedy Algorithm for Best Candidate Image Selection

In the context of fMRI-based visual reconstruction, we introduce a greedy algorithm that selects the most representative subset of images from the common space. We first decompose (SVD) the adapter's weight matrix of the reference subject. Then we project the embeddings of the shared set of 1000 images onto the d principal dimensions corresponding to the principal singular vectors. For each dimension j, we partition the range into B_j bins that is determined by the ratio of singular value λ_j to the largest singular value λ_1:

$$B_j = \left\lfloor w \cdot \frac{\lambda_j}{\lambda_1} \right\rfloor, \quad j = 1, 2, \ldots, d$$

where w is a predefined scaling parameter. Dimensions with larger singular values are partitioned into more bins, reflecting their greater variance and importance in capturing meaningful data features.

Each image (and its fMRI) maps to exactly one bin in each dimension and our goal is to find the smallest subset of images that covers all bins in each dimension. This subset is critical in selecting the most representative images for a new subject, ensuring that the selected images capture the key activity patterns across subjects. The problem as stated is an NP-hard set-cover instance. The maximum "set" size is d, the number of principal eigenvector dimensions. The greedy heuristic below therefore achieves an $H(d)$-approximation (Johnson, 1974; Lovász, 1975), where $H(d)$ is the harmonic series sum up to d. Stopping at a fixed budget, as in our experiments, gives a $(1 - 1/e)$-approximation for maximum coverage (Nemhauser et al., 1978).

Given a subset S of images, let $Gap(S, j)$ be the number of empty bins in dimension j, and let $Gap(S) = \Sigma_j Gap(S, j)$ be the total number of empty bins across all dimensions. At every step, the greedy algorithm chooses image $i \notin S$ that minimizes Gap when added to S.

After applying the image selection algorithm, the chosen subset simply replaces random sampling for fine-tuning. The training procedure itself is unchanged: the selected images can be used with either baseline fine-tuning or AAMax. The algorithm serves purely as a strategy to identify the most representative subset of images, reducing the data requirement without altering the reconstruction pipeline. Formal proofs for the algorithm are presented in Appendix N.

4 RESULTS

We now evaluate the performance of our proposed framework across a range of experimental conditions. We begin with complete data experiments, where AAMax is compared against baseline training. We then turn to limited data settings, where we report results from both AAMax and AAMax combined with image selection in a unified comparison. We also analyze how AAMax scales with increasing amounts of training data, highlighting that its advantage diminishes as larger datasets allow even unaligned models to generalize. Finally, we present experiments showing that AAMax is effective across architectures using MindEye1 and MindEye2 (Scotti et al., 2024; 2023) architectures.

4.1 TESTS ON COMPLETE DATA

We first validate our training method using the full dataset. Subject 1 is used as the reference subject, while fine-tuning is performed on Subjects 2, 5, and 7. The results averaged over all 3 subjects are presented in Tables 1 and 27. AA can be used with any reference subject, but the quality of the reference affects fine-tuning performance (Appendix E). We use the same set of metrics as Scotti et al. (2024) to evaluate our reconstructions. Pixel Correlation, SSIM and 2-way percent correct for the 2nd and 5th layer of AlexNet are considered low level metrics. 2-way percent correct for Inception(Szegedy et al., 2016), CLIP, EfficientNet(Tan, 2019) and SwAV(Caron et al., 2020) score are considered high level metrics. When trained to completion, both baseline fine-tuning and AAMax show similar overall performance. However, the MSE loss at the adapter level is significantly lower (by an order of magnitude as shown in Appendix K.2) when using adapter alignment. This allows for more precise post-hoc analysis after extracting embeddings from the common space.

Table 1: Quantitative results on training the reconstruction models on 40 hours of data. The fine-tuning results are averaged over subjects 2,5 and 7. AAMax and baseline fine-tuning perform similarly across all metrics. Std. deviations are provided in Appendix U.

	Low-Level				**High-Level**			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Subject 1 Pre-train	0.345	0.346	92.80%	96.88%	94.40%	90.02%	0.692	0.399
Baseline FT AVG	0.258	**0.339**	**89.64%**	95.05%	92.63%	**89.64%**	0.717	0.427
AAMax FT AVG	**0.259**	0.337	89.20%	**95.16%**	**92.68%**	89.49%	**0.713**	**0.423**

4.2 TESTS ON LIMITED DATA

A key objective of multi-subject pipelines is to achieve accurate models of new subjects with limited data. Because acquiring fMRI data is costly, reducing the data requirements makes reconstruction pipelines more practical and broadly applicable.

Table 2: Limited data fine-tuning results (1 hour, 250 images) using Subject 1 as the reference and Subjects 2, 5, and 7 as fine-tuning subjects. Each subject is evaluated under four settings: Baseline with random sampling, AAMax with random sampling, Baseline with image selection (IS), and AAMax with IS. Results are averaged over three runs with fixed random splits. AAMax consistently outperforms the baseline, IS provides additional gains in both training regimes, and the best performance is achieved by combining AAMax with IS. Pixel-level metrics are shown in gray since they are not optimized in these runs but are still valid.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Subject 2 FT 1HR Baseline	0.078±0.010	0.264±0.007	75.33%±1.62%	84.18%±1.40%	74.15%±0.99%	73.66%±1.3%	0.862±0.011	0.524±0.009
Subject 2 FT Selected 1HR Baseline	0.097±0.003	0.288±0.004	78.78%±0.41%	85.78%±0.69%	77.13%±0.76%	76.17%±0.03%	0.842±0.003	0.513±0.005
Subject 2 FT 1HR AAMax	0.112±0.005	0.259±0.025	78.75%±1.29%	88.04%±1.19%	78.99%±0.97%	78.28%±0.39%	0.824±0.003	0.513±0.014
Subject 2 FT Selected 1HR AAMax	0.102±0.003	0.267±0.11	**81.41%±0.27%**	**89.55%±0.26%**	**80.43%±0.13%**	**80.33%±0.01%**	**0.823±0.002**	**0.496±0.002**
Subject 5 FT 1HR Baseline	0.082±0.002	0.267±0.002	75.05%±0.08%	83.28%±0.27%	76.04%±1.12%	76.40%±0.64%	0.855±0.002	0.526±0.006
Subject 5 FT Selected 1HR Baseline	0.098±0.001	0.286±0.003	77.07%±1.0%	86.31%±0.1%	78.09%±0.30%	79.55%±0.06%	0.827±0.004	0.499±0.003
Subject 5 FT 1HR AAMax	0.105±0.001	0.278±0.005	79.90%±0.34%	88.68%±0.07%	80.63%±0.31%	82.20%±0.38%	0.808±0.004	0.493±0.02
Subject 5 FT Selected 1HR AAMax	0.117±0.004	0.285±0.009	**81.05%±0.24%**	**89.88%±0.78%**	**81.92%±0.41%**	**82.47%±0.43%**	**0.799±0.001**	**0.480±0.004**
Subject 7 FT 1HR Baseline	0.079±0.006	0.252±0.15	72.98%±3.03%	79.98%±1.39%	71.57%±1.04%	72.07%±0.86%	0.871±0.009	0.539±0.014
Subject 7 FT Selected 1HR Baseline	0.099±0.004	0.286±0.005	78.41%±0.08%	84.77%±1.13%	74.91%±1.52%	75.79%±0.21%	0.844±0.006	0.515±0.001
Subject 7 FT 1HR AAMax	0.104±0.002	0.285±0.001	80.04%±0.22%	87.47%±0.78%	79.17%±0.38%	78.93%±0.53%	0.826±0.001	0.509±0.001
Subject 7 FT Selected 1HR AAMax	0.123±0.003	0.298±0.019	**81.24%±1.36%**	**88.34%±0.30%**	**79.94%±1.00%**	**79.81%±0.91%**	**0.820±0.001**	**0.495±0.007**

In our experiments, we designate Subject 1 as the reference and fine-tune Subjects 2, 5, and 7 using the shared images. For consistency, we use the same reference subject checkpoint across all experiments and report averaged results over three runs with fixed random splits. Table 2 presents the unified results for fine-tuning with 1 hour of data (250 images), comparing baseline training and AAMax with both random sampling and image selection (IS). Across all subjects, AAMax significantly outperforms the baseline, while IS consistently improves performance regardless of the fine-tuning strategy. As expected, the strongest results are obtained when combining AAMax with IS, showing that the two approaches are complementary. With all other settings identical, this isolates the contributions of alignment and IS (further ablations in Appendices G and I). For IS, we select images based on the top 40 eigenvector dimensions, with ablations on this choice provided in Appendix F. We run experiments with scaling the amount of training data in the following section. Ablations with 2 and 4 hours of data are included in Appendix H, where the same trend holds. We also run experiments on convergence in Appendix H and find that AA followed by just 1 epoch of End-to-End fine-tuning is good enough to outperform the baseline. Appendix F compares IS with standard selection baselines, and Appendix L shows that AAMax also improves pre-diffusion retrieval, so its gains are not due to the generative prior.

We further investigate the efficiency of our approach by testing how far data requirements can be reduced while maintaining reconstruction performance. In this setup, Subjects 2 and 5 were fine-tuned from Subject 1 using 100, 150, and 200 training images. At 100 images, both AAMax and AAMax+IS already match the baseline trained with 1 hour of data, and at 150 images they clearly surpass it. As shown in Figure 3, AAMax+IS consistently achieves the strongest results, though even AAMax alone outperforms the baseline in these low-data regimes. These findings highlight the robustness of adapter alignment and demonstrate that effective reconstructions can be achieved with as little as 40% of the data previously considered necessary.

4.3 Integrating with Existing Alignment Methods

AA combined with IS can also be integrated with existing alignment methods, enabling a modular training strategy. This is possible when the methods operate at different stages of the decoding pipeline and thus address complementary sources of cross-subject variability.

For alignment methods that operate at earlier stages of the pipeline, such as voxel-space or categorical alignment, AAMax+IS can be applied downstream without interference. MindAligner (Dai et al., 2025) performs coarse alignment directly in fMRI voxel space using category-level supervision. In contrast, AAMax operates in the learned common representation space after the subject-specific adapter, enforcing fine-grained geometric alignment. Because these methods act on distinct representations, they can be combined sequentially: coarse alignment reduces subject variability at the signal level, while AAMax+IS refines alignment in the shared space. As shown in Table 3,

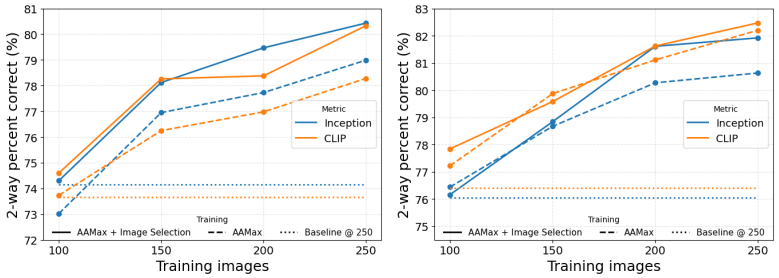


Figure 3: Pushing the limits of data efficiency with AAMax+Image Selection. Left: Subject 2 fine-tuned from Subject 1. Right: Subject 5 fine-tuned from Subject 1. In both cases, AAMax and AAMax+Image Selection outperform the baseline trained with 250 images, even when using only 150 images (a 40% reduction) and match performance with 100 images (a 60% reduction). Image selection in these experiments uses the top 20 eigenvector dimensions (Appendix F).

Table 3: **Comparison and combination of MindAligner and Adapter Alignment.** All methods use MindEye1 with pretraining on Subject 1 and fine-tuning on Subject 2 for 1 hour of data. MindAligner follows the original paper, using categorical stimulus pairs. AA uses our 250-image Image Selection subset that spans the breadth of the principal eigenspace. The combined method applies MindAligner for coarse voxel-space alignment followed by AAMax + Image Selection for fine-grained common-space alignment. The combined approach achieves the strongest overall performance.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
MindAligner (MA)	0.116	0.280	82.11%	88.68%	81.36%	79.69%	0.809	0.483
AAMax + IS	0.118	0.285	81.78%	89.61%	**81.49%**	79.55%	0.810	0.491
MA + AAMax + IS	0.124	0.284	**82.24%**	**89.95%**	**81.49%**	**80.99%**	**0.795**	**0.480**

combining MindAligner with AAMax+IS yields the strongest overall performance, demonstrating that Adapter Alignment is not a competing alternative but a complementary refinement stage. A detailed breakdown of how we integrate MindAligner and AAMax+IS is provided in Appendix T.

In contrast, methods such as MindBridge (Wang et al., 2024) and Ferrante et al. (2024) perform alignment at the same level of the pipeline as AA, namely through subject-specific adapters mapping into a shared space. For these methods, direct integration of the alignment method is not possible; instead, the relevant comparison is between alignment training strategies. To isolate this effect, we compare three regimes under identical architectural and data conditions in Appendix G. We find that AAMax consistently outperforms both alternatives across low-level and high-level metrics. This demonstrates that, when alignment is performed at the same stage of the pipeline, the training paradigm is the key differentiator, and AAMax is the most effective of the strategies compared.

Finally, we emphasize that AAMax+IS is compatible with many additional components proposed in prior work. Techniques such as synthetic fMRI generation, auxiliary regularization, or improved diffusion backbones are orthogonal to our contribution and can be incorporated alongside AAMax for further gains. Taken together, these results position Adapter Alignment not as a replacement for existing alignment methods, but as a general-purpose alignment module that can either (i) be combined with upstream alignment approaches or (ii) outperform alternative strategies when operating at the same representational level.

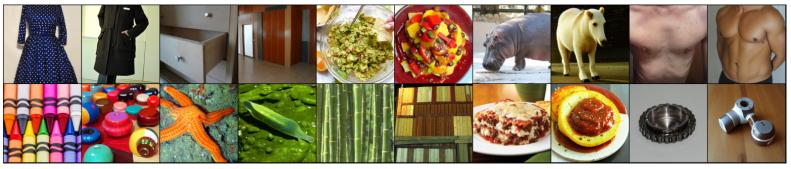


Figure 4: Examples of reconstructed images from Subject 1 of the THINGS dataset after fine-tuning. Each pair depicts the ground truth (left) and the reconstructed image (right).

5 TRANSFER TO OTHER DATASETS

To test whether a reference model trained on one dataset can be used to adapt subjects from another, we conducted experiments on the THINGS dataset. We used a model pre-trained on Subject 1 from the NSD dataset as the reference and fine-tuned it on data from Subject 1 of the THINGS dataset. Since no images are shared between THINGS and NSD, we construct an approximate common set: we compute the CLIP similarity between each THINGS training image and the 1,000 NSD shared images and retain pairs with cosine similarity above 0.90, yielding approximately 150 semantically matched pairs (Appendix S). As shown in Table 4, AAMax again outperforms conventional fine-tuning, although correspondence quality matters (Appendix S).

Several factors may account for the performance differences between the two datasets. Recent fMRI reconstruction reports have largely relied on the NSD, which provides high-SNR 7T fMRI acquired at 1.8 mm resolution, but this level of data quality is often out of reach for many neuroscience labs. THINGS uses a more widely available 3T acquisition at 2 mm resolution. The datasets also differ in SNR (7T vs 3T), stimulus presentation protocol, stimulus set (NSD: MS-COCO, THINGS: curated object images), and task performed by participants (NSD: recognition memory on each image; THINGS: object or not?). These differences make variation in decoding and reconstruction performance expected, rather than evidence that models fail to transfer across datasets.

We further evaluated cross-dataset transfer using BOLD5K (Chang et al., 2019), with results reported in Appendix S. Because the COCO images in BOLD5K overlap with the NSD shared images, these experiments use exact image correspondences, and they compare reference datasets under an identical fine-tuning protocol. Consistent with our design, NSD-trained models provide substantially stronger references, improving fine-tuning on both NSD and BOLD5K target subjects. By contrast, BOLD5K-trained references yield weaker transfer, even for BOLD5K targets. These findings support our central premise that high-SNR reference datasets induce more stable and informative shared representations, making them better suited for efficient cross-subject and cross-dataset alignment.

Table 4: Quantitative results on fine-tuning THINGS-fMRI Subject 1 using NSD Subject 1 as the reference subject. Only the high level pipeline is trained and therefore, pixelwise metrics are excluded. AAMax outperforms the baseline even when an explicit data mapping is not available.

	Low-Level		High-Level			
Method	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
NSD Subj1 Pre-train	92.80%	96.88%	94.40%	90.02%	0.692	0.399
THINGS Subj1 Individual	68.02%	72.6%	58.8%	64.29%	0.947	0.649
THINGS Subj1 Baseline Fine-Tune	77.46%	81.07%	71.06%	73.16%	0.911	0.619
THINGS Subj1 AAMax Fine-Tune	**78.27%**	**85.55%**	**74.10%**	**81.37%**	**0.902**	**0.595**

6 CONCLUSIONS & DISCUSSION

In this work, we provide evidence that fMRI signals from different subjects can be structurally aligned within a shared representation space. Building on this, we introduce Adapter Alignment (AA), a staged training procedure for low-data adaptation of new subjects that consistently outperforms traditional end-to-end pipelines in low-data regimes while achieving comparable performance in high-data settings. We further propose a greedy image selection algorithm that identifies representative training images, together reducing the data required for fine-tuning by up to 60% while matching baseline performance, and by 40% while significantly outperforming it. These contributions demonstrate that accurate visual reconstructions can be achieved with substantially less data than previously thought feasible, greatly improving the efficiency of fMRI-based reconstruction.

Our experiments show that AA is robust across reference–target subject pairs, architectures, and datasets. As a few-shot method, it requires a small calibration set shared with or matched to a reference subject, its performance depends on the quality of that reference and correspondence, and our main evaluation uses a modified NSD split (Appendix X).

By leveraging the shared representation space, AA makes fMRI-based visual reconstruction more efficient and accessible, and it can be combined with other alignment strategies for further gains. Future work could investigate richer adapter designs and extend these strategies to other brain signal modalities, advancing data-efficient brain-to-image reconstruction.

AI USE STATEMENT

In this work, we used generative AI tools (ChatGPT) to edit the paper to improve readability (grammar, spelling, and word choice) and to assist with formatting the manuscript and drafting parts of the appendix. Generative AI tools were also used to generate and run a small portion of the experiments. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available fMRI datasets (NSD, THINGS-fMRI, and BOLD5000), which were collected by their original authors with informed consent under institutional ethics approval; we collected no new human-subject data and use no personally identifying information. Because fMRI-to-image reconstruction processes sensitive brain data, it raises privacy, consent, and misuse concerns. These methods should not be interpreted as general-purpose "mind reading": reconstructions are strongly shaped by the pre-trained generative model, and Adapter Alignment requires each new participant to cooperatively provide fMRI data in the scanner, so it cannot be applied to an individual without their participation. By reducing the scan time required per participant, our approach lowers the burden on research participants and makes this research more accessible to laboratories with standard 3T scanners. Any future deployment, especially in clinical populations, must include safeguards for informed consent, data security, transparency about reconstruction uncertainty, and protection against coercive or unauthorized use.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

Thomas Bazeille, Elizabeth DuPre, Hugo Richard, Jean-Baptiste Poline, and Bertrand Thirion. An empirical evaluation of functional alignment using inter-subject decoding. *NeuroImage*, 245: 118683, 2021.

Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

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, 2019.

Po-Hsuan Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James V Haxby, and Peter J Ramadge. A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Thomas B Christophel, Martin N Hebart, and John-Dylan Haynes. Decoding the contents of visual short-term memory from human visual and parietal cortex. *Journal of Neuroscience*, 32(38): 12983–12989, 2012.

Yuqin Dai, Zhouheng Yao, Chunfeng Song, Qihao Zheng, Weijian Mai, Kunyu Peng, Shuai Lu, Wanli Ouyang, Jian Yang, and Jiamin Wu. Mindaligner: Explicit brain functional alignment for cross-subject visual decoding from limited fmri data. *arXiv preprint arXiv:2502.05034*, 2025.

Edward F Ester, David E Anderson, John T Serences, and Edward Awh. A neural measure of precision in visual working memory. *Journal of cognitive neuroscience*, 25(5):754–761, 2013.

Matteo Ferrante, Tommaso Boccato, Furkan Ozcelik, Rufin VanRullen, and Nicola Toschi. Through their eyes: multi-subject brain decoding with simple alignment techniques. *Imaging Neuroscience*, 2:1–21, 2024.

Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022.

Stephenie A Harrison and Frank Tong. Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238):632–635, 2009.

James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

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

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

Sirawaj Itthipuripat, Thomas C Sprague, and John T Serences. Functional mri and eeg index complementary attentional modulations. *Journal of Neuroscience*, 39(31):6162–6179, 2019.

David S. Johnson. Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9(3):256–278, 1974.

Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024.

Hsin-Hung Li, Thomas C Sprague, Aspen H Yoo, Wei Ji Ma, and Clayton E Curtis. Joint representation of working memory and uncertainty in human cortex. *Neuron*, 109(22):3699–3712, 2021.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636, 2022.

Yulong Liu, Yongqiang Ma, Guibo Zhu, Haodong Jing, and Nanning Zheng. See through their minds: Learning transferable neural representation from cross-subject fmri, 2024. URL `https://arxiv.org/abs/2403.06361`.

Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38):E9015–E9024, 2018.

László Lovász. On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, 13(4): 383–390, 1975.

Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5899–5908, 2023.

Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

Xuelin Qian, Yun Wang, Jingyang Huo, Jianfeng Feng, and Yanwei Fu. fmri-pte: A large-scale fmri pretrained transformer encoder for multi-subject brain activity decoding. *arXiv preprint arXiv:2311.00342*, 2023.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Ziqi Ren, Jie Li, Xuetong Xue, Xin Li, Fan Yang, Zhicheng Jiao, and Xinbo Gao. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228:117602, 2021.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Alexander Schrijver et al. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer, 2003.

Miranda Scolari, Anna Byers, and John T Serences. Optimal deployment of attentional gain during fine discriminations. *Journal of Neuroscience*, 32(22):7723–7733, 2012.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 37, 2023.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ Van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018.

John T Serences and Geoffrey M Boynton. Feature-based attentional modulations in the absence of direct visual stimulation. *Neuron*, 55(2):301–312, 2007.

John T Serences, Edward F Ester, Edward K Vogel, and Edward Awh. Stimulus-specific delay activity in human primary visual cortex. *Psychological science*, 20(2):207–214, 2009.

Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in computational neuroscience*, 13:432276, 2019.

Thomas C Sprague and John T Serences. Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nature neuroscience*, 16(12):1879–1887, 2013.

Thomas C Sprague, Edward F Ester, and John T Serences. Reconstructions of information in visual spatial working memory degrade with memory load. *Current Biology*, 24(18):2174–2180, 2014.

Thomas C Sprague, Sirawaj Itthipuripat, Vy A Vo, and John T Serences. Dissociable signatures of visual salience and behavioral relevance across attentional priority maps in human cortex. *Journal of neurophysiology*, 119(6):2153–2165, 2018.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, 2023.

Mingxing Tan. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced gromov wasserstein. *Advances in neural information processing systems*, 35:21792–21804, 2022.

Alexis Thual, Yohann Benchetrit, Felix Geilert, Jérémy Rapin, Iurii Makarov, Hubert Banville, and Jean-Rémi King. Aligning brain functions boosts the decoding of visual semantics in novel subjects. *arXiv preprint arXiv:2312.06467*, 2023.

Brian A Wandell and Jonathan Winawer. Computational neuroimaging and population receptive fields. *Trends in cognitive sciences*, 19(6):349–357, 2015.

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

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8226–8235, 2024.

Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7754–7765, 2023.

APPENDIX CONTENTS

- **A** Related Work
- **B** Evaluation Protocol and Comparison with the Standard NSD Split
- **C** Fine-tuning on Different Subjects with MindEye2
- **D** Generalizing to Different Adapter Architectures
- **E** Generalizing to Different Reference Subjects
- **F** Image Selection Algorithm Ablations and Baselines
- **G** Comparing Different Alignment Training Strategies
- **H** Extended Limited Data Experiments
- **I** Ablation Study on the two phases of Adapter Alignment Fine-Tuning
- **J** Additional Scaling Analysis
- **K** Common Space and Neural Correspondence Analysis
- **L** Pre-Diffusion Retrieval
- **M** Emergence of Common Concepts: Additional Details
- **N** Image Selection Algorithm: Formulations and Proofs
- **O** Training Time and Computational Cost
- **P** High-Level Pipeline Training Process
- **Q** Low-Level Pipeline Training Process
- **R** Visual Comparison Results among Different Methods
- **S** Cross-Dataset Transfer and Correspondence Quality
- **T** Combining Adapter Alignment and Image Selection with MindAligner
- **U** Summary Statistics
- **V** Scaling With Training Data
- **W** Generalizing To Different Architectures
- **X** Limitations
- **Y** Sample Reconstructions

A RELATED WORK

Recent works have achieved impressive results by mapping fMRI data to latent diffusion model (LDM) spaces (Takagi & Nishimoto, 2023; Scotti et al., 2023; Lu et al., 2023; Xia et al., 2024), while simultaneously integrating multiple modalities. Despite this progress, these methods have not been thoroughly tested for their generalization performance across a larger population of subjects. In other words, constructing a model for a new subject still requires retraining the model from scratch to ensure optimal performance.

Approaches like Ferrante et al. (2024) have demonstrated cross-subject brain decoding using limited data samples by aligning multiple subjects. Although the authors, like us, leveraged visual stimuli commonly used for multiple subjects, the fidelity of their reconstructed images lags behind current advancements. Scotti et al. (2024), produced significant results in image reconstruction under limited data conditions but did not fully exploit the properties of the common representation space. Adapter Alignment builds on this shared-space design and explicitly aligns each new subject to it, which enables faster convergence and improved performance in the limited-data setting. Wang et al. (2024) introduced a cross-subject brain-decoding framework that unifies fMRI dimensions through adaptive aggregation and learns subject-invariant representations via cyclic reconstruction. Its joint optimization of alignment and reconstruction resembles the second stage of our procedure; under matched conditions, a MindBridge-style joint training variant is outperformed by explicit alignment followed by end-to-end fine-tuning (Appendix G). Finally, Dai et al. (2025) manages to align subjects directly in voxel space using categorical priors and many-to-many mappings, motivated by the

absence of shared stimuli, whereas our strategy operates within the learned common representation space after the subject-specific adapter. Because the two methods act on different representations, they can be combined (Section 4.3, Appendix T).

Qian et al. (2023) refers to a 'common space'; however, their representation is primarily anatomical and only implicitly semantic. Thual et al. (2022) focuses on aligning the entire cortical surfaces of different individuals in a way that preserves both anatomical structure and functional activation patterns. However many downstream tasks like mapping fMRI signals to semantic embeddings for image reconstruction benefit from representations that capture high-level semantic content. Therefore the suggested approach doesn't necessarily capture abstract, semantic features (such as object categories, scenes, or other high-level concepts) that are crucial for decoding complex stimuli. Thual et al. (2023) is an extension of Thual et al. (2022). They apply functional alignment on top of the anatomical alignment. The method is very efficient even in limited data settings but it's less integrated, meaning that it appears more as a preprocessing step. Piecewise shared response modeling Bazeille et al. (2021) provides significant performance in terms of aligning neural data from different subjects. However, the independent alignment of segments might not integrate seamlessly into a coherent global representation. Similarly to MindEye2, but in the context of M/EEG data, Li et al. (2024) manage to reconstruct visual stimuli using cross subject decoding. They tackle inter-subject variability primarily by incorporating subject-specific tokens into their M/EEG encoder.

Functional alignment in neuroscience. Aligning functional responses across individuals using shared stimuli has a long history in cognitive neuroscience. Hyperalignment (Haxby et al., 2011) learns subject-specific transformations into a common high-dimensional model of representational space, and the shared response model (Chen et al., 2015) learns a reduced-dimensional shared response with subject-specific bases. Adapter Alignment operates within this tradition, but in the learned embedding space of a reconstruction pipeline: the subject-specific adapter absorbs differences in voxel count, anatomy, and cortical organization, and the alignment objective is applied to its output rather than to voxel responses. Treating each subject's fMRI pattern as a vector is therefore an architectural abstraction and not a claim that cortical geometry is irrelevant; anatomically grounded alignment methods (Thual et al., 2022; 2023) operate upstream of this stage and are complementary to it.

Positioning. Table 5 summarizes where and how related methods enforce cross-subject alignment. The individual components of Adapter Alignment, namely subject-specific adapters, a reference or shared space, and alignment on shared stimuli, are established. Our contribution is their organization into a staged, architecture-agnostic procedure: an explicit alignment stage followed by end-to-end fine-tuning that retains the alignment loss, together with a selection algorithm that exploits the aligned space to choose calibration images. Appendix G shows that this organization outperforms adapter-only alignment and joint end-to-end alignment under matched architecture, data, and hyperparameters.

Table 5: Where and how related methods enforce cross-subject alignment.

Method	Where alignment acts	Alignment signal	How alignment is trained
Hyperalignment / SRM (Haxby et al., 2011; Chen et al., 2015)	Voxel / response space	Shared stimuli	Separate alignment model (no reconstruction)
MindEye2 (Scotti et al., 2024)	Shared space after linear adapters	Implicit, via reconstruction losses	Jointly end-to-end; no explicit alignment
Ferrante et al. (2024)	Adapter output (reference space)	Shared stimuli	Explicit alignment only; no subsequent end-to-end training
MindBridge (Wang et al., 2024)	Subject-invariant shared space	Cyclic reconstruction	Jointly with the reconstruction objective
MindAligner (Dai et al., 2025)	Voxel space	Category-level pairs	Explicit voxel-space alignment module
Adapter Alignment (ours)	Adapter output (reference space)	Shared or semantically matched calibration images	Explicit alignment, then end-to-end fine-tuning retaining the alignment loss

B Evaluation Protocol and Comparison with the Standard NSD Split

Modified split. In the standard NSD protocol, the 1,000 images viewed by all subjects are reserved for testing, and each subject is trained on images unique to that subject. Adapter Alignment requires a one-to-one correspondence between the reference and the new subject during alignment, and it targets a practical setting in which a small calibration set, shared with a pre-trained reference subject, is collected for every new participant. We therefore move the shared images into the training set and replace them in the test set with an equal number of previously unique training images (Section 3.1). Each subject is thus evaluated on its own set of unseen, subject-specific images, and the test split is

constructed to maintain a roughly balanced category distribution. Calibration images, reconstruction-training images, and evaluation images remain disjoint. All baselines and AAMax are retrained on exactly the same modified split, and results are averaged over three independently generated splits.

Calibration against the standard split. To quantify how the modified split affects evaluation difficulty, we reran the MindEye2 baseline ourselves on the standard split under a matched architectural setup. Across the standard-split and modified-split runs, everything was kept the same, including the architecture, the training and fine-tuning procedure, and the hyperparameters, for which we used the setup recommended in the MindEye2 manuscript (Scotti et al., 2024); only the dataset split was changed. Table 6 reports the results, averaged over Subjects 2, 5, and 7 fine-tuned from Subject 1 with 1 hour of data.

Table 6: Effect of the dataset split on MindEye2 fine-tuning (1 hour of data, Subject 1 as the reference). The standard-split baseline was rerun with the hyperparameters recommended by MindEye2; only the split differs between the standard and modified rows. All rows are averaged over Subjects 2, 5, and 7; per-subject modified-split results are given in Appendices W and C.

Method	**Split**	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓
MindEye2 Baseline FT	Standard	**82.94%**	**89.68%**	80.36%	78.74%	**0.814**
MindEye2 Baseline FT	Modified	79.54%	86.46%	78.50%	75.13%	0.847
MindEye2 AAMax FT	Modified	81.23%	88.80%	**82.79%**	**79.08%**	0.819

Discussion. Under identical training conditions, the baseline performs worse on every metric under the modified split than under the standard split, indicating that the subject-specific unique-image test set is a more difficult benchmark than the shared-image test set. Despite this harder protocol, AAMax recovers most of the low-level gap and exceeds the standard-split baseline on Inception and CLIP. Because all within-paper comparisons use identical modified splits, the gains of AAMax over the baseline are attributable to the training procedure rather than to the split redesign.

We note two limitations of this comparison. First, the baselines use the hyperparameter configuration recommended in the original MindEye1 and MindEye2 papers and were not independently retuned for the modified split; AAMax inherits exactly the same backbone hyperparameters. This controlled practice isolates the effect of the training procedure but does not establish that each baseline is optimally tuned for the modified protocol. Second, our goal is not a leaderboard comparison under the standard NSD benchmark, but a practical protocol for future data collection and low-data adaptation to new participants, in which a small shared calibration set can be collected by design.

C Fine-tuning on Different Subjects with MindEye2

The main text presents results on fine-tuning Subject 2 using Subject 1 as reference. Here we present the results for Subject 5 and Subject 7 too with 1 hour of training data. In both cases, AAMax outperforms the baseline on most metrics.

Table 7: Fine-tuning on Subjects 5 and 7 using MindEye2's architecture. AAMax outperforms the baseline on most metrics. The same reference subject checkpoint was used in all the runs and all models were trained with the exact same hyperparameter setup.

	Low-Level				**High-Level**			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Subject 5 FT 1HR Baseline	0.102	**0.358**	78.40%	86.61%	79.85%	77.23%	0.839	0.494
Subject 5 FT 1HR AAMax	**0.121**	0.354	**80.68%**	**88.86%**	**85.12%**	**81.89%**	**0.799**	**0.469**
Subject 7 FT 1HR Baseline	0.103	0.356	78.50%	84.70%	76.81%	74.02%	0.858	**0.506**
Subject 7 FT 1HR AAMax	**0.131**	**0.390**	**80.80%**	**87.86%**	**80.29%**	**76.55%**	**0.836**	0.508

D GENERALIZING TO DIFFERENT ADAPTER ARCHITECTURES

To test the sensitivity of Adapter Alignment to the choice of adapter, we compared several lightweight variants. As shown in Table 8, the choice of adding or removing a non-linearity, or substituting different activation functions (e.g., GELU vs. ReLU), does not significantly affect alignment or reconstruction performance. In practice, simpler architectures tend to perform slightly better, with a single-layer adapter outperforming deeper variants. Our adapter choice in the main experiments was guided by the underlying backbone: we use a nonlinear (linear + GELU) adapter with MindEye1 and retain the linear adapter design of MindEye2 for consistency with the original paper. Overall, these results indicate that the specific adapter design does not affect the validity of Adapter Alignment, and performance is robust across lightweight variants.

Table 8: Ablation on adapter variants for Subject 2 fine-tuned with 1 hour of data under AAMax training. We compare a nonlinear adapter with GELU, a nonlinear adapter with ReLU, and a two-layer linear adapter. Results show that adapter choice has little impact on performance, with all variants yielding similar alignment and reconstruction quality.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Subject 2 FT 1HR AAMAX linear	0.110	0.277	79.70%	88.61%	80.35%	79.88%	0.815	0.486
Subject 2 FT 1HR AAMAX GELU	0.112	0.259	78.75%	88.04%	78.99%	78.28%	0.824	0.513
Subject 2 FT 1HR AAMAX ReLU	0.101	0.273	80.44%	88.18%	78.72%	78.17%	0.827	0.508
Subject 2 FT 1HR AAMAX 2 linear	0.107	0.268	78.28%	87.12%	78.78%	78.62%	0.827	0.504

E GENERALIZING TO DIFFERENT REFERENCE SUBJECTS

We conducted additional experiments (Table 9) to examine how the choice of reference subject influences fine-tuning performance. In these experiments, Subjects 2, 5, and 7 were used as reference subjects, and Subject 1 was fine-tuned with one hour of data. The results show that the reference subject does affect reconstruction quality: Subject 5, which typically achieves the strongest individual reconstructions, yields the best fine-tuning performance, while Subject 7 produces the weakest, consistent with its lower single-subject performance. Nonetheless, Adapter Alignment remains effective in all cases, as AAMax consistently outperforms baseline fine-tuning.

Table 9: Effect of reference subject choice on limited-data fine-tuning. We compare AAMax and normal fine-tuning (160 epochs) when Subjects 2, 5, and 7 are used as reference subjects and Subject 1 is fine-tuned. AAMax consistently outperforms normal fine-tuning regardless of the reference, though overall performance reflects the quality of the chosen reference subject.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Subject 1 FT From Subject 2 1HR Baseline	0.007	0.268	75.55%	84.00%	75.20%	74.60%	0.862	0.520
Subject 1 FT From Subject 2 1HR AAMAX	0.060	0.277	**81.13%**	**88.99%**	**79.71%**	**78.89%**	**0.824**	**0.494**
Subject 1 FT From Subject 5 1HR Baseline	0.094	0.239	71.54%	80.10%	73.11%	74.02%	0.875	0.537
Subject 1 FT From Subject 5 1HR AAMAX	0.105	0.259	**76.77%**	**86.20%**	**80.08%**	**80.14%**	**0.828**	**0.496**
Subject 1 FT From Subject 7 1HR Baseline	0.010	0.266	74.96%	81.55%	71.71%	73.72%	0.874	0.528
Subject 1 FT From Subject 7 1HR AAMAX	0.012	0.283	**79.58%**	**87.78%**	**77.48%**	**77.87%**	**0.835**	**0.502**

Additional reference–target combinations. To examine whether these trends extend beyond Subject 1 as the target, we additionally fine-tuned Subjects 2, 5, and 7 with 1 hour of data from each of the other two subjects as the reference (Table 10). Together with Table 9 and the Subject 1-reference results in Table 2, this covers every ordered reference–target pair among Subjects 1, 2, 5, and 7. We report these AAMax results for reference and to show how reconstruction quality depends on the reference–target pairing, rather than as a comparison against baseline fine-tuning. Performance varies with the quality and compatibility of the pair, but AAMax remains effective for weaker targets such as Subject 7, and no single reference is best for every target and metric.

Table 10: AAMax fine-tuning (1 hour of data) for all remaining reference–target combinations among Subjects 2, 5, and 7. Results are shown for reference; performance depends on the pairing of reference and target subject. Low-level metrics are shown in gray because they are not optimized in these runs.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Subject 5 FT From Subject 2 1HR AAMAX	0.097	0.275	77.31%	86.98%	81.41%	80.06%	0.829	0.488
Subject 7 FT From Subject 2 1HR AAMAX	0.112	0.272	77.05%	84.75%	78.39%	76.95%	0.843	0.502
Subject 2 FT From Subject 5 1HR AAMAX	0.092	0.264	76.74%	86.27%	78.80%	78.24%	0.823	0.489
Subject 7 FT From Subject 5 1HR AAMAX	0.093	0.252	76.49%	85.31%	79.21%	77.99%	0.834	0.504
Subject 2 FT From Subject 7 1HR AAMAX	0.098	0.264	75.73%	84.33%	77.69%	77.36%	0.838	0.494
Subject 5 FT From Subject 7 1HR AAMAX	0.091	0.264	76.70%	85.75%	81.72%	79.86%	0.823	0.487

Multiple reference subjects. We also tested whether pre-training the reference model on more than one subject improves fine-tuning (Table 11). Pre-training on Subjects 1 and 2 jointly and fine-tuning Subject 5 with 1 hour of data does not consistently improve over using Subject 1 alone as the reference. This is consistent with MindEye2 (Appendix A.11 of Scotti et al. (2024)), where adding pre-training subjects yielded limited gains. We hypothesize that once a strong reference space is learned, additional subjects contribute mostly redundant rather than complementary structure. In practice, this suggests that a single high-quality reference subject suffices, while a weak reference, such as a lower-SNR dataset (Appendix S), degrades fine-tuning performance.

Table 11: Single- versus multi-subject reference pre-training. Subject 5 is fine-tuned with AAMax on 1 hour of data. The single-reference row corresponds to Subject 5 FT 1HR AAMax in Table 2.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Pre-train Subjects 1+2 → FT Subject 5 1HR AAMAX	0.102	0.263	78.79%	88.45%	**81.08%**	81.41%	**0.804**	0.495
Pre-train Subject 1 → FT Subject 5 1HR AAMAX	0.105	0.278	**79.90%**	**88.68%**	80.63%	**82.20%**	0.808	**0.493**

F Image Selection Algorithm Ablations and Baselines

F.1 Algorithm Coverage

We next evaluate how well the image selection algorithm covers the common space compared to random sampling. Using the top 20 eigenvectors, we discretize each dimension into bins and, for a subset of $K = 250$ images, count the number of empty bins (bins without a selected image). Random selection leaves, on average, around 200 empty bins with a standard deviation of about 10 across 1,000 trials. In contrast, our algorithm consistently reduces this number to 15 empty bins. A permutation test yields a p-value of 0.0002, confirming that the improved coverage is statistically significant. This demonstrates that our algorithm selects images that span the representation space more uniformly along high-variance directions. It is important to note that the number of empty bins in our algorithm remain relatively similar when changing the number of top eigenvectors.

F.2 Varying the Eigenvector Dimension

We conducted an ablation study on the image selection algorithm to evaluate the effect of varying the number of principal eigenvector dimensions used in the binning process. Specifically, we compared subsets selected using the top 10, 20, 30, and 40 eigenvectors. As shown in Table 12, performance is sensitive to this choice. For Subject 2, 20 and 40 eigenvectors are competitive and clearly outperform 10 and 30 dimensions, with 40 eigenvectors achieving the best score on most metrics. More generally, we observed that using a larger number of eigenvector dimensions tended to yield better fine-tuning performance. We therefore adopt 40 eigenvectors for the image-selection experiments in Tables 2 and 3; the reduced-data experiments in Figure 3 were run with 20 eigenvectors.

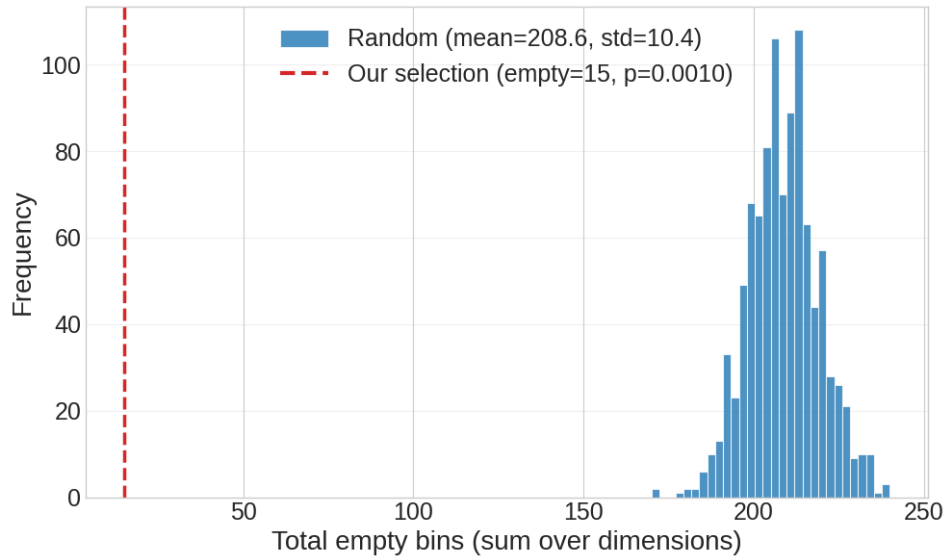


Figure 5: **Coverage of the image selection algorithm compared to random sampling.** We project embeddings onto the top 20 eigenvectors and count empty bins across 250 selected images. The histogram shows empty-bin counts for 1,000 random subsets, while the dashed line marks our algorithm's selection. Our method leaves only 15 bins empty, whereas random sampling leaves $\approx 200 \pm 10$ on average, indicating significantly more uniform coverage of the representation space ($p = 0.0002$).

Table 12: Ablation on the number of eigenvector dimensions used for image selection. We compare random image selection and Adapter Alignment with images selected using our algorithm across different eigenvector settings. Results are reported for Subject 1 as the reference and Subject 2 fine-tuned with 250 images under AAMax training, averaged over 3 runs. 20 and 40 eigenvectors are competitive, and 40 eigenvectors are used for the image-selection results in Tables 2 and 3.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Random Selection	0.112	0.259	78.75%	88.04%	78.99%	78.28%	0.824	0.513
10 EV Dimensions	0.114	0.303	78.99%	86.53%	77.30%	75.23%	0.850	0.541
20 EV Dimensions	0.102	0.267	81.41%	89.55%	80.43%	80.33%	0.823	0.496
30 EV Dimensions	0.112	0.275	79.60%	88.03%	78.58%	76.89%	0.850	0.537
40 EV Dimensions	0.118	0.285	81.78%	89.61%	81.49%	79.55%	0.810	0.491

F.3 Comparison with Standard Selection Baselines

Our greedy coverage objective resembles coreset and diversity-based selection over a learned representation. We therefore compare it against four standard selection strategies under the same image budget:

- **K-Center Greedy:** cosine farthest-first traversal over the normalized 4,096-dimensional Subject 1 adapter embeddings. At every step it selects the image farthest from its nearest already-selected image, minimizing the largest uncovered radius (minimax coverage) in the reference representation.
- **K-Means:** K-means with 250 clusters in the same normalized Subject 1 adapter space, selecting the real image nearest to each centroid. This favors representative prototypes and density coverage rather than boundary points.

- **DPP Diversity:** greedy maximization of the determinant of a linear kernel over the same normalized Subject 1 adapter embeddings, favoring a jointly diverse and non-redundant subset with large volume.
- **CLIP K-Center:** the same cosine farthest-first rule applied to 768-dimensional CLIP ViT-L/14 image embeddings, targeting visual–semantic diversity in pre-trained image space independently of the fMRI representation.

All methods select 250 images (1 hour of data) from the same pool of 1,000 shared images. We use the same Subject 1 reference model, Subject 2 fine-tuning budget, MindEye1 architecture with AAMax training, random seed, checkpoint epoch, test set, reconstruction settings, and evaluation pipeline, so only the image-selection rule changes. Our method uses 40 eigenvector dimensions.

Table 13: Comparison of image-selection strategies. Subject 1 is the reference and Subject 2 is fine-tuned with AAMax on 250 selected images; only the selection rule differs between rows. Best results are in bold.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Ours	**0.118**	0.285	81.78%	**89.61%**	**81.49%**	79.55%	**0.810**	0.491
K-Means	0.110	0.281	**82.40%**	89.14%	80.51%	**80.26%**	0.815	0.490
K-Center Greedy	0.114	0.271	81.93%	89.54%	80.10%	79.35%	0.817	**0.489**
DPP Diversity	0.115	0.270	80.80%	89.22%	80.21%	78.66%	0.820	0.493
CLIP K-Center	0.112	**0.290**	81.15%	88.73%	78.56%	79.92%	0.831	0.499

As shown in Table 13, the baselines are competitive, and no single method dominates across all metrics. Our algorithm achieves the best PixCorr, AlexNet(5), Inception, and EfficientNet scores (4 of 8 metrics), K-Means is best on AlexNet(2) and CLIP, K-Center Greedy is best on SwAV, and CLIP K-Center is best on SSIM. This pattern indicates that the outcome is not explained by a single generic notion of diversity: both the selection geometry and the representation in which diversity is measured affect different aspects of reconstruction quality. The selectors computed in the aligned reference space (ours, K-Means, K-Center Greedy, DPP) perform best on the high-level metrics, whereas CLIP K-Center, which ignores the fMRI representation, performs worst on Inception, EfficientNet, and SwAV. Our greedy algorithm is thus one effective instantiation of coverage-based selection in the aligned space, with the approximation guarantees of Appendix N, rather than the only viable choice.

F.4 Overlap Between Selected Subsets

To verify that these methods do not recover the same calibration subset, we computed the pairwise overlap between their selected images (Table 14). Most cross-method overlaps range from 17.2% to 30.8%. The larger overlap between DPP and K-Center Greedy (52.4%) is expected, since both operate on the same Subject 1 adapter representation, although they optimize different subset objectives. Only one image is selected by all five methods. The baselines therefore explore materially different calibration subsets rather than minor variants of the same selection.

Table 14: Pairwise overlap between the 250-image subsets selected by each method. Entries are the number of shared images, with percentages in parentheses.

	Ours	K-Center	K-Means	DPP	CLIP K-Center
Ours	250 (100.0%)	73 (29.2%)	74 (29.6%)	65 (26.0%)	58 (23.2%)
K-Center	73 (29.2%)	250 (100.0%)	43 (17.2%)	131 (52.4%)	77 (30.8%)
K-Means	74 (29.6%)	43 (17.2%)	250 (100.0%)	43 (17.2%)	60 (24.0%)
DPP	65 (26.0%)	131 (52.4%)	43 (17.2%)	250 (100.0%)	64 (25.6%)
CLIP K-Center	58 (23.2%)	77 (30.8%)	60 (24.0%)	64 (25.6%)	250 (100.0%)

G Comparing Different Alignment Training Strategies

To directly compare our approach against related methods, we conducted an experiment fine-tuning Subject 2 with 250 training images under three different training regimes (Table 15). First, we evaluated AA Step 1 only, which corresponds to the setting of Ferrante et al. (Ferrante et al., 2024), where adapters are overfitted to the common space without any subsequent end-to-end training. Second, we implemented a variant inspired by MindBridge (Wang et al., 2024), where the adapter remains unfrozen while the MLP and prior are frozen, and the full end-to-end training is carried out using a combined MSE, CLIP, and prior loss. Finally, we compared these baselines to our full AAMax training regime, which explicitly aligns adapters in the first phase and then fine-tunes the entire pipeline end-to-end.

The results show that AAMax achieves the strongest performance across both low-level and high-level metrics, outperforming both AA Step 1 and the MindBridge-style end-to-end setup. Importantly, this experiment was run under identical architectural settings and without auxiliary enhancements such as additional synthetic data or architectural add-ons. This isolates the effect of the training paradigm itself, demonstrating that when compared purely on training objectives, AAMax provides the most effective strategy for fine-tuning new subjects.

Table 15: Comparison of different training regimes for fine-tuning Subject 2 with 250 training images. We evaluate (1) AAMax (full Adapter Alignment with staged alignment + end-to-end fine-tuning), (2) AA Step 1 only (adapter overfitting without subsequent end-to-end training), and (3) End-to-End with MSE+Prior+CLIP Loss (MLP Frozen) (reset-tuning). AAMax achieves the best performance across both low-level and high-level metrics, demonstrating that explicit staged alignment followed by end-to-end fine-tuning is more effective than either alignment-only or combined end-to-end training when compared under identical architectural conditions.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
AAMAX	0.112	0.259	**78.75%**	**88.04%**	**78.99%**	**78.28%**	**0.824**	0.513
AA Step 1 only	0.008	0.240	77.14%	87.04%	77.00%	75.43%	0.857	0.510
End-to-End with MSE+Prior+CLIP Loss (MLP Frozen)	0.007	0.238	76.40%	86.42%	77.72%	75.04%	0.852	**0.507**

H Extended Limited Data Experiments

To further validate the efficiency of Adapter Alignment, we repeated our limited-data experiments using Subject 5 as the reference and Subject 1 for fine-tuning. We fine-tune with 1, 2, and 4 hours of data and compare baseline end-to-end training to AAMax. As always, AAMax first performs Step 1 adapter alignment by overfitting the adapter, followed by end-to-end fine-tuning. We track performance after 1 epoch of fine-tuning and after 160 epochs of fine-tuning. As shown in Table 16, even after a single epoch, AAMax consistently outperforms baseline models trained for 160–200 epochs.

We also track performance over fine-tuning epochs for all three data conditions (1, 2, and 4 hours). As shown in Figure 6, Adapter Alignment converges rapidly and maintains a consistent advantage throughout training, while baseline training requires hundreds of epochs to reach comparable accuracy. These results reinforce that AAMax substantially reduces both the data and computational requirements for fine-tuning new subjects.

I Ablation Study on the Two Phases of Adapter Alignment Fine-Tuning

We analyze the role of Step 1 (adapter alignment) relative to full end-to-end fine-tuning. In this experiment, we fine-tuned Subject 2 with 1 hour of data using Subject 1 as the reference. Figure 7 illustrates the reconstruction performance across epochs for both phases.

During Step 1, where only the adapter is optimized, performance improves steadily as the adapter overfits to the shared images. This phase alone provides a substantial boost over the unaligned baseline,

Table 16: Quantitative results on fine-tuning a new subject with limited data. FT=Fine tuning. The first section shows the result of fine-tuning with 250 (1 hour) images after 1 epoch and after 160 epochs on the baseline and using AAMax. The second section compares performance on 500 (2 hours) images and the third section compares performance on 1000 (4 hours) images. In all cases AAMax, significantly outperforms the baseline. The greyed out low level metrics are not optimized for in these experiments.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Subject 1 FT 1HR Baseline E1	0.031	0.234	58.31%	62.56%	58.51%	58.71%	0.947	0.611
Subject 1 FT 1HR AAMAX E1	0.095	0.254	**77.20%**	85.98%	78.80%	79.39%	0.833	0.503
Subject 1 FT 1HR Baseline E160	0.094	0.239	71.54%	80.10%	73.11%	74.02%	0.875	0.537
Subject 1 FT 1HR AAMAX E160	**0.105**	**0.259**	76.77%	**86.20%**	**80.08%**	**80.14%**	**0.828**	**0.496**
Subject 1 FT 2HR Baseline E1	0.049	0.242	62.41%	68.06%	62.64%	64.80%	0.925	0.583
Subject 1 FT 2HR AAMAX E1	0.112	0.275	79.46%	88.29%	**83.28%**	82.61%	0.803	0.485
Subject 1 FT 2HR Baseline E160	0.110	0.270	75.32%	84.62%	76.66%	78.55%	0.845	0.516
Subject 1 FT 2HR AAMAX E160	**0.114**	**0.285**	**80.00%**	**88.50%**	83.19%	**83.05%**	**0.799**	**0.480**
Subject 1 FT 4HR Baseline E1	0.079	0.241	65.84%	73.04%	64.19%	68.10%	0.914	0.574
Subject 1 FT 4HR AAMAX E1	**0.129**	0.265	80.83%	90.93%	**86.66%**	86.17%	0.767	0.460
Subject 1 FT 4HR Baseline E160	0.124	0.253	77.86%	89.18%	83.10%	83.26%	0.801	0.480
Subject 1 FT 4HR AAMAX E160	0.127	**0.273**	**81.73%**	**91.29%**	86.23%	**86.26%**	**0.763**	**0.453**

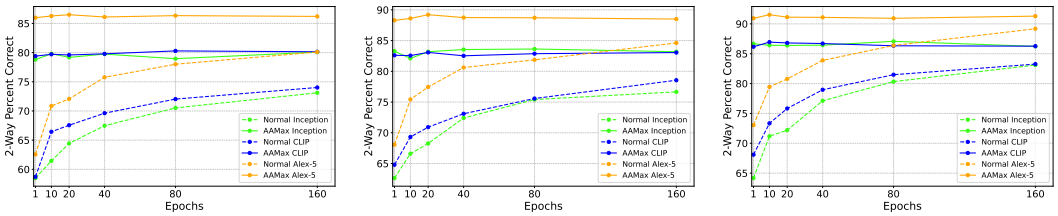


Figure 6: Comparing finetuning reconstruction performance over epochs with limited data. From left to right: (a) Finetuning Subject 1 with 250 images or 1 hour of data (b) Finetuning Subject 1 with 500 images or 2 hours of data (c) Finetuning Subject 1 with 1000 images or 4 hours of data. AAMax not only outperforms normal end-to-end training, but the performance at Epoch 1 is better than the baseline's performance at Epoch 160.

and after sufficient epochs the adapter can achieve strong alignment across subjects. However, the improvement plateaus, as Step 1 only modifies the adapter and does not allow deeper components of the pipeline to adapt.

Introducing end-to-end fine-tuning after adapter alignment leads to an immediate jump in performance. Even after a single epoch of end-to-end updates, the model surpasses what hundreds of adapter-only epochs can achieve. We stop at around 200 epochs as the relative gain plateaus again. While the improvement in performance by a few percent from end-to-end fine-tuning might not seem significant when compared to the gain from Step 1, it can make the difference between coherent images and gibberish when working with limited data and when reconstruction fidelity is low.

J Additional Scaling Analysis

The architectural setup used by MindEye2 leverages a 4096-dimensional shared space with BigCLIP and Stable Diffusion XL (SDXL), requiring substantially greater computational resources than our main experiments, which employ Versatile Diffusion. To evaluate whether the benefits of Adapter Alignment (AAMax) persist across different model capacities, we conducted a scaling law analysis varying both the dimensionality of the shared space and the choice of generative backbone.

J.1 Empirical Observations

1. **Architectural Improvements at 1024 Dimensions:** Using a 1024-dimensional shared space, we find that BigCLIP + SDXL outperforms Versatile Diffusion, confirming that

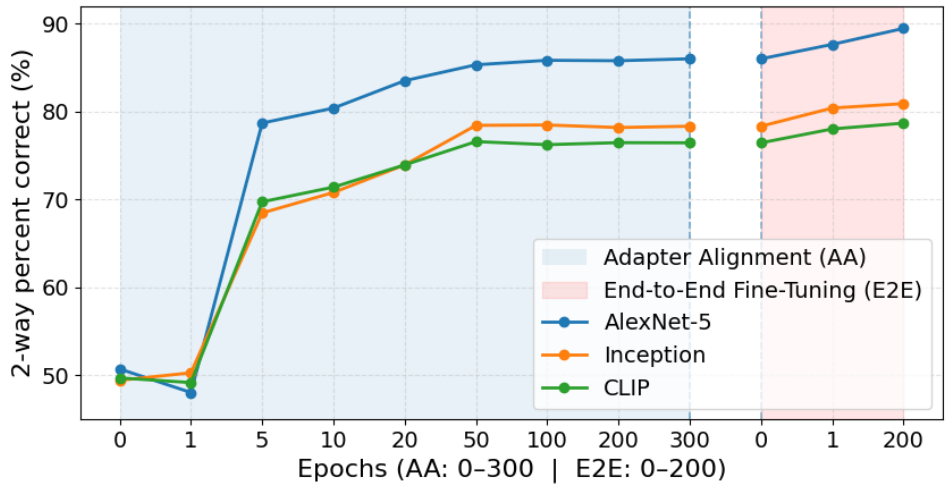


Figure 7: Reconstruction performance across epochs for Subject 2 fine-tuned with 1 hour of data using Subject 1 as the reference. The shaded regions denote Step 1 (adapter alignment) and Step 2 (end-to-end fine-tuning). Performance steadily increases during adapter alignment but plateaus short of optimal results. Transitioning to end-to-end fine-tuning yields an immediate jump in accuracy and ultimately outperforms adapter-only training. The dashed lines indicate the transition from AA to E2E. Performance at Epoch 300 of AA is the same as Epoch 0 of E2E, since no further fine-tuning has been done yet.

a higher-capacity CLIP space and diffusion model contribute to improved reconstruction performance.

2. **Scaling Behavior with Versatile Diffusion:** We vary the shared space dimensionality from $512 \rightarrow 1024 \rightarrow 4096$, using adapter alignment (AAMax) fine-tuning and observe consistent improvements in reconstruction quality as dimensionality increases.

These findings indicate that AAMax is robust to architectural choice and continues to provide benefits regardless of the CLIP encoder or diffusion model employed.

J.2 Projection to 4096 Dimensions with Stable Diffusion XL

The key insight from these experiments is that **adapter alignment is a training paradigm, not an architectural modification**. Our results demonstrate that AA is agnostic to the representational scale and generative backbone, and scales effectively with larger models. While large-scale pipelines like MindEye2 set upper bounds on reconstruction quality under high compute budgets, adapter alignment provides a complementary framework that improves efficiency and accessibility. Importantly, its benefits persist even when extended to high-capacity architectures, making it a broadly applicable strategy for fMRI-based reconstruction.

K Common Space and Neural Correspondence Analysis

We run several empirical tests to validate the effectiveness of Adapter Alignment in aligning the common space across subjects. We distinguish two kinds of evidence. Analyses on the shared images that are used for alignment (Sections K.2 and K.3) verify that the alignment objective is achieved; because Step 1 explicitly fits the new subject's adapter to the reference embeddings of these images, high agreement on them is expected by construction. Analyses on images that are not used for alignment, namely each subject's held-out test images (Section K.1) and shared images excluded from the calibration set (Section K.4), test whether the induced cross-subject structure generalizes beyond the calibration pairs. These analyses provide empirical evidence of transferable functional representational structure; they do not by themselves establish an intrinsic shared neural space or anatomical correspondence.

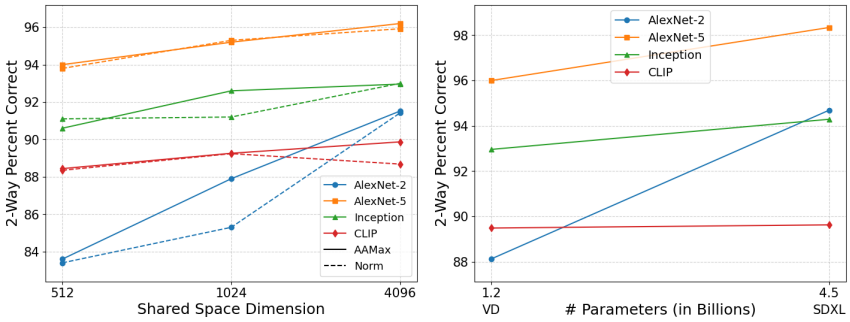


Figure 8: Comparing performance with scale for AAMax. **Left:** AAMax vs regular finetuning performance over increasing shared space size. AAMax scales with shared space dimensionality. **Right:** AAMax finetuning performance at 1024 shared space with Versatile Diffusion and Stable Diffusion XL. All metrics improve when moving to a larger diffusion model.

K.1 Eigenvector Analysis

To first investigate the structural alignment of embeddings across subjects, we analyzed the subject-wise cosine similarity of the principal eigenvectors in the embedding spaces produced by two techniques: the standard end-to-end training (Baseline) and our proposed approach (AAMax). For both techniques, we extracted the top five principal eigenvectors from the test data for each subject and computed the cosine similarity between the eigenvectors from different subjects. In normal

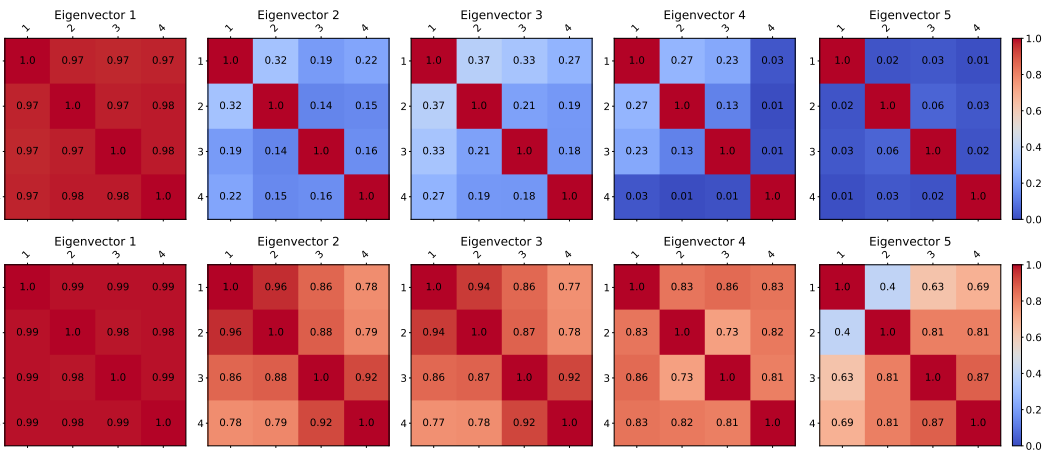


Figure 9: Comparing eigenvector cosine similarity for the test data across subjects in both training techniques. We use the first 5 principal eigenvectors. Row 1 shows the results for normal end-to-end training. Except for the first eigenvector, cosine similarity is very low across subjects. Row 2 shows the results for AAMax training. Cosine similarity is high even at the 5th principal eigenvector.

end-to-end training, we observed consistently low cosine similarity between all but the first principal eigenvector of different subjects, suggesting that the primary directions of variance captured by the embeddings were highly subject-specific and lacked alignment across subjects. This indicates that the embedding space learned in this approach does not generalize well to a common structure, resulting in subject-specific representations that diverge significantly in terms of principal variance.

In contrast, AAMax training showed significantly higher cosine similarity between the principal eigenvectors of different subjects. This indicates that AAMax effectively aligns the embeddings across subjects in terms of the primary directions of variance, achieving a shared representation space that better captures common structural features. The high similarity across subjects indicates that AAMax encodes the principal semantic and structural patterns consistently. Because these eigenvectors are computed on each subject's held-out test images, which are subject-specific, unseen

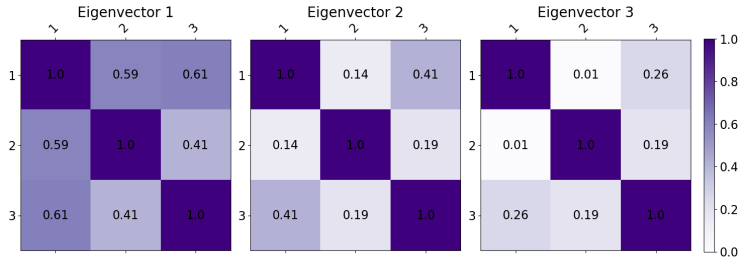


Figure 10: Comparing eigenvector cosine similarity for the test data across datasets in both training techniques. 1. Reference Subject 1 NSD 2. Normal Finetune Subject 1 THINGS 3. AAMax Finetune Subject 1 THINGS. AAMax training generates better aligned embedding spaces.

during alignment, and have no one-to-one correspondence across subjects, this shows that the aligned geometry generalizes beyond the image pairs on which alignment is explicitly enforced.

We also perform a similar analysis across datasets in Figure 10. We compare the alignment between Subject 1 from NSD and Subject 1 from THINGS when the latter has been finetuned on a model that was pre-trained with the former. We present the difference between both training techniques. AAMax training generates better aligned embedding spaces.

K.2 Mean Squared Error

We compare the end-to-end training approach with AAMax by evaluating the MSE between subjects using shared images after training as shown in Figure 11. In the end-to-end model, we observe relatively high MSE between subjects, indicating poor alignment in the common space. In contrast, the MSE for AAMax is an order of magnitude lower, demonstrating a much stronger alignment of subject-specific embeddings. Because these shared images include the calibration images on which the adapter-level MSE is optimized, this result primarily confirms that the alignment objective is achieved. Section K.4 shows that AAMax also reduces cross-subject MSE on shared images excluded from the calibration set.

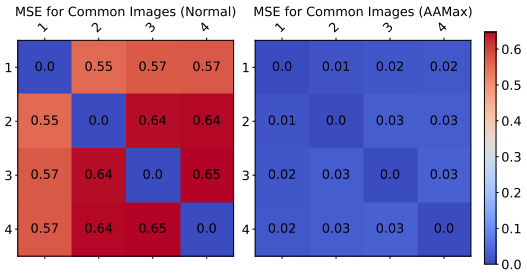


Figure 11: Comparing Mean Squared Error for the shared images across subjects in both training techniques. The MSE is compared after training is completed in both techniques. MSE for AAMax is an order of magnitude lower than the baseline

K.3 k-Nearest Neighbor Consistency

We also evaluate the k-nearest neighbors (k-NN) consistency (Figure 12) between subjects for shared images after training, using 50 nearest neighbors as the metric. In the end-to-end model, the k-NN consistency is relatively low, indicating that the nearest neighbors of embeddings from one subject do not align well with those from another subject. However, with AAMax, the k-NN consistency significantly improves, demonstrating much better alignment of subject-specific embeddings. AAMax consistently outperforms the normal approach. As with the MSE analysis, this evaluation includes the calibration images; the corresponding result on held-out shared images is reported in Section K.4.

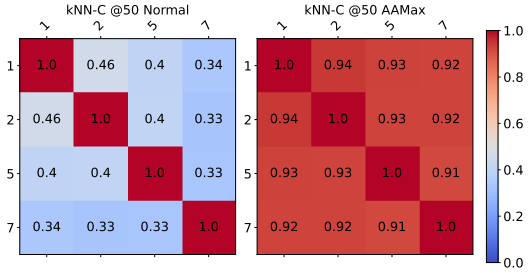


Figure 12: Comparing kNN consistency @ 50 for the shared images across subjects in both training techniques. The kNN-C is compared after training is completed in both techniques. AAMax significantly outperforms the baseline.

K.4 Neural Correspondence: pRF- and fLOC-Resolved Alignment

The analyses above treat each subject's `nsdgeneral` voxel vector as a single input. Here we test whether AAMax improves cross-subject alignment of signals originating from homologous, independently defined functional regions of visual cortex: retinotopic areas and eccentricity bands from population receptive field (pRF) mapping, and category-selective regions from the functional localizer (fLOC) experiment. The subject-specific pRF and fLOC labels are distributed with NSD (Allen et al., 2022) and were estimated independently of our model.

Setup. We compare baseline fine-tuning and AAMax using MindEye2, with Subject 1 (pre-trained on its full NSD data) as the reference and Subject 2 fine-tuned on 250 shared images for 200 epochs. The two conditions differ only in the training procedure: the baseline performs standard end-to-end fine-tuning, and AAMax adds the adapter alignment stage and the adapter-level MSE loss. The Subject 1 adapter weights and bias are bit-for-bit identical in the two checkpoints, so the reference representation is fixed and any difference reflects how Subject 2 is mapped toward it. All analyses use the 750 shared images that are *not* part of the 250-image calibration set, averaging the three trial repetitions of each image. We analyze the 4,096-dimensional output of the subject-specific adapter, before the shared residual network, the CLIP projection, the diffusion prior, or the image generator.

Region-resolved decomposition. Applying the same `nsdgeneral` selection to the pRF and fLOC label volumes assigns every model input voxel its subject-specific labels. Individual voxels are not matched across subjects; instead, we compare contributions from homologous regions. Because the MindEye2 adapter is linear, $z_s = W_s x_s + b_s$, the bias-free signal decomposes exactly into regional contributions, $W_s x_s = \sum_r W_{s,r} x_{s,r}$. For each region r we compute the paired cross-subject cosine similarity $\cos(W_{1,r}x_{1,r}, W_{2,r}x_{2,r})$ for each image, excluding the bias so that a shared offset cannot inflate the similarity of weak regional signals. We also report 750-way cross-subject identification (whether Subject 1's embedding of an image is closest to Subject 2's embedding of the same image; chance $= 0.13\%$), a leave-one-region-out analysis that removes a region's exact contribution from the full embedding, paired-bootstrap 95% confidence intervals over 2,000 image resamples, and a shuffled-correspondence null with 1,000 permutations.

Table 17: Cross-subject alignment of the full adapter output on 750 held-out shared images (Subject 1 $\rightarrow$ Subject 2, MindEye2). Identification is 750-way.

Metric	Baseline	AAMax
Paired cosine ↑ (95% CI)	0.366 [0.360, 0.373]	**0.576** [0.569, 0.582]
Paired MSE ↓	8.42	**6.30**
kNN consistency @50 ↑	0.339	**0.427**
Subject 1 $\rightarrow$ 2 identification, top-1 / top-5 ↑	50.80% / 76.27%	**92.40% / 98.93%**
Subject 2 $\rightarrow$ 1 identification, top-1 / top-5 ↑	49.47% / 78.13%	**87.73% / 97.47%**

Full-space results. On held-out shared images, AAMax improves every cross-subject alignment measure (Table 17), and the paired cosine increases for 99.9% of the images. In both conditions, correctly paired embeddings are far more similar than shuffled pairs (shuffled cosine ≈ 0.005–0.007; $p \leq 0.001$), so the alignment is stimulus-specific rather than a shared average direction.

Table 18: Region-resolved paired cosine similarity on 750 held-out shared images. Each entry compares only the contribution of the given pRF-defined area, eccentricity band, or fLOC-defined category system. 95% paired-bootstrap confidence intervals of the difference are shown for the fLOC systems.

	Region	Baseline	AAMax	Δ
pRF areas	All pRF voxels	0.344	**0.505**	+0.160
	V1	0.286	**0.394**	+0.108
	V2	0.241	**0.335**	+0.094
	V3	0.186	**0.266**	+0.080
	hV4	0.171	**0.250**	+0.079
Eccentricity	0.5°	0.102	**0.180**	+0.078
	1°	0.094	**0.194**	+0.100
	2°	0.243	**0.360**	+0.117
	4°	0.270	**0.372**	+0.102
	$> 4^\circ$	0.339	**0.441**	+0.102
fLOC systems	All fLOC voxels	0.342	**0.488**	+0.146 [0.142, 0.151]
	Faces	0.217	**0.303**	+0.085 [0.080, 0.090]
	Places	0.278	**0.357**	+0.079 [0.075, 0.084]
	Bodies	0.305	**0.437**	+0.132 [0.126, 0.137]
	Words	0.101	**0.168**	+0.067 [0.062, 0.073]

pRF-resolved results. AAMax increases the paired cosine of the contributions from every retinotopic area (V1, V2, V3, and hV4) and every eccentricity band (Table 18), and it does so in all 20 area $\times$ eccentricity cells. Using only pRF-defined voxels, 750-way Subject $1 \rightarrow 2$ identification improves from 28.93% to 64.53%. The improvement is therefore not confined to an undifferentiated whole-brain vector: it is present within independently defined early and intermediate visual areas across the visual field. Absolute values should not be compared across regions, since regions differ in voxel count, pRF reliability, signal-to-noise ratio, and cortical magnification; the informative comparison is AAMax versus the baseline within the same region.

Table 19: Paired cosine similarity for individual bilateral fLOC subregions on 750 held-out shared images, with 95% paired-bootstrap confidence intervals of the difference. Subregions absent in either subject (mTL-faces, aTL-faces, RSC, FBA-1, mTL-bodies, mTL-words) are excluded.

Subregion	Baseline	AAMax	Δ [95% CI]
OFA	0.121	**0.133**	+0.011 [0.006, 0.016]
FFA-1	0.120	**0.186**	+0.066 [0.061, 0.070]
FFA-2	0.156	**0.236**	+0.079 [0.074, 0.086]
OPA	0.099	**0.173**	+0.074 [0.069, 0.080]
PPA	0.292	**0.359**	+0.068 [0.064, 0.072]
EBA	0.271	**0.395**	+0.124 [0.118, 0.130]
FBA-2	0.122	**0.193**	+0.071 [0.065, 0.077]
OWFA	0.086	**0.121**	+0.035 [0.031, 0.040]
VWFA-1	0.065	**0.123**	+0.057 [0.052, 0.063]
VWFA-2	0.007	**0.016**	+0.009 [0.005, 0.014]
mfs-words	0.054	**0.069**	+0.015 [0.011, 0.019]

fLOC-resolved results. AAMax improves the paired cosine in all four category-selective systems and in every retained bilateral subregion, and every paired-bootstrap confidence interval excludes zero (Tables 18 and 19). For every system and both conditions, correctly paired contributions exceed all 1,000 shuffled-correspondence permutations ($p \leq 0.001$). Using only fLOC-defined voxels, 750-way identification improves from 14.53% to 31.73%. The gains are largest in EBA, FFA-2, OPA, and FBA-2, and small in OFA, VWFA-2, and mfs-words. Not every measure improves uniformly: for word-selective voxels, kNN consistency is essentially unchanged (0.164 vs. 0.163) despite the improved paired cosine and identification.

Table 20: Leave-one-region-out analysis: decrease in the full paired cosine similarity when a region's exact contribution is removed from the adapter output. Larger values indicate that the region contributes more to the cross-subject alignment.

	Region removed	Baseline	AAMax
pRF	V1	0.020	0.026
	V2	0.011	0.018
	V3	0.008	0.019
	hV4	0.005	0.010
	All pRF voxels	0.029	0.056
fLOC	Faces	0.006	0.011
	Places	0.023	0.037
	Bodies	0.033	0.047
	Words	−0.006	0.000
	All fLOC voxels	0.051	0.081

Leave-one-region-out. Removing the contributions of retinotopic areas or of face-, place-, or body-selective regions reduces cross-subject alignment more under AAMax than under the baseline (Table 20), indicating that these homologous signals contribute more consistently to the aligned representation. Removing word-selective signals has essentially no effect, suggesting that they carry stimulus-specific correspondence in isolation but contribute little to the full cross-subject similarity.

Scope. These results use independently obtained functional atlases to decompose and validate the learned representation. They show that AAMax improves the stimulus-specific functional alignment of signals arising from homologous retinotopic and category-selective regions, on images not used for alignment. They do not show that AAMax establishes cortical vertex correspondence, and AAMax does not use cortical mesh geometry during training: anatomical alignment operates upstream of the reconstruction pipeline, whereas AAMax operates after the subject-specific adapter has absorbed anatomical variability. The regional decomposition is exact for the linear adapter, but the downstream nonlinear network may combine regional contributions non-additively.

L Pre-Diffusion Retrieval

Reconstruction metrics depend on both the fMRI-derived conditioning signal and the pre-trained generative model, so they cannot on their own show how much image-specific information is recovered from the fMRI signal. To measure this directly, we evaluate retrieval on the fMRI-derived embeddings before any diffusion prior or image generator is applied.

Setup. We fine-tune Subject 2 with 1 hour of data (250 images) using Subject 1 as the reference, for both MindEye1 (CLIP ViT-L/14) and MindEye2 (OpenCLIP ViT-bigG/14), and compare baseline fine-tuning with AAMax using the epoch-200 checkpoints. Retrieval uses the contrastive embedding produced immediately after the subject-specific adapter and the shared voxel-to-CLIP network, and is evaluated on the 1,000-image Subject 2 test set. The baseline uses the same architecture, subject-specific adapter, voxel-to-CLIP network, contrastive objective, data, and checkpoint epoch as AAMax, so the comparison isolates the effect of the training procedure. We report two protocols:

- **Global retrieval:** top-1, top-5, and top-10 accuracy among all 1,000 test images.

- **MindEye protocol** (Scotti et al., 2023; 2024): top-1 accuracy among 300 candidates (the correct match and 299 randomly sampled distractors), repeated 30 times with different candidate samples.

Image retrieval denotes fMRI→image matching, and brain retrieval denotes the reverse direction.

Table 21: Pre-diffusion retrieval for Subject 2 fine-tuned with 1 hour of data from Subject 1. Global retrieval reports top-1/5/10 accuracy (%) among all 1,000 test images. The 300-way protocol reports mean ± standard deviation (%) of top-1 accuracy across 30 random candidate samplings of a single trained model.

		1,000-way (T1 / T5 / T10)		**300-way (top-1)**	
Model	**Training**	Image retrieval ↑	Brain retrieval ↑	Image retrieval ↑	Brain retrieval ↑
MindEye1	Baseline	21.7 / 45.2 / 55.1	25.4 / 50.1 / 61.7	34.78 ± 0.76	39.24 ± 0.99
MindEye1	AAMax	**29.0 / 55.9 / 66.6**	**36.1 / 64.8 / 74.2**	**44.20 ± 0.80**	**52.08 ± 0.92**
MindEye2	Baseline	33.2 / 63.8 / 75.0	39.8 / 70.0 / 80.2	50.87 ± 0.88	56.18 ± 1.23
MindEye2	AAMax	**39.6 / 69.6 / 78.6**	**45.1 / 75.2 / 85.5**	**56.60 ± 0.79**	**63.00 ± 0.74**

Results. AAMax improves retrieval consistently for both architectures, in both retrieval directions, and under both candidate-pool protocols (Table 21). Under the 300-way protocol, AAMax improves image retrieval by 9.4 and 5.7 percentage points for MindEye1 and MindEye2, respectively, and brain retrieval by 12.8 and 6.8 points. Because these measurements are taken before diffusion sampling, they show that the improvements of AAMax reflect additional image-specific information in the fMRI-derived embeddings, rather than a stronger reliance on the generative prior. The reported standard deviations capture variability across candidate samplings of a single trained model per condition, not variability across training runs.

M Emergence of Common Concepts: Additional Details

The existence of a semantic common representation space is suggested from the emergence of semantic categories in this space. For this, we extracted the most significant left singular vectors from the ridge regression weight matrix. This matrix was extracted from the pre-trained MindEye2 (Scotti et al., 2024) model which was trained using the 1024-dimensional common space. Subsequently, we projected the 1024-dimensional embeddings of each image onto these vectors and sampled from the two extremes (set 1 and set 3) and the middle (set 2). In order to obtain an objective classification, we presented these sets of images to GPT-4o (Achiam et al., 2023) and asked it to find the semantic categories that are well represented in set 1, somewhat represented in set 2 and not represented in set 3. Results for the first five singular vectors are shown in Table 22. As is evident, images in the principal dimensions of the representation space have a common theme.

Previous studies, such as those by Huth et al. (2012), have indicated that animal categories, including humans, are represented distinctly from non-animal categories within the brain. In our study, we observed that the common representational space also preserves these semantic properties. As demonstrated in Figure 13, after projecting the fMRI signals of Subject 2 into the common space and onto the first two singular vectors, we can see the animal vs. non-animal distinction, consistent with previous hypotheses. Extending this analysis across additional dimensions, we observed similar separations between other hypothesized semantic categories of "Social interaction," "man-made vs nature," and "biological vs non-biological."

To determine if dominant semantic categories were present within these sets, we employed ChatGPT with the query: “Is there a dominant semantic category to which the majority of images in a set belong?” The methodology for defining semantic categories was consistent with that used in Huth et al. (2012). Table 22 illustrates our findings.

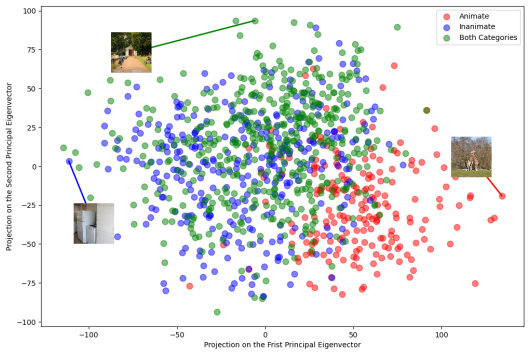


Figure 13: Visualization of the Animate/Inanimate semantic category on the two principal singular vectors. Each point (embedding) is colored based on whether its corresponding image contains 'objects' that fall into the Animate category (red), Inanimate category (blue) or both categories (green).

Table 22: Concept analysis of images along principal dimensions.

DIMENSION 1	ANIMALS (GIRAFFES, ZEBRAS, AND ELEPHANTS), NATURE, AND OUTDOOR SCENES
DIMENSION 2	CATS AND FOOD
DIMENSION 3	KITE FLYING, AIRPLANES, AND FLYING OBJECTS
DIMENSION 4	SURFING, WATER SPORTS, BEACH, AND OCEAN SCENES
DIMENSION 5	AIRPLANES, FLYING OBJECTS, OUTDOOR ACTIVITIES AND EVENTS, TRANSPORTATION SCENES

After projecting the fMRI signals of individual images from Subject 2 into the common space and onto the first singular vectors, we observe that certain concepts are strongly represented along these dimensions defined by the singular vectors, aligning with previous hypotheses (Figures 14, 15).

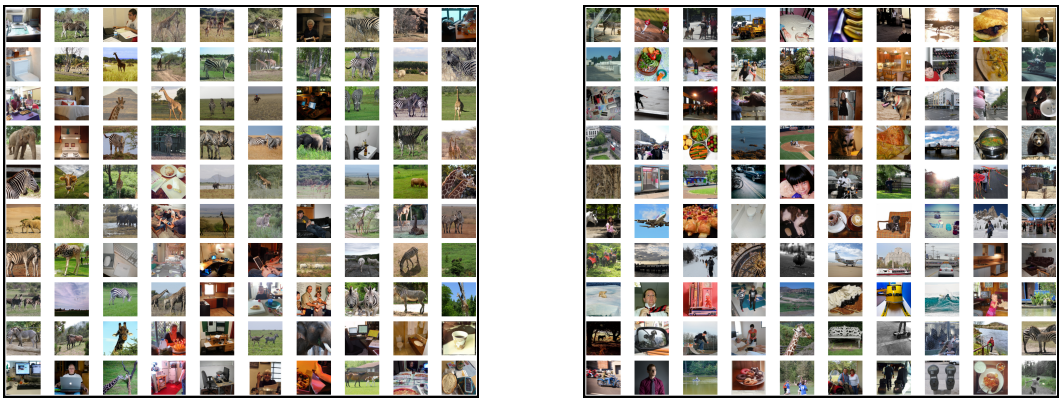


Figure 14: Comparison of images from Subject 2 with the highest and lowest projection values along the first principal left eigenvector. (a) Images whose corresponding 1024-dimensional embeddings yielded the highest projection values. (b) Images whose corresponding 1024-dimensional embeddings yielded the lowest projection values. The presence of the animate concept (e.g., giraffes, zebras, elephants) is prominent in the first set, while it fades in the second.

N IMAGE SELECTION ALGORITHM: FORMULATIONS AND PROOFS

This appendix formalizes the image selection problem of Section 3.4, shows that it is NP-hard, and states the approximation guarantees of the greedy procedure (Algorithm 1). The same greedy procedure supports two formulations that differ only in their stopping criterion: a *set cover* formula-

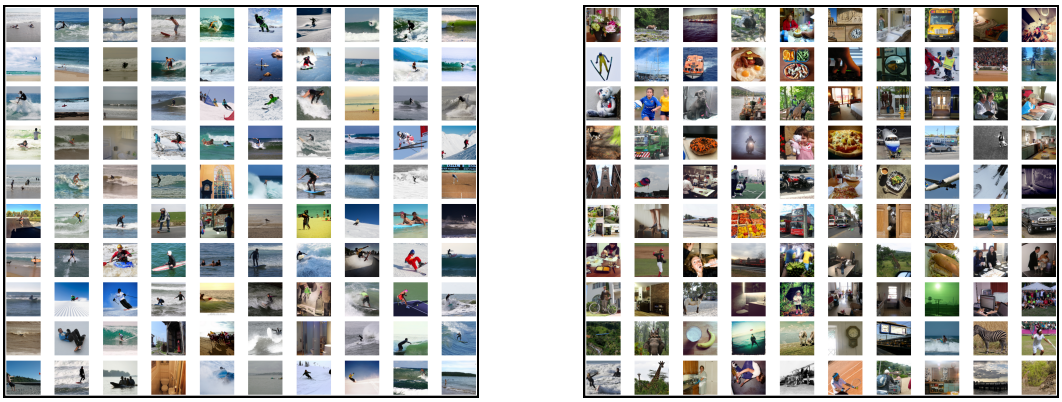


Figure 15: Comparison of images from Subject 2 with the highest and lowest projection values along the 4th principal left eigenvector. (a) Images whose corresponding 1024-dimensional embeddings yielded the highest projection values. (b) Images whose corresponding 1024-dimensional embeddings yielded the lowest projection values. Again the presence of common concepts (e.g. surfing, water sports, beach, and ocean scenes) is prominent in the first set, while it fades in the second.

tion, which seeks the smallest subset spanning the discretized eigenspace, and a *maximum coverage* formulation, which selects a fixed budget of images to maximize coverage.

N.1 Setup and Formulations

Let $\mathcal{I}$ denote the pool of candidate images (the 1,000 NSD shared images). The reference subject's shared-space embeddings of these images are projected onto the top d principal directions of the reference adapter, and the projected range of each dimension $j \in \{1, \dots, d\}$ is partitioned into $B_j = \lfloor w \cdot \lambda_j / \lambda_1 \rfloor$ bins, where λ_j is the j-th singular value and w is a scaling parameter. The universe of bins is

$$U = \bigcup_{j=1}^{d} \{(j, b) : b \in [B_j]\}, \qquad |U| = \sum_{j=1}^{d} B_j.$$

Each image $i \in \mathcal{I}$ falls into exactly one bin in every dimension, so it covers the set $C_i = \{(j, \mathrm{bin}_j(i)) : j = 1, \dots, d\} \subseteq U$ with $|C_i| = d$. For a subset $S \subseteq \mathcal{I}$, the coverage function and the gap function of Section 3.4 are

$$f(S) = \Big| \bigcup_{i \in S} C_i \Big|, \qquad \mathrm{Gap}(S) = \sum_{j=1}^{d} \mathrm{Gap}(S, j) = |U| - f(S),$$

where $\mathrm{Gap}(S, j)$ is the number of bins in dimension j that contain no image of S.

Formulation 1 (Set Cover). Find the smallest subset that covers every bin:

$$\min_{S \subseteq \mathcal{I}} |S| \quad \text{s.t.} \quad f(S) = |U|.$$

Formulation 2 (Maximum Coverage). Given a budget k, find the subset of at most k images that covers the most bins:

$$\max_{S \subseteq \mathcal{I},\ |S| \le k} f(S).$$

Algorithm 1 solves both formulations with the same greedy rule: at every step it adds the image that minimizes $\mathrm{Gap}(S \cup \{i\})$, which is equivalent to maximizing the marginal coverage gain $f(S \cup \{i\}) - f(S)$. Running until $\mathrm{Gap}(S) = 0$ yields a set cover (Formulation 1); stopping after k images yields a maximum coverage solution (Formulation 2). Because the greedy procedure is deterministic, the k-image solution is exactly the first k images of the full-coverage sequence. In our experiments we use the fixed-budget variant with $k = 250$ images (one hour of data). With $d = 40$ and $w = 105$, a full cover requires 255 images, so the two formulations select nearly identical subsets in practice.

N.2 NP-HARDNESS

Definition 1 (Bin Covering Problem)**.** *Given a set of vectors, each of which places one element into one of the bins of every dimension, and an integer k, decide whether there exists a subset of at most k vectors that covers every bin of every dimension.*

Definition 2 (Set Cover Problem)**.** *Given a finite universe $\mathcal{U}$, a collection $\mathcal{S} = \{S_1, \ldots, S_l\}$ of subsets of $\mathcal{U}$ with $\bigcup_i S_i = \mathcal{U}$, and an integer k, decide whether there exists a subcollection $\mathcal{S}' \subseteq \mathcal{S}$ with $|\mathcal{S}'| \leq k$ and $\bigcup_{S \in \mathcal{S}'} S = \mathcal{U}$.*

Theorem 1. *The bin covering problem is NP-hard, even when every dimension has exactly three bins.*

Proof. We reduce from Set Cover. Given an instance $(\mathcal{U}, \mathcal{S}, k)$ with $\mathcal{U} \neq \emptyset$, we create one dimension for every element $e \in \mathcal{U}$ and two additional dimensions p and q, each with three bins. We construct one vector per input set and two special vectors:

- For each $S_i \in \mathcal{S}$, the vector V_i selects bin 1 in dimension e if $e \in S_i$ and bin 2 otherwise, and selects bin 2 in dimensions p and q.
- The vector N_p selects bin 2 in every element dimension, bin 1 in dimension p, and bin 3 in dimension q.
- The vector N_q selects bin 3 in every element dimension, bin 3 in dimension p, and bin 1 in dimension q.

The construction is polynomial in $|\mathcal{U}|$ and l. By construction, bin 1 of dimension p is covered only by N_p, bin 1 of dimension q is covered only by N_q, and bin 1 of an element dimension e is covered only by the vectors V_i with $e \in S_i$. We show that $\mathcal{U}$ has a set cover of size at most k if and only if there is a bin cover of size at most $k + 2$.

($\Rightarrow$) Let $\mathcal{S}'$ be a set cover with $|\mathcal{S}'| \leq k$, and take the vectors $\{V_i : S_i \in \mathcal{S}'\} \cup \{N_p, N_q\}$. In every element dimension, bin 1 is covered because $\mathcal{S}'$ covers $\mathcal{U}$, bin 2 is covered by N_p, and bin 3 is covered by N_q. In dimensions p and q, bin 1 is covered by N_p and N_q respectively, bin 2 is covered by any V_i in the selection (which is non-empty because $\mathcal{U} \neq \emptyset$), and bin 3 is covered by N_q and N_p respectively. The selection therefore covers every bin and has size at most $k + 2$.

($\Leftarrow$) Let $\mathcal{V}'$ be a bin cover with $|\mathcal{V}'| \leq k + 2$. Because bin 1 of dimensions p and q is covered only by N_p and N_q, both special vectors belong to $\mathcal{V}'$, so the remaining vectors are at most k vectors of the form V_i. Neither special vector covers bin 1 of any element dimension, so for every $e \in \mathcal{U}$ some remaining V_i has $e \in S_i$. The corresponding sets therefore form a set cover of size at most k.

Since Set Cover is NP-hard, the bin covering problem is NP-hard. The general problem, with an arbitrary number of bins B_j in each dimension, contains this three-bin case and is therefore NP-hard as well. Deciding whether k images can cover all of U is exactly this decision problem, so the optimization versions of both Formulation 1 and Formulation 2 are NP-hard. □

N.3 MONOTONICITY AND SUBMODULARITY OF THE COVERAGE FUNCTION

A set function f is monotone if $f(A) \leq f(B)$ for all $A \subseteq B$, and submodular if it has diminishing returns (Schrijver et al., 2003):

$$f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B), \qquad \forall A \subseteq B,\ x \notin B.$$

Theorem 2. *The coverage function $f(S) = |\bigcup_{i \in S} C_i|$ is monotone and submodular. Equivalently, $\mathrm{Gap}(S) = |U| - f(S)$ is non-increasing and supermodular.*

Proof. Write $f(S) = \sum_{j=1}^{d} f_j(S)$, where $f_j(S)$ is the number of covered bins in dimension j. Adding an image x to a subset A increases f_j by

$$\Delta_A(x, j) = f_j(A \cup \{x\}) - f_j(A) = \begin{cases} 1 & \text{if no image of } A \text{ lies in bin } \mathrm{bin}_j(x), \\ 0 & \text{otherwise.} \end{cases}$$

Since $\Delta_A(x, j) \geq 0$, f is monotone. If $A \subseteq B$, every bin covered by A is also covered by B, so $\Delta_A(x, j) \geq \Delta_B(x, j)$ for every dimension j. Summing over dimensions gives

$$f(A \cup \{x\}) - f(A) = \sum_{j=1}^{d} \Delta_A(x, j) \;\geq\; \sum_{j=1}^{d} \Delta_B(x, j) = f(B \cup \{x\}) - f(B),$$

so f is submodular. Because $\mathrm{Gap}(S) = |U| - f(S)$, the marginal change of the gap is $\mathrm{Gap}(A \cup \{x\}) - \mathrm{Gap}(A) = -\sum_j \Delta_A(x, j) \leq 0$, and the inequality above reverses, i.e., $\mathrm{Gap}(A \cup \{x\}) - \mathrm{Gap}(A) \leq \mathrm{Gap}(B \cup \{x\}) - \mathrm{Gap}(B)$. The gap function is therefore non-increasing and supermodular, and minimizing $\mathrm{Gap}(S \cup \{i\})$ is the same as maximizing the marginal gain of f. □

N.4 Approximation Guarantees

Theorem 3 (Set Cover)**.** *Run until* $\mathrm{Gap}(S) = 0$*, Algorithm 1 returns a cover of size at most* $H(d) \cdot \mathrm{OPT}$*, where* OPT *is the size of a minimum cover and* $H(d) = \sum_{m=1}^{d} 1/m \leq \ln d + 1$.

This is the classical greedy guarantee for set cover, which is $H(s)$ in the size s of the largest input set (Johnson, 1974; Lovász, 1975). In our setting every image covers exactly one bin per dimension, so $s = |C_i| = d$. For $d = 40$, $H(40) \approx 4.28$.

Theorem 4 (Maximum Coverage)**.** *Stopped after* k *images, Algorithm 1 returns a subset* S_k *with*

$$f(S_k) \;\geq\; \left(1 - \frac{1}{e}\right) \max_{S \subseteq \mathcal{I},\ |S| \leq k} f(S).$$

This follows from the monotonicity and submodularity of f (Nemhauser et al., 1978).

N.5 Greedy Image Selection Algorithm

Algorithm 1 presents the greedy procedure. Setting $k = \infty$ gives the set cover formulation; a finite k gives the maximum coverage formulation used in our experiments.

Algorithm 1 Greedy Heuristic for Bin Coverage

Require: Set of images $\mathcal{I}$, number of dimensions d, scaling parameter w, budget k ($k = \infty$ for full coverage)
Ensure: S: selected subset of images
1: Initialize $S \leftarrow \emptyset$
2: **for** each dimension $j = 1, 2, \ldots, d$ **do**
3: Compute the number of bins $B_j = \left\lfloor w \cdot \frac{\lambda_j}{\lambda_1} \right\rfloor$
4: **end for**
5: Define $Gap(S, j)$: number of empty bins in dimension j
6: Define $Gap(S) \leftarrow \sum_j Gap(S, j)$
7: **while** $Gap(S) > 0$ **and** $|S| < k$ **do**
8: Select $i \in \mathcal{I} \setminus S$ minimizing $Gap(S \cup \{i\})$
9: Update $S \leftarrow S \cup \{i\}$
10: **end while**

O Training Time and Computational Cost

This appendix reports the training time, model size, inference time, and hyperparameter-search scope of AAMax relative to baseline fine-tuning. All experiments were run on a single NVIDIA H200 GPU with 141 GB of memory. We separate the cost of the underlying architecture, which is shared by both methods, from the incremental cost of AAMax.

O.1 Training Time

We measured the run time of each step of the training process (Table 23). The adapters are first allowed to overfit to the reference embeddings by training them for 300 epochs with the MSE loss, which takes around one minute. End-to-end fine-tuning then has essentially the same per-epoch cost for the baseline and for AAMax: apart from the adapter-level MSE loss, both pipelines run the same architecture on the same data. For a complete 160-epoch run, baseline fine-tuning requires 590.4 minutes (9.84 GPU hours) and AAMax requires 597.4 minutes (9.96 GPU hours). AAMax therefore adds approximately 7.0 minutes, or 1.19%, to the training run: 1.1 minutes for the alignment stage and about 5.9 minutes accumulated from the 2.2-second per-epoch difference. Moreover, AAMax does not require the full schedule: one epoch of end-to-end fine-tuning after alignment already outperforms 160 epochs of baseline training (Appendix H).

Table 23: Training times for baseline and AAMax fine-tuning, averaged over the three fine-tuned subjects (Subjects 2, 5, and 7). The total time is the time to train for 160 epochs of end-to-end fine-tuning.

Training Method	Adapter Alignment (Total Time)	End-to-End Fine-tuning		Total Time to Train
		Time per Epoch	Total Time (160 epochs)	
Baseline	-	221.40 seconds	590.40 minutes	590.40 minutes
AAMax	67 seconds	223.61 seconds	596.29 minutes	597.41 minutes

O.2 Model Size

AAMax changes the optimization procedure and does not add a layer or an inference module. In every paired comparison, the baseline and AAMax use the same architecture and the same subject-specific adapters, so they have identical parameter counts (Table 24). The counts include the subject-specific adapters, the shared voxel-to-CLIP network, and the diffusion prior, and exclude the frozen image generator (Versatile Diffusion/UnCLIP), which is identical for both methods. If AAMax is applied to a backbone that has no subject-specific adapter, the only additional subject-specific cost is a linear projection plus bias, with $D_s \times 4096 + 4096$ parameters for a subject with D_s voxels (58.5M parameters for Subject 2, with 14,278 voxels).

Table 24: Learned parameters of the brain-decoding checkpoints, which are identical for baseline fine-tuning and AAMax. Subject adapters are those for Subjects 1 and 2.

Architecture	Subject adapters	Shared voxel-to-CLIP	Diffusion prior	Total
MindEye1	122,912,768	883,162,112	56,056,564	1,062,131,444
MindEye2	122,896,384	1,903,020,028	259,866,596	2,285,783,008

O.3 Inference Time

Because the baseline and AAMax checkpoints share the same inference graph, their inference costs are identical. Measured on the same H200 GPU, high-level MindEye1 reconstruction takes approximately 4.44 seconds per image (1 hour 13 minutes 58 seconds for 1,000 test images) with batch size 1, eight reconstruction candidates per sample, 100 diffusion-prior timesteps, and 20 UnCLIP inference steps. Most of this time is spent in diffusion-prior sampling and image generation rather than in the subject-specific adapter.

O.4 Hyperparameter-Search Scope

AAMax inherits the backbone fine-tuning hyperparameters of the baseline it improves, including the optimizer, learning-rate schedule, batch construction, number of fine-tuning epochs, and loss weights. We did not perform a separate per-subject search over these hyperparameters for either

method; the baselines use the configurations recommended in the original MindEye1 and MindEye2 papers. The AAMax-specific settings are held fixed across all paired experiments: the alignment stage uses AdamW with a learning rate of 3×10^{-4} for 300 epochs, and end-to-end fine-tuning uses an adapter-level MSE weight of 10. The adapter-architecture comparisons in Appendix D are robustness ablations rather than a hyperparameter search for the reported results.

O.5 Amortized Cost

The reference model and the diffusion stack are part of the underlying architecture and are required equally by baseline fine-tuning and AAMax. Training the reference subject is a one-time cost that is amortized over every subsequently adapted subject, whereas the reductions in scan time and adaptation time achieved by AAMax apply to each new subject.

P High-Level Pipeline Training Process

The high-level pipeline aims to map input fMRI signals to the CLIP embedding space. The training process differs based on whether the subject is the reference subject or a new subject being aligned to the reference space.

P.1 Training the Reference Subject

For the reference subject, the loss function combines Diffusion Prior Loss and CLIP Loss. The overall loss for the reference subject is:

$$\mathcal{L}_{\text{ref}} = \lambda_1 \mathcal{L}_{\text{Prior}} + \lambda_2 \mathcal{L}_{\text{CLIP}}$$

where λ_1 and λ_2 are weighting coefficients for the respective loss components.

P.2 Training a New Subject

For a new subject, the training involves two stages:

Stage 1: Adapter-Level Alignment. Initially, the adapter is trained using the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{adapter}} = \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{adapter, new}}, \mathbf{z}_{\text{adapter, ref}})$$

where $\mathbf{z}_{\text{adapter, new}}$ and $\mathbf{z}_{\text{adapter, ref}}$ are the adapter-level embeddings for the new and reference subjects, respectively, on the shared common images.

Stage 2: End-to-End Training. After aligning the adapter, the entire pipeline is trained end-to-end. During this stage:

- For non-common images, Prior Loss and CLIP Loss are applied to align the new subject's output embeddings to the CLIP space.
- For common images, the MSE Loss is applied at the adapter level to retain the alignment to the reference subject as much as possible.

The overall loss function in this stage is:

$$\mathcal{L}_{\text{new}} = \begin{cases} \lambda_1 \mathcal{L}_{\text{Prior}} + \lambda_2 \mathcal{L}_{\text{CLIP}}, & \text{for non-common images} \\ \lambda_1 \mathcal{L}_{\text{Prior}} + \lambda_2 \mathcal{L}_{\text{CLIP}} + \lambda_3 \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{adapter,new}}, \mathbf{z}_{\text{adapter,ref}}), & \text{for common images} \end{cases}$$

where $\mathbf{z}_{\text{adapter, new}}$ and $\mathbf{z}_{\text{adapter, ref}}$ are the adapter-level embeddings for the new and reference subjects, respectively, on the shared common images.

Q Low-Level Pipeline Training Process

The low-level pipeline aims to map input fMRI signals to the latent space of an Autoencoder (AE), which is trained to compress the NSD train images into a 1024-dimensional latent representation and reconstruct them. The training process differs based on whether the subject is the reference subject or a new subject being aligned to the reference space.

Q.1 Training the Reference Subject

For the reference subject, the goal is to map the fMRI signals to the AE latent space using the Mean Squared Error (MSE) loss. The AE latent representations of the corresponding images serve as the ground truth.

The loss function for the reference subject is:

$$\mathcal{L}_{\text{ref}} = \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{fMRI}}, \mathbf{z}_{\text{AE}})$$

where $\mathbf{z}_{\text{fMRI}}$ is the mapped output embedding from fMRI signals, and $\mathbf{z}_{\text{AE}}$ is the AE latent space representation of the corresponding image.

Q.2 Training a New Subject

For a new subject, the training involves two stages:

Stage 1: Adapter-Level Alignment. In this stage, only the adapter is trained using MSE Loss to align the new subject's adapter-level embeddings with those of the reference subject for shared common images:

$$\mathcal{L}_{\text{adapter}} = \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{adapter, new}}, \mathbf{z}_{\text{adapter, ref}})$$

where $\mathbf{z}_{\text{adapter, new}}$ and $\mathbf{z}_{\text{adapter, ref}}$ are the adapter-level embeddings for the new and reference subjects, respectively.

Stage 2: End-to-End Training. In this stage, the entire pipeline is trained end-to-end. MSE Loss is applied at 2 points in the pipeline:

- At the adapter level, to maintain alignment between the new subject's embeddings and the reference subject's embeddings for common images.
- At the mapper output level, to ensure the new subject's fMRI embeddings align with the AE latent representations for corresponding images.

The loss function for this stage is:

$$\mathcal{L}_{\text{new}} = \begin{cases} \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{fMRI, new}}, \mathbf{z}_{\text{AE}}), & \text{for non-common images} \\ \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{fMRI, new}}, \mathbf{z}_{\text{AE}}) + \mathcal{L}_{\text{MSE}}(\mathbf{z}_{\text{adapter, new}}, \mathbf{z}_{\text{adapter, ref}}), & \text{for common images} \end{cases}$$

where:

- $\mathbf{z}_{\text{fMRI, new}}$: Output embedding for the new subject.
- $\mathbf{z}_{\text{AE}}$: AE latent space representation of the image.
- $\mathbf{z}_{\text{adapter, new}}$ and $\mathbf{z}_{\text{adapter, ref}}$: Adapter-level embeddings for the new and reference subjects, respectively.

R VISUAL COMPARISON RESULTS AMONG DIFFERENT METHODS

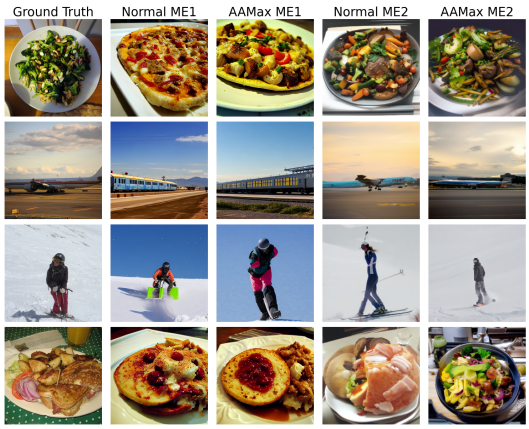


Figure 16: Visual comparison among MindEye 1 without AAMax (Normal ME1), MindEye1 with AAMax (AAMax ME1), MindEye 2 without AAMax (Normal ME2) and MindEye2 with AAMax (AAMax ME2).

S CROSS-DATASET TRANSFER AND CORRESPONDENCE QUALITY

Adapter Alignment requires a set of image pairs on which the new subject's adapter is aligned to the reference. Within NSD, these are the shared images viewed by all subjects. Across datasets, exact correspondences may or may not exist. This appendix describes how correspondences are constructed for BOLD5000 and THINGS-fMRI, and quantifies how the quality of the correspondence affects AAMax.

S.1 TRANSFER BETWEEN NSD AND BOLD5000

We conducted additional transfer experiments using the BOLD5000 Chang et al. (2019) dataset. BOLD5000 includes a subset of COCO images that overlaps directly with the NSD shared images, so these experiments use exact image correspondences between the reference and the fine-tuned subject. Table 25 summarizes the results. These experiments allow us to directly compare how well models trained on NSD vs. BOLD5000 serve as reference subjects for fine-tuning. All fine-tuning runs use AAMax with 500 training samples for the fine-tuned subject on the ME1 architecture; because all rows use AAMax, the comparison isolates the effect of the reference dataset rather than the training procedure. The reference subjects across both datasets were trained on all subject specific available data.

- Row 1 (B5K Subj1 Pre-train): Performance of a model trained from scratch on Subject 1 of BOLD5000.
- Row 2 (FT B5K Subj2 on B5K Subj1): Fine-tuning Subject 2 of BOLD5000 using Subject 1 of BOLD5000 as the reference.
- Row 3 (FT B5K Subj2 on NSD Subj1): Fine-tuning Subject 2 of BOLD5000 using NSD Subject 1 as the reference.
- Row 4 (FT NSD Subj2 on NSD Subj1): Fine-tuning NSD Subject 2 using NSD Subject 1 as the reference (the standard NSD→NSD setting).
- Row 5 (FT NSD Subj2 on B5K Subj1): Fine-tuning NSD Subject 2 using BOLD5000 Subject 1 as the reference.

Across these conditions, we consistently observe that NSD serves as a substantially stronger reference dataset. NSD-trained reference models yield higher-quality fine-tuning outcomes on both NSD and BOLD5000 target subjects. In contrast, BOLD5000-trained reference models lead to noticeably

weaker reconstruction performance even when transferring within the same dataset. These results confirm the intuition that NSD's high-resolution 7T acquisitions give rise to richer, more stable subject-specific representation spaces, which in turn make NSD the most reliable and effective choice for constructing a reference model. This finding aligns with our original rationale for using NSD as the backbone dataset for the common space.

Table 25: Cross-dataset and within-dataset transfer performance between NSD and BOLD5000.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
B5k Subj1 Pre-train	0.041	0.211	65.71%	76.87%	66.24%	74.77%	0.928	0.612
FT B5k Subj2 on B5k Subj1	0.047	0.247	60.53%	66.23%	**60.06%**	64.63%	0.952	**0.631**
FT B5k Subj2 on NSD Subj1	0.053	0.252	**63.32%**	**69.13%**	59.75%	**66.07%**	**0.948**	0.634
FT NSD Subj2 on B5k Subj1	0.115	0.289	79.76%	86.73%	79.57%	79.88%	0.825	0.499
FT NSD Subj2 on NSD Subj1	0.114	0.270	**83.70%**	**91.76%**	**83.71%**	**82.00%**	**0.789**	**0.469**

S.2 Constructing Approximate Correspondences for THINGS-fMRI

No images are shared between THINGS-fMRI and NSD. For the experiment in Section 5, we therefore construct an approximate common set. We compute the CLIP cosine similarity between each THINGS training image and each of the 1,000 NSD shared images used to define the reference embeddings, and retain the pairs with similarity above 0.90, which yields approximately 150 semantically matched pairs. During adapter alignment, the THINGS subject's response to each retained image is aligned to the NSD reference subject's embedding of its matched NSD image. We chose the 0.90 threshold because Figure 3 indicates that roughly 150 images suffice for strong performance, and this threshold yields that budget while keeping the matched pairs highly similar. THINGS images are mostly object-centric and visually simple, so pairs above this threshold typically depict the same object category in similar views, and high CLIP similarity behaves much like category-level matching.

S.3 Effect of Correspondence Quality

To quantify the cost of replacing exact correspondences with semantic ones, we repeated the NSD Subject 1 → Subject 2 experiment (1 hour of data, MindEye1) using 250 CLIP-matched image pairs instead of 250 identical shared images. Because NSD images are complex natural scenes, close semantic matches are much rarer than for the object-centric THINGS images: whereas every THINGS pair exceeds the 0.90 threshold by construction, over 90% of the matched NSD pairs fall below it (median CLIP similarity of approximately 0.79–0.86 across the matching directions used). Table 26 shows that AAMax with approximate pairs still improves over the baseline on every metric, but the gain is much smaller than with exact pairs, which are substantially better on all metrics except SSIM.

Table 26: Effect of correspondence quality on AAMax. Subject 1 is the reference and Subject 2 is fine-tuned with 1 hour of data. Approximate pairs replace the identical shared images with CLIP-matched images. All rows are averaged over three independently generated splits. Low-level metrics are shown in gray because they are not optimized in these runs.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Baseline	0.078	0.264	75.33%	84.18%	74.15%	73.66%	0.862	0.524
AAMax, approximate pairs	0.086	0.275	78.11%	85.74%	74.79%	75.29%	0.843	0.518
AAMax, exact pairs	0.112	0.259	**78.75%**	**88.04%**	**78.99%**	**78.28%**	**0.824**	**0.513**

These results support a more precise conclusion than exact matching alone would: shared images are not strictly required, but the quality of the correspondence strongly affects the quality of the alignment, especially for complex natural scenes, where CLIP-level semantic similarity does not

guarantee proximity in the learned fMRI representation. We do not include random matching as a baseline, because it would explicitly force unrelated stimuli toward the same reference embedding and thus inject contradictory supervision rather than provide a neutral control. Category-level matching is represented by MindAligner (Section 4.3, Appendix T), which aligns subjects in voxel space using category-level pairs; combining it with AAMax gives the strongest results, suggesting that coarse category-level alignment and fine-grained common-space alignment are complementary. When only weak matches are available, a hybrid strategy is therefore likely to be most effective: coarse semantic or voxel-space alignment, followed by Adapter Alignment on the highest-confidence subset of pairs.

T Combining Adapter Alignment and Image Selection with MindAligner

We conducted a direct comparison between MindAligner and our Adapter Alignment (AA) method using the MindEye1 architecture under the same experimental protocol used throughout the paper. As before, we pretrain the full model on Subject 1 and fine-tune on Subject 2 using 1 hour of data. We evaluate three configurations:

1. **MindAligner fine-tuning:** We applied MindAligner exactly as described in (Dai et al., 2025), including the generation of categorical stimulus pairs, even though the training data contain exact shared stimuli. This procedure yields significantly more training examples than the 250-image subset used in our low-data experiments, and the model is trained using the original MindAligner voxel-space loss functions.
2. **Adapter Alignment + Image Selection:** Using our proposed image-selection algorithm, we generate a 250-image subset that spans the breadth of the principal eigenspace of Subject 1's shared representation. This ensures coverage across all principal axes of variation, enabling AA to perform exact one-to-one alignment in the shared space. This configuration mirrors the setup used in all other AA experiments.
3. **MindAligner + Adapter Alignment + Image Selection:** MindAligner and AA operate at different levels of the pipeline. MindAligner performs coarse categorical voxel-space alignment, and AA performs fine-grained common-space alignment. This lets us build and evaluate a multi-stage pipeline. Training a new subject is decomposed into four steps:
   (a) fine-tune the BFA (Brain Feature Adapter) to reconstruct category-level fMRI for Subject 1;
   (b) apply Adapter Alignment by overfitting the new subject's adapter to Subject 1's shared-space embeddings on the new 1-hour dataset;
   (c) perform end-to-end training of the BFA module with the remainder of the architecture frozen;
   (d) perform AA end-to-end fine-tuning with the BFA frozen.

   This sequence allows MindAligner to establish coarse alignment at the voxel level, while AA and Image Selection refine the geometry of the shared space.

U Summary statistics

In this section, we present summary statistics of the results reported in Table 1.

Table 27: Summary statistics of the quantitative results on training the reconstruction models on all 40 hours of data. The fine-tuning results are averaged over subjects 2,5 and 7. AAMax and baseline fine-tuning perform similarly across all metrics.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Subject 1 Pre-train	0.345	0.346	92.80%	96.88%	94.40%	90.02%	0.692	0.399
Baseline FT AVG	0.258±0.03	**0.339±0.005**	**89.64%±1.8%**	95.05%±0.87%	92.63%±1.62%	**89.64%±1.55%**	0.717±0.017	0.427±0.01
AAMax FT AVG	**0.259±0.03**	0.337±0.00	89.20%±2.27%	**95.16%±1.03%**	**92.68%±1.69%**	89.49%±1.25%	**0.713±0.02**	**0.423±0.02**

V SCALING WITH TRAINING DATA

We further examine how the benefits of Adapter Alignment scale with training data size. Using Subject 2 as the fine-tuning subject, we progressively increase the training set from 250 to 500, 1000, and 4000 images, before extending to the full dataset. Up to 1000 images, only common images are used for alignment and fine-tuning; for 4000 images, alignment is still performed on the 1000 shared images, while the additional 3000 images are used only for fine-tuning. As shown in Figure 17, AAMax achieves substantially higher performance than baseline fine-tuning in low-data regimes, with the gap narrowing as training data increases. At full data, both approaches converge, consistent with prior multi-subject pipelines such as MindEye2, where large amounts of subject-specific data allow the model to generalize even without explicit alignment.

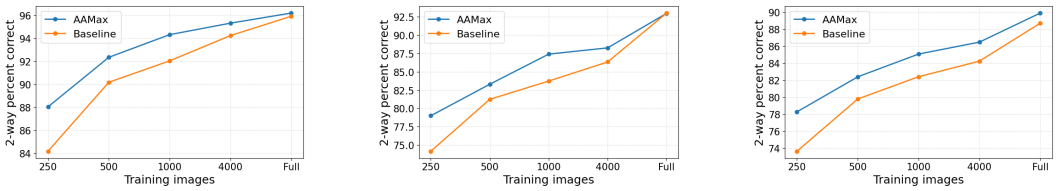


Figure 17: Performance of AAMax compared to baseline fine-tuning as a function of training data size. Left: AlexNet-5 2-way percent correct. Middle: Inception 2-way percent correct. Right: CLIP 2-way percent correct. Across all three metrics, AAMax achieves higher accuracy in limited-data regimes, with the advantage diminishing as more data becomes available.

W GENERALIZING TO DIFFERENT ARCHITECTURES

Adapter Alignment is inherently architecture-agnostic: any reconstruction model can be extended with subject-specific adapters, enabling alignment to a common space. For efficiency and comparability with prior work, all of our main experiments are conducted using a slightly modified MindEye1 pipeline. To demonstrate the architecture-agnostic nature of Adapter Alignment, we additionally evaluate a limited-data setting using 1 hour of training data to fine-tune Subject 2. We compare baseline fine-tuning and AAMax under both the MindEye1 and MindEye2 architectures, and observe consistent improvements with AAMax across both models. Additional experiments using MindEye2 are presented in Appendix C.

Table 28: Evaluating Adapter Alignment across architectures. Subject 1 is trained as the reference, and Subject 2 is fine-tuned with 1 hour of training data. Results are shown for MindEye1 and MindEye2 under baseline and AAMax fine-tuning. Adapter Alignment consistently improves performance and scales with stronger architectures.

	Low-Level				High-Level			
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
MindEye1								
Baseline	0.078	0.264	75.33%	84.18%	74.15%	73.66%	0.862	0.524
AAMax	0.112	0.259	78.75%	88.04%	78.99%	78.28%	0.824	0.513
MindEye2								
Baseline	0.104	0.354	81.71%	88.06%	78.84%	74.15%	0.845	0.499
AAMax	0.140	0.355	**82.22%**	**89.67%**	**82.97%**	**78.79%**	**0.821**	**0.483**

X LIMITATIONS

Few-shot adaptation, not zero-shot generalization. Adapter Alignment reduces the amount of subject-specific data needed to model a new participant; it does not remove the need for such data. Every new subject requires a lightweight subject-specific adapter and a small calibration set viewed by, or semantically matched to, the reference subject. Our evaluations therefore concern low-data

adaptation to subjects that were not used to pre-train the reference model, rather than zero-shot transfer to unseen subjects. The comparable reconstruction methods we build on or compare against likewise require subject-specific data for every new subject.

Calibration correspondences. Explicit alignment requires image pairs between the new subject and the reference. With exact shared images the method is most effective. When only semantically matched images are available, AAMax still improves over baseline fine-tuning, but its advantage shrinks as the quality of the correspondence decreases, particularly for complex natural scenes (Appendix S). The CLIP-based matching used for THINGS-fMRI may impose part of the shared structure through the CLIP representation rather than recovering it from the brain data alone.

Dependence on the reference subject. AAMax outperforms baseline fine-tuning for every reference subject we evaluated with Subject 1 as the target, but absolute performance depends on the quality and compatibility of the reference (Appendix E), and a reference model trained on a lower-SNR dataset yields weaker fine-tuning (Appendix S). Pre-training the reference on multiple subjects did not improve over a single strong reference. In practice, this favors building the reference from a high-quality dataset such as NSD; when no strong reference is available for a given scanner or stimulus set, the gains of AAMax are expected to be smaller.

Evaluation protocol. Our main NSD experiments use a modified split in which the shared images are moved into training. All methods are retrained and evaluated on identical splits, and a matched rerun shows that the modified test set is harder than the standard one (Appendix B), but the protocol still limits direct comparison with results reported under the standard NSD benchmark. The baselines use the hyperparameters recommended in their original papers and were not independently retuned for the modified split; this isolates the effect of the training procedure but does not establish that each baseline is optimally tuned for this protocol. Several supporting analyses, namely the image-selection baselines (Appendix F), pre-diffusion retrieval (Appendix L), and the neural correspondence analysis (Appendix K), use a single training run and a single target subject.

Neural interpretation. Like most reconstruction pipelines, AA treats each subject's fMRI pattern as a vector, and alignment is learned after the subject-specific adapter has absorbed differences in voxel count, anatomy, and cortical organization. Our pRF- and fLOC-resolved analyses show improved functional alignment of signals from homologous visual regions, but they do not establish anatomical or vertex-level correspondence, and AA does not use cortical mesh geometry during training.

System-level cost. AA reduces the scan time and computation needed to adapt each new subject. It does not reduce the one-time cost of training the reference model or of the frozen generative stack, which are determined by the underlying architecture (Appendix O).

Generative prior and signal quality. The ultimate image fidelity hinges on the strengths and biases of the pre-trained diffusion model, so any generative shortcomings propagate into our outputs. In addition, variability and noise in the fMRI signals (e.g., scanner differences, low SNR) can further degrade adapter alignment and reconstruction accuracy.

Ethical considerations. Because fMRI-to-image reconstruction systems process sensitive brain data, their development raises important privacy, consent, and misuse concerns. These methods should not be interpreted as general-purpose "mind reading," and any future deployment, especially in patient or clinical populations, must include strict safeguards for informed consent, data security, transparency about reconstruction uncertainty, and protection against coercive or unauthorized use. This work should therefore be viewed as a step toward more efficient scientific and assistive neural decoding, rather than as a technology ready for unconstrained real-world application.

Y RECONSTRUCTION RESULTS



Figure 18: Reconstruction results after fine-tuning on 250 common images.