# MINDBANK: DATA-EFFICIENT CROSS-SUBJECT FMRI DECODING BY LEVERAGING SHARED SEMANTIC STRUCTURE ACROSS SOURCE SUBJECTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Cross-subject functional magnetic resonance imaging (fMRI) visual decoding aims to adapt decoders pretrained on source subjects to a new target subject using only a limited amount of paired fMRI–image data. However, limited data and large differences in fMRI responses across individuals can cause the target representation to lose semantic information during adaptation. We recognize that despite individual variability in fMRI responses, the underlying structure of visual semantics is shared across individuals. Consistent with this view, we observe that the decoded target representation is more similar to the average decoded representation across source subjects. We propose MindBank, which addresses the challenges by leveraging this similarity and the variability through two complementary mechanisms. (1) MindBank aggregates representations from frozen source decoders into a shared semantic reference and adaptively integrates the reference with target features. This mechanism keeps target adaptation consistent with the semantic structure shared across subjects. (2) MindBank exploits variations across the source representations to generate controlled perturbations of target representations. The perturbations expand the coverage of the target feature space and alleviate the limitations imposed by scarce target data. Experiments on the Natural Scenes Dataset (NSD) under a unified one-hour adaptation protocol demonstrate that MindBank uses the fewest trainable parameters, yet outperforms state-of-the-art methods in both retrieval and reconstruction. Code is included in the supplement and will be released.

## 1 INTRODUCTION

Decoding perceived visual content from human brain activity is a long-standing goal of cognitive neuroscience and brain–computer interface research. Functional magnetic resonance imaging (fMRI) non-invasively measures brain responses during visual processing, enabling the inference of content from neural activity (Cox & Savoy, 2003). Large-scale fMRI datasets (Allen et al., 2022), multimodal representation learning (Radford et al., 2021) and generative modeling (Rombach et al., 2022; Xu et al., 2023) have accelerated progress in this field. Decoders map fMRI signals into pretrained vision or vision–language representation spaces and use generative models as image priors, enabling the recovery of high-level semantics and high-fidelity image reconstructions (Ozcelik & VanRullen, 2023; Zeng et al., 2024).

Nevertheless, most advances have focused on the one-subject-one-model paradigm, training a dedicated decoder for each subject (Takagi & Nishimoto, 2023; Scotti et al., 2023). Since brain responses vary substantially across individuals, models trained on one subject often fail to generalize to others, constraining their broader use in real-world scenarios (Wang et al., 2024a). Meanwhile, collecting sufficient data for new subjects is costly and time-consuming (Scotti et al., 2024). These constraints motivate cross-subject models that adapt effectively with limited target data. Recent cross-subject methods typically adapt decoders pretrained on source subjects to new target subjects, using decoding objectives to implicitly align target representations with the source representation space (Scotti et al., 2024; Gong et al., 2025). However, existing methods still often produce reconstructions whose semantic content does not match that of the corresponding ground-truth images.



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

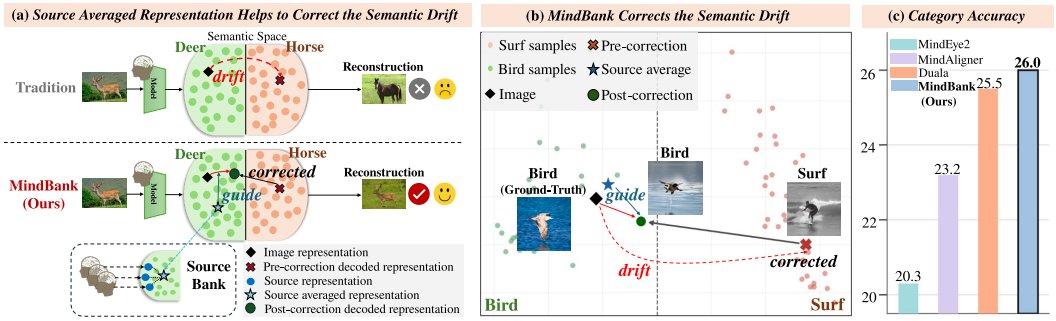



Figure 1: (a) MindBank uses shared semantic structure to mitigate semantic drift. (b) Visual instantiation of the semantic drift phenomenon and MindBank's correction. (c) Category accuracy of reconstruction results on the NSD.

Accordingly, we examine existing models through the lens of category-level semantics and identify a phenomenon that we term *semantic drift*, where decoded features shift toward incorrect categories during decoding, as illustrated in Figure 1(a) and (b). We attribute semantic drift to inter-subject variability in neural response patterns and the scarcity of data, which together prevent the model from accurately learning the target subject's response patterns. Consequently, the model may misinterpret the target subject's responses and fail to preserve category-discriminative semantic information, leading to semantically mismatched retrievals and reconstructions. We recognize that the underlying structure of visual semantics is shared across individuals, which is further supported by the empirical results in Section 3.2. The results show that although the target representation is not aligned with the representation of any source subject, after averaging the source representations, the target representation is well-aligned with the average representation, with an increase in similarity from 0.57 to 0.73. The averaging process partially offsets the inter-subject variations, yielding a cross-subject shared structure.

To address the two challenges, we propose **MindBank**, a data-efficient method for cross-subject fMRI decoding. MindBank constructs a Source Bank by extracting source representations, which are integrated into the target representation to supervise the fine-tuning, thereby mitigating semantic information loss during cross-subject adaptation. These source representations capture not only shared semantic structure but also natural inter-subject variation. Source Disagreement Augmentation (SDA) converts the variation into controlled perturbations that augment the target feature distribution, thereby mitigating the effects of limited target data and improving decoding robustness. We also introduce Category-aware Semantic Alignment (CSA) to better mitigate the semantic drift. CSA assigns the category label to each target feature and its SDA-perturbed counterpart, and applies a semantic alignment loss that pulls representations with the same label closer while pushing those with different labels apart, preserving clear semantic boundaries between categories. As shown in Figure 1(b) and (c), our methods help to maintain the semantic information during adaptation and improve the category-level accuracy of the reconstructed images. In summary, our main contributions are as follows:

- We analyze the similarity of decoded representations across subjects. The results indicate that the average source representation can serve as an effective supervisory signal for target-subject adaptation.
- We propose **MindBank**, a framework that models semantic interactions between source and target branches, leveraging shared source semantics to improve the fine-tuning and mitigate the semantic drift.
- We evaluate our method on the NSD dataset. MindBank achieves state-of-the-art performance under one-hour adaptation, with 81.64% brain retrieval accuracy and an improvement of 5.7 percentage points in category accuracy over the baseline.

# 2 Related Work

## 2.1 fMRI-Based Visual Decoding

fMRI-based visual decoding seeks to infer perceived visual content from brain responses (Lin et al., 2022; Xia et al., 2024b). Early methods mapped fMRI responses into visual feature spaces (Kay et al., 2008; Naselaris et al., 2011; Horikawa & Kamitani, 2017) and later exploited the hierarchical correspondence between visual cortical activity and deep neural network (DNN) representations (Yamins & DiCarlo, 2016; DiCarlo et al., 2012). Recent work has shifted toward generative reconstruction (Seeliger et al., 2018; Lin et al., 2019; Chen et al., 2023; Beliy et al., 2019), mapping brain responses into the latent spaces of pretrained DNNs, vision–language models, and diffusion models (Shen et al., 2019; Takagi & Nishimoto, 2023; Scotti et al., 2023). Their semantic and structural priors enable more accurate and coherent reconstructions (Quan et al., 2024; Shen et al., 2024; Wu et al., 2025). However, most methods remain subject-specific and require separate decoders trained with extensive paired fMRI–image data, limiting decoders' generalizability across subjects (Xia et al., 2024a; Mai & Zhang, 2023). In contrast, MindBank is a data-efficient cross-subject fMRI decoding method that achieves strong performance with only 1 hour of fMRI–image data.

## 2.2 Cross-Subject Brain Decoding

Inter-subject variability in brain anatomy and function limits visual decoder generalization. Early shared-stimulus mappings scale poorly because only a small subset of Natural Scenes Dataset (NSD) images is shared across subjects (Allen et al., 2022; Bazeille et al., 2019). Recent methods project subject-specific brain responses into shared latent spaces (Scotti et al., 2024; Thual et al., 2023; Ferrante et al., 2024; Bao et al., 2025). MindBridge synthesizes pseudo-shared stimuli, whereas MindTuner models nonlinear inter-subject relationships; both reduce source–target discrepancies (Wang et al., 2024a; Gong et al., 2025). Overall, existing methods focus on minimizing inter-subject representation discrepancies. In contrast, MindBank uses agreement between the target and average source representations as semantic guidance while preserving subject-specific variation.

# 3 Methods

## 3.1 Problem Setting and Overview

In cross-subject visual decoding, the goal is to transfer a decoder learned from existing subjects to a new subject with limited fMRI data (Liu et al., 2025; Wang et al., 2024b; Mansour et al., 2008). For a subject $s$, let $\mathcal{D}_s = \{(I_s, V_s)\}$ denote paired fMRI–image data, where $I_s \in \mathbb{R}^{3\times H\times W}$ is the visual stimulus and $V_s \in \mathbb{R}^{d_s}$ is the voxel response with $d_s$ voxels. A subject-specific ridge regression layer $R_s : \mathbb{R}^{d_s} \to \mathbb{R}^U$ maps the voxel responses into a shared latent space, yielding $x_s = R_s(V_s) \in \mathbb{R}^U$. $x_s$ is processed by a shared multilayer perceptron (MLP) backbone $G_{\text{src}}$ to produce the hidden latents $z_s = G_{\text{src}}(x_s)$ and then decoded by a shared linear layer $J$, yielding $Z_s = J(z_s)$. Let $K$ be the number of source subjects. The decoding model is first pretrained on $K$ source subjects and then fine-tuned on a target subject $t$. After decoding, a diffusion prior predicts CLIP image embeddings from brain features, enabling end-to-end reconstruction. The baseline training objective follows MindEye2 and combines the diffusion-prior loss $\mathcal{L}_{\text{prior}}$, the low-level blurry reconstruction loss $\mathcal{L}_{\text{lowlevel}}$ and a bidirectional contrastive loss $\mathcal{L}_{\text{BiMixCo}}$.

Building upon this foundation, we focus on the new-subject fine-tuning stage. We conduct an analysis to examine the relationships among the target representation, source representations, and their average. Based on the analysis, we propose **MindBank**, a framework that leverages the variability among source subjects and the similarity between source and target representations in adaptation. Figure 3 illustrates the overall pipeline. MindBank constructs a Source Bank from source-path responses to each target input. SDA uses cross-source variation to perturb target features, while CSA encourages intra-category compactness and inter-category separation in the target feature space.

## 3.2 Cross-Subject Representation Analysis

To examine cross-subject representational relationships, we designate Subject 2 as the target, use $\mathcal{S} = \{1, 3, 4, 5, 6, 7, 8\}$ as source subjects, and sample 1,000 paired fMRI–image examples. For

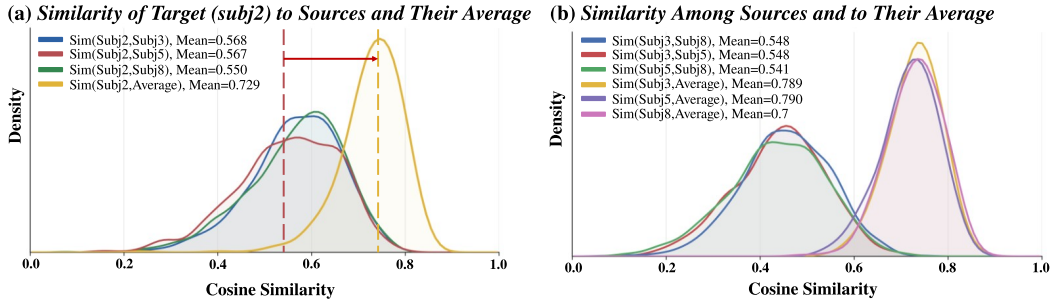



Figure 2: (a) Similarity of the target representation to representations of Subjects 3, 5, and 8 and to the average of all source representations. (b) Similarity of the average to representations of Subjects 3, 5, and 8 and pairwise similarity among representations of Subjects 3, 5, and 8.

each example, the target decoder produces $Z_t$, while seven source decoders produce $\{Z_s\}_{s\in\mathcal{S}}$ after the input is padded or truncated to match the ridge layer. Figure 2(a) reports sample-wise similarities of $Z_t$ to $Z_s$ from Subjects 3, 5, and 8 and to the average of source representations, while Figure 2(b) reports similarities of the average to these three $Z_s$ and their pairwise similarities. The same pattern holds for other source subjects and for Subjects 1, 5, and 7 when each serves as the target.

Figure 2(a) demonstrates that averaging source representations substantially increases the similarity to the target representation. Before averaging, the target representation is poorly aligned with individual source representations; for example, the mean similarity between Subjects 2 and 3 is $0.568$, which limits the direct fusion of a single source representation with the target representation. After averaging, the similarity increases to $0.729$, yielding better target–source alignment. The averaging process partially offsets subject-specific discrepancies among the source representations and produces a source consensus shared across source subjects. Figure 2(b) also supports this interpretation, as the source representations exhibit higher similarities to their average than to individual source representations. We therefore interpolate this average with the target representation to inject shared source semantics and mitigate semantic drift, corresponding to our Source Bank method below. Figure 2(b) also shows relatively low pairwise similarity among source representations, indicating diverse subject-specific variations. SDA converts this diversity into controlled target perturbations that expand target feature coverage under limited data.

### 3.3 Our Proposed MindBank

MindBank uses reliability-calibrated Source Bank supervision to guide target adaptation, Source Disagreement Augmentation (SDA) to augment target features with cross-source variation, and Category-aware Semantic Alignment (CSA) to preserve category structure under limited target data.

**Source Bank.** As discussed in Section 3.2, to mitigate semantic drift during target adaptation, we first obtain source references. For each source subject $s \in \{1, \ldots, K\}$, the target fMRI input is padded or truncated to match the input dimension of the frozen source ridge $R_s$, then projected by $R_s$ to obtain $x_s \in \mathbb{R}^U$. Padding/truncation ensures dimensional compatibility but does not establish voxel-wise anatomical or functional correspondence across subjects. The target ridge $R_t$ maps the original target input to $x_t \in \mathbb{R}^U$. Let $G_{\text{src}}$ denote the frozen source backbone, $G_t$ the target backbone and $J$ the shared linear layer. The source-conditioned representations are

$$SS^s = J(G_{\text{src}}(x_s)), \qquad ST^s = J(G_t(x_s)), \tag{1}$$

while $x_t$ is processed through both backbones and $J$:

$$TT = J(G_t(x_t)), \qquad TS = J(G_{\text{src}}(x_t)). \tag{2}$$

We aggregate the source-conditioned representations as

$$SS = \frac{1}{K}\sum_{s=1}^{K} SS^s, \qquad ST = \frac{1}{K}\sum_{s=1}^{K} ST^s. \tag{3}$$

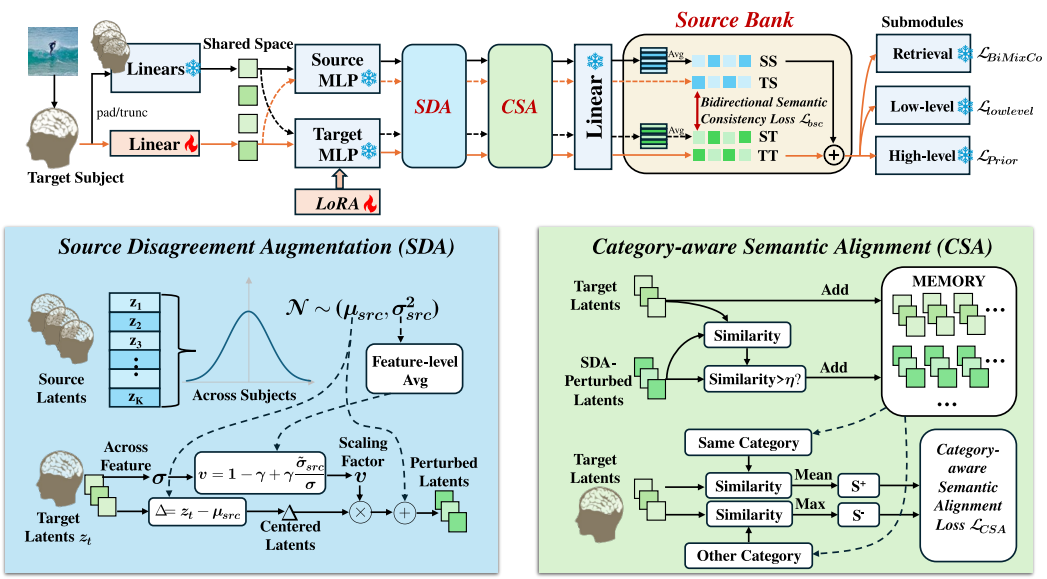



Figure 3: Overview of MindBank. Source Bank aggregates source-path responses to guide target adaptation. Source Disagreement Augmentation (SDA) converts cross-source dispersion into controlled perturbations. Category-aware Semantic Alignment (CSA) stores features in a memory bank and promotes intra-category compactness and inter-category separation.

Source Bank comprises $SS$, $ST$, $TS$, and $TT$, where $SS$ represents the source consensus and $TT$ retains the target representation. We obtain $\widetilde{TT}$ by linearly interpolating $TT$ and $SS$ with weights $1-\alpha$ and $\alpha$, respectively, where $\alpha$ controls source contribution. This interpolation injects shared structure to mitigate semantic drift. We further apply a bidirectional semantic consistency (BSC) loss. The main constraint $\ell_{\text{main}}$ is the cosine distance between $TT$ and $SS$, where cosine distance $\ell_{\cos}$ denotes one minus cosine similarity. The auxiliary constraint $\ell_{\text{aux}}$ averages the cosine distances for the pairs $(TS, TT)$ and $(ST, SS)$. $\ell_{\text{main}}$ anchors target representations to the source average, while $\ell_{\text{aux}}$ compares representations obtained through the two backbones, reducing cross-backbone discrepancies for the same inputs. The per-sample BSC loss is

$$\ell_{\text{bsc}} = \lambda_1 \ell_{\text{main}} + \lambda_2 \mathbf{1}_{\{e<E/2\}} \ell_{\text{aux}}. \tag{4}$$

Here, $\lambda_1$ and $\lambda_2$ weight the main and auxiliary constraints, respectively, while $e$ and $E$ denote the current and total epochs. Applying the auxiliary constraint only during the first half of training thereby preserves target-specific knowledge after early cross-backbone alignment.

Furthermore, Reliability Calibration (RC) adjusts constraint strength by calibrating Source Bank supervision at the sample and channel levels, preventing potentially misleading supervision. At the sample level, we measure the agreement score $u$ by computing similarity of $SS^s$ to $SS$, then use $u$ to compute the normalized sample-weighted loss:

$$\mathcal{L}_{\text{bsc}} = \frac{\sum_{b\in\mathcal{B}} u_b \ell_{\text{bsc},b}}{\max\left(\sum_{b\in\mathcal{B}} u_b, \epsilon\right)}. \tag{5}$$

Here, $b$ indexes samples in the mini-batch $\mathcal{B}$, and $\epsilon > 0$ is a small constant for numerical stability.

At the channel level, $C$ is the number of feature channels. RC combines source agreement $u_c$ with target–source alignment $q_c$ to obtain channel reliability $\rho_c$. The normalized reliability determines the interpolation coefficient:

$$\alpha_c = \alpha_{\text{base}} \frac{\rho_c}{C^{-1}\sum_{j=1}^{C} \rho_j}, \tag{6}$$

where $\alpha_{\text{base}}$ controls the overall source contribution. The calibrated target representation is

$$\widetilde{TT}_c = (1-\alpha_c) TT_c + \alpha_c SS_c. \tag{7}$$

This design incorporates more source semantics into channels with high source agreement and strong target–source alignment while preserving target-specific information in less reliable channels.

**Source Disagreement Augmentation.** Limited target data cannot adequately cover cross-subject variation, while source representations in the bank naturally provide diverse subject-specific variations. We therefore introduce Source Disagreement Augmentation (SDA) to use the variation among these representations to expand the target feature distribution.

For each sample, let $z_t \in \mathbb{R}^{1\times D}$ and $\{z_s\}_{s=1}^{K}$ denote the target and source latents, respectively. Their center and disagreement are

$$\mu_{\mathrm{src}} = \frac{1}{K}\sum_{s=1}^{K} z_s, \qquad \sigma_{\mathrm{src}} = \sqrt{\mathrm{Var}_s(z_s) + \epsilon}. \tag{8}$$

$\mathrm{Var}_s$ denotes the variance across source subjects. The feature-wise disagreement is aggregated as

$$\widetilde{\sigma}_{\mathrm{src}} = \sqrt{\frac{1}{D}\sum_{d=1}^{D} \sigma_{\mathrm{src},d}^2 + \epsilon}. \tag{9}$$

We then calculate the centered target latent and its scale:

$$\Delta = z_t - \mu_{\mathrm{src}}, \qquad \sigma = \sqrt{\mathrm{Var}_d(\Delta_d) + \epsilon}. \tag{10}$$

SDA rescales the latent by source disagreement:

$$v = (1-\gamma) + \gamma\frac{\widetilde{\sigma}_{\mathrm{src}}}{\sigma}, \qquad \widehat{z}_t = \mu_{\mathrm{src}} + v\Delta. \tag{11}$$

With augmentation strength $\gamma = 0.5$, SDA is applied with probability $p = 0.5$.

By preserving the source consensus and the direction of the target residual, SDA expands the target feature distribution along plausible cross-subject variations, improving generalization.

**Category-aware Semantic Alignment.** To better reduce semantic drift, we introduce Category-aware Semantic Alignment (CSA), which promotes intra-category compactness and inter-category separation in the target feature space. Semantically consistent SDA perturbations provide additional intra-category samples, enabling CSA to preserve category structure under limited target data.

Specifically, a frozen CLIP model assigns each image a category label $y$. CSA constructs a memory bank $\mathcal{M}$ by storing the target intermediate features and their semantically consistent SDA-perturbed features with labels. For a normalized target feature $z$, entries in $\mathcal{M}$ with label $y$ form the positive set $\mathcal{M}^+$, while the remaining entries form the negative set $\mathcal{M}^-$. CSA compares each anchor with its positive features and its closest negative feature. The positive similarity and hardest negative similarity are computed as

$$s^+ = \frac{1}{|\mathcal{M}^+|}\sum_{b\in\mathcal{M}^+} z^\top b, \qquad s^- = \max_{b\in\mathcal{M}^-} z^\top b. \tag{12}$$

CSA applies the margin-ranking loss

$$\ell_{\mathrm{CSA}} = \max\left(0, m + s^- - s^+\right), \tag{13}$$

where $m$ is the separation margin. The overall $\mathcal{L}_{\mathrm{CSA}}$ averages this loss over valid anchors.

The SDA-perturbed feature $\widehat{z}$ is stored in the memory bank only when $\cos(z, \widehat{z}) > \eta$, where $\eta = 0.9$. This process brings same-category features closer and separates features from different categories.

### 3.4 Training Objective

MindBank combines the baseline task objective with $\mathcal{L}_{\mathrm{bsc}}$ and $\mathcal{L}_{\mathrm{CSA}}$:

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \mathcal{L}_{\mathrm{bsc}} + \lambda_{\mathrm{CSA}}\mathcal{L}_{\mathrm{CSA}}. \tag{14}$$

The baseline objective $\mathcal{L}_{\mathrm{base}}$ contains the retrieval, high-level diffusion-prior, and low-level reconstruction losses. $\mathcal{L}_{\mathrm{bsc}}$ contains the weights $\lambda_1$ and $\lambda_2$, while $\lambda_{\mathrm{CSA}}$ weights CSA.

Table 1: Average and subject-wise results on the NSD for Subjects 1, 2, 5, and 7 under the one-hour target-adaptation protocol. Arrows indicate the preferred direction. The best result in each subject block is shown in bold, and the second-best average result for each metric is underlined.

| Method | Venue | Subject | Low-Level | | | | High-Level | | | | Retrieval | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | PixCorr↑ | SSIM↑ | Alex(2)↑ | Alex(5)↑ | Incep↑ | CLIP↑ | Eff.↓ | SwAV↓ | Image↑ | Brain↑ |
| MindEye2 | ICML'24 | Avg. | 0.195 | 0.419 | 84.21% | 90.55% | 80.66% | 80.01% | 0.810 | 0.468 | 78.97% | 57.39% |
| MindAligner | ICML'25 | Avg. | 0.206 | 0.414 | 85.60% | 91.63% | 82.13% | 81.08% | 0.802 | 0.463 | 78.94% | 75.26% |
| MindTuner | AAAI'25 | Avg. | 0.224 | **0.420** | 87.75% | 93.63% | 84.75% | 83.45% | **0.780** | **0.440** | 83.00% | 75.95% |
| Duala | CVPR'26 | Avg. | **0.230** | 0.416 | 87.92% | 93.47% | 85.42% | 83.51% | 0.781 | 0.445 | **84.49%** | 81.11% |
| MindBank (Ours) | – | Avg. | 0.228 | 0.412 | **88.33%** | **94.12%** | **85.62%** | **83.64%** | 0.784 | 0.445 | 83.83% | **81.64%** |
| MindEye2 | ICML'24 | Subj. 1 | 0.235 | **0.428** | 88.02% | 93.33% | 83.56% | 81.76% | 0.798 | 0.459 | 93.96% | 77.63% |
| MindAligner | ICML'25 | Subj. 1 | 0.226 | 0.415 | 88.19% | 93.26% | 83.48% | 81.76% | 0.800 | 0.459 | 90.90% | 86.88% |
| MindTuner | AAAI'25 | Subj. 1 | **0.262** | 0.422 | 90.60% | 94.90% | 85.80% | 84.60% | 0.774 | 0.433 | 94.20% | 87.40% |
| Duala | CVPR'26 | Subj. 1 | 0.253 | 0.418 | 90.81% | 95.29% | 86.62% | 85.11% | **0.768** | 0.433 | **94.77%** | 91.22% |
| MindBank (Ours) | – | Subj. 1 | 0.259 | 0.417 | **91.04%** | **95.40%** | **87.22%** | **85.87%** | 0.774 | **0.432** | 94.23% | **93.72%** |
| MindEye2 | ICML'24 | Subj. 2 | 0.200 | **0.433** | 85.00% | 92.13% | 81.86% | 79.89% | 0.807 | 0.454 | 90.53% | 67.18% |
| MindAligner | ICML'25 | Subj. 2 | 0.218 | 0.426 | 88.08% | 93.33% | 82.47% | 81.62% | 0.791 | 0.452 | 90.04% | 85.61% |
| MindTuner | AAAI'25 | Subj. 2 | 0.225 | 0.425 | 89.10% | 95.10% | 84.80% | 83.70% | 0.781 | **0.440** | 93.70% | 82.80% |
| Duala | CVPR'26 | Subj. 2 | **0.236** | 0.428 | 89.21% | 94.35% | 85.35% | **83.74%** | **0.778** | 0.443 | **94.64%** | 91.47% |
| MindBank (Ours) | – | Subj. 2 | 0.233 | 0.425 | **89.98%** | **95.14%** | **85.59%** | 83.63% | 0.785 | 0.445 | 92.47% | **92.37%** |
| MindEye2 | ICML'24 | Subj. 5 | 0.175 | 0.405 | 83.11% | 90.83% | 82.32% | 78.53% | 0.781 | 0.444 | 66.94% | 46.96% |
| MindAligner | ICML'25 | Subj. 5 | 0.197 | 0.409 | 84.69% | 91.61% | 82.63% | 80.12% | 0.784 | 0.454 | 70.62% | 65.95% |
| MindTuner | AAAI'25 | Subj. 5 | 0.208 | **0.415** | **86.80%** | **93.70%** | **87.70%** | **85.90%** | **0.750** | **0.422** | 72.20% | 68.10% |
| Duala | CVPR'26 | Subj. 5 | **0.214** | 0.413 | 86.50% | 93.07% | 86.36% | 84.46% | 0.772 | 0.439 | 76.59% | **71.20%** |
| MindBank (Ours) | – | Subj. 5 | 0.203 | 0.409 | 86.01% | 93.53% | 87.11% | 84.60% | 0.770 | 0.438 | **76.76%** | 71.19% |
| MindEye2 | ICML'24 | Subj. 7 | 0.170 | 0.408 | 80.70% | 85.90% | 74.90% | 79.84% | 0.834 | 0.504 | 64.44% | 37.77% |
| MindAligner | ICML'25 | Subj. 7 | 0.183 | 0.407 | 81.45% | 88.31% | 79.92% | **80.83%** | 0.834 | 0.487 | 64.18% | 62.58% |
| MindTuner | AAAI'25 | Subj. 7 | 0.202 | **0.417** | 84.50% | 90.80% | 80.70% | 79.60% | 0.817 | 0.465 | 71.90% | 65.50% |
| Duala | CVPR'26 | Subj. 7 | **0.215** | 0.405 | 85.17% | 91.16% | **83.34%** | 80.72% | **0.805** | **0.463** | **71.94%** | **70.53%** |
| MindBank (Ours) | – | Subj. 7 | **0.215** | 0.398 | **86.30%** | **92.42%** | 82.56% | 80.45% | 0.806 | **0.463** | 71.87% | 69.30% |

Table 2: Comparison of trainable parameters and inference time for different methods on Subject 1.

| **Method** | **Training Parameters** Ridge | MLP | Prior | Total | **Inference** |
|---|---|---|---|---|---|
| MindEye2 | 64.41M | 1903M | 260M | 2.2G | 9.45s |
| MindAligner | 139.23M | 0M | 0M | 139.23M | 9.57s |
| MindTuner | 64.41M | 12.30M | 0M | 76.71M | – |
| Duala | 64.41M | 4.68M | 0M | 69.09M | – |
| **MindBank** | **64.41M** | **4.04M** | **0M** | **68.45M** | **9.65s** |

# 4 Experiments and Results

## 4.1 Dataset

We evaluate our method on the Natural Scenes Dataset (NSD) (Allen et al., 2022), a 7 Tesla (7T) fMRI dataset of eight subjects viewing natural images from Microsoft Common Objects in Context (MS-COCO) (Lin et al., 2014). Subjects completed 30–40 one-hour sessions, with approximately 750 trials per session and 3 seconds per image. We fine-tune on one session and evaluate Subjects 1, 2, 5, and 7, who completed all 40 sessions.

## 4.2 Implementation Details

Our implementation follows MindEye2 and MindTuner (Scotti et al., 2024; Gong et al., 2025). We optimize the ridge layer and rank-8 low-rank adaptation (LoRA) adapters with a scaling factor of 8 (Hu et al., 2022). Table 2 summarizes the trainable parameters of Subject 1: MindBank updates 64.41M ridge parameters and 4.04M MLP parameters, for 68.45M in total. Compared with other methods, our approach achieves better performance with fewer trainable parameters. For MindBank, the weights are $\lambda_1 = 0.05$ and $\lambda_2 = 0.02$, and the CSA weight is $\lambda_{\mathrm{CSA}} = 0.30$. The base channel-interpolation strength $\alpha_{\mathrm{base}}$ is 0.10 for Subjects 1 and 2 and 0.30 for Subjects 5 and 7, with an

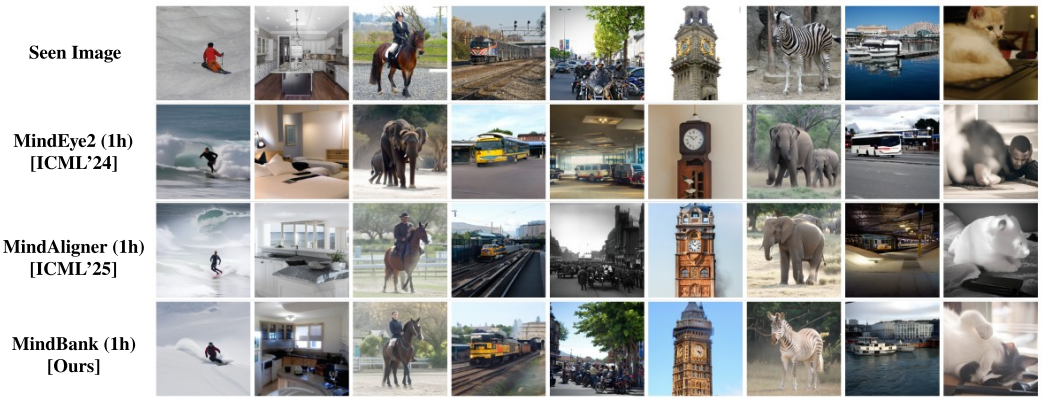



Figure 4: Qualitative reconstruction comparison under one-hour adaptation. Rows show the seen image, MindEye2, MindAligner, and MindBank. MindBank more consistently recovers the stimulus category and global scene composition.

Table 3: Component ablation on Subject 2 under the one-hour adaptation protocol. The best result is shown in bold.

| Component | | | | Low-Level | | | | High-Level | | | | Retrieval | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BANK | RC | SDA | CSA | PixCorr↑ | SSIM↑ | Alex(2)↑ | Alex(5)↑ | Incep↑ | CLIP↑ | Eff.↓ | SwAV↓ | Image↑ | Brain↑ |
| – | – | – | – | 0.217 | 0.423 | 88.23% | 94.01% | 83.56% | 82.30% | 0.808 | 0.455 | **95.12%** | 91.00% |
| √ | – | √ | √ | **0.233** | **0.425** | 89.75% | 95.08% | 85.55% | 83.43% | **0.785** | 0.445 | 92.09% | **92.49%** |
| √ | √ | – | √ | 0.231 | 0.421 | 89.64% | 94.88% | 85.03% | 83.35% | **0.785** | **0.444** | 90.04% | 92.42% |
| √ | √ | √ | – | 0.228 | **0.425** | 89.09% | 94.62% | 85.55% | 82.88% | 0.794 | 0.449 | 94.62% | 91.99% |
| √ | √ | √ | √ | **0.233** | **0.425** | **89.98%** | **95.14%** | **85.59%** | **83.63%** | **0.785** | 0.445 | 92.47% | 92.37% |

alignment floor of $\tau = 0.5$. SDA uses an augmentation strength of $\gamma = 0.5$. CSA uses a threshold of $\eta = 0.9$ to admit perturbed features. The experimental settings for all results reported are detailed in the supplement, which also provides additional results.

### 4.3 Results

**Image and Brain Retrieval** Table 1 compares MindBank with cross-subject visual decoding methods (Dai et al., 2025; Li et al., 2026). MindBank obtains the second-best average image retrieval of 83.83% and the best average brain retrieval of 81.64%. This represents a 5.69-point gain in brain retrieval over MindTuner, while the image retrieval remains within 0.66 points of the best-performing Duala. On Subject 1, brain retrieval rises from 77.63% for the MindEye2 model to 93.72% for MindBank. This gain is accompanied by stronger high-level reconstruction scores, indicating that the bank supervision improves the semantic organization of the target representation.

**Image Reconstruction** Across four subjects, MindBank performs best on Alex(2), Alex(5) (Krizhevsky et al., 2012), Inception (Szegedy et al., 2016), and CLIP, reaching 88.33%, 94.12%, 85.62%, and 83.64%. It improves Alex(2) and Alex(5) over Duala by 0.4 and 0.7 points while remaining competitive in PixCorr, EfficientNet (Tan & Le, 2019), and SwAV (Caron et al., 2020); the corresponding gains on Subject 7 reach 1.13 and 1.26 points. Figure 4 further shows more consistent category and scene-layout recovery, including the correct cat reconstruction in the final example.

## 5 Ablation Studies and Analysis

**Effectiveness of Key Components** Table 3 evaluates the contributions of each component in MindBank under the one-hour adaptation setting. All Source Bank variants improve PixCorr, Alex(2), Alex(5), and Inception over the baseline. Specifically, PixCorr increases from 0.217 to 0.228–0.233, while Inception rises from 83.56% to 85.03%–85.59%. SDA raises image retrieval

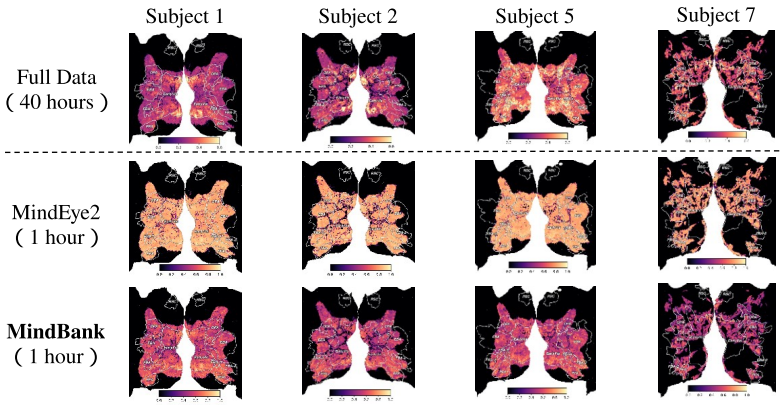



Figure 5: Visualization of transfer quantity in brain heatmaps.

Table 4: Complete sensitivity results for the MindBank loss weights on Natural Scenes Dataset Subject 2 under the one-hour adaptation protocol.

| Loss Weight | | | Low-Level | | | | High-Level | | | | Retrieval | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_1$ | $\lambda_2$ | $\lambda_{\text{CSA}}$ | PixCorr↑ | SSIM↑ | Alex(2)↑ | Alex(5)↑ | Incep↑ | CLIP↑ | Eff.↓ | SwAV↓ | Image↑ | Brain↑ |
| 0.05 | 0.02 | 0.30 | 0.233 | 0.425 | 89.98% | 95.14% | 85.59% | 83.63% | 0.785 | 0.445 | 92.47% | 92.37% |
| 0.00 | 0.02 | 0.30 | 0.232 | 0.426 | 89.74% | 94.65% | 85.71% | 83.01% | 0.790 | 0.446 | 92.39% | 91.42% |
| 0.10 | 0.02 | 0.30 | 0.232 | 0.425 | 89.76% | 94.87% | 84.87% | 83.73% | 0.788 | 0.445 | 90.16% | 92.68% |
| 0.05 | 0.00 | 0.30 | 0.230 | 0.428 | 89.83% | 94.81% | 84.50% | 82.81% | 0.797 | 0.452 | 91.12% | 91.84% |
| 0.05 | 0.05 | 0.30 | 0.232 | 0.425 | 89.60% | 94.91% | 85.37% | 83.60% | 0.782 | 0.445 | 91.38% | 92.21% |
| 0.05 | 0.02 | 0.50 | 0.234 | 0.422 | 89.62% | 95.19% | 85.89% | 84.19% | 0.779 | 0.443 | 90.04% | 92.38% |

from 90.04% to 92.47%, while CSA increases CLIP from 82.88% to 83.63% and brain retrieval from 91.99% to 92.37%. Overall, the components provide complementary gains, and their combination yields the strongest balance across reconstruction and retrieval.

**Sensitivity to Loss Weights** Table 4 evaluates the main bank weight $\lambda_1$, bidirectional auxiliary weight $\lambda_2$, and Category-aware Semantic Alignment weight $\lambda_{\text{CSA}}$ on Subject 2. MindBank is robust across the tested settings, with stable reconstruction quality and brain retrieval. A larger $\lambda_1$ favors brain retrieval but reduces image retrieval, reaching 92.68% and 90.16%, respectively, at 0.10. The auxiliary weight $\lambda_2$ has a moderate effect, and its default value 0.02 provides the most balanced low-level, semantic, and retrieval performance. Increasing $\lambda_{\text{CSA}}$ strengthens semantic metrics, raising CLIP to 84.19%, while image retrieval decreases to 90.04% and brain retrieval remains stable.

**Neuroscience Interpretability** To further investigate cross-subject functional alignment, we compute Transfer Quantity (TQ) values from ridge regression and use them to visualize brain regions associated with stimulus representations. As shown in Figure 5, high TQ values are concentrated in the occipital place area (OPA), extrastriate body area (EBA), and parahippocampal place area (PPA). MindBank recovers more clearly defined cortical patterns that resemble those obtained with the full dataset, in contrast to the diffuse patterns produced by MindEye2.

## 6 Conclusion

MindBank addresses semantic drift in limited-data cross-subject fMRI decoding by transferring shared structure from pretrained source decoders. Source Bank aggregates source representations, while RC weights supervision by sample- and channel-level reliability. SDA turns cross-source variation into controlled target perturbations, while CSA preserves category structure. On NSD, MindBank achieves strong retrieval and reconstruction using one hour of target data. Representation and cortical analyses reveal coherent semantic organization and clearer activation patterns. MindBank enables data-efficient decoder adaptation across individuals.

## AI USE STATEMENT

We used ChatGPT to assist with manuscript formatting, check for formatting issues, and search for relevant literature. The authors take responsibility for the final content and presentation of the paper.

## REFERENCES

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. doi: 10.1038/s41593-021-00962-x.

Guangyin Bao, Qi Zhang, Zixuan Gong, Jialei Zhou, Wei Fan, Kun Yi, Usman Naseem, Liang Hu, and Duoqian Miao. Wills Aligner: Multi-subject collaborative brain visual decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):14194–14202, 2025. doi: 10.1609/aaai.v39i13.33554.

Thomas Bazeille, Hugo Richard, Hicham Janati, and Bertrand Thirion. Local optimal transport for functional brain template estimation. In *Information Processing in Medical Imaging*, volume 11492 of *Lecture Notes in Computer Science*, pp. 237–248. Springer, 2019. doi: 10.1007/978-3-030-20351-1_18.

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. In *Advances in Neural Information Processing Systems*, volume 32, pp. 6514–6524, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924, 2020.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22710–22720, 2023.

David D. Cox and Robert L. Savoy. Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2):261–270, 2003. doi: 10.1016/S1053-8119(03)00049-1.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12214–12228. PMLR, 2025.

James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. doi: 10.1016/j.neuron.2012.01.010.

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. doi: 10.1162/imag_a_00170.

Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. MindTuner: Cross-subject visual decoding with visual fingerprint and semantic correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):14247–14255, 2025. doi: 10.1609/aaai.v39i13.33560.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):15037, 2017. doi: 10.1038/ncomms15037.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. doi: 10.1038/nature06713.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pp. 1097–1105, 2012.

Shumeng Li, Jintao Guo, Jian Zhang, Yulin Zhou, Luyang Cao, and Yinghuan Shi. Duala: Dual-level alignment of subjects and stimuli for cross-subject fMRI decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42722–42731, 2026.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Yunfeng Lin, Jiangbei Li, and Hanjing Wang. DCNN-GAN: Reconstructing realistic image from fMRI. In *2019 16th International Conference on Machine Vision Applications*, pp. 1–6, 2019. doi: 10.23919/MVA.2019.8757985.

Yulong Liu, Yongqiang Ma, Guibo Zhu, Haodong Jing, and Nanning Zheng. See through their minds: Learning transferable brain decoding models from cross-subject fMRI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):5730–5738, 2025. doi: 10.1609/aaai.v39i6.32611.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Weijian Mai and Zhijun Zhang. UniBrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity, 2023. arXiv preprint arXiv:2308.07428.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, volume 21, pp. 1041–1048, 2008.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011. doi: 10.1016/j.neuroimage.2010.07.073.

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

Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 233–243, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Abraham. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24705–24728, 2023.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059. PMLR, 2024.

Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel A. J. van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018. doi: 10.1016/j.neuroimage.2018.07.043.

Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 98083–98110, 2024. doi: 10.52202/079017-3113.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15(1):e1006633, 2019. doi: 10.1371/journal.pcbi.1006633.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pp. 369–386. SPIE, 2019. doi: 10.1117/12.2520589.

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 and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 2019.

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

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, 2024a.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Zicheng Wang, Zhen Zhao, Luping Zhou, and Parashkev Nachev. UniBrain: A unified model for cross-subject brain decoding, 2024b. arXiv preprint arXiv:2412.19487.

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

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, 2024a.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. UMBRAE: Unified multimodal brain decoding. In *Computer Vision – ECCV 2024*, volume 15065 of *Lecture Notes in Computer Science*, pp. 242–259. Springer, 2024b. doi: 10.1007/978-3-031-72667-5_14.

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.

Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.

Bohan Zeng, Shanglin Li, Xuhui Liu, Sicheng Gao, Xiaolong Jiang, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. Controllable mind visual diffusion model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):6935–6943, 2024. doi: 10.1609/aaai.v38i7.28519.

# SUPPLEMENTARY MATERIAL

## A IMPLEMENTATION DETAILS

This section describes the dataset and architecture, training configuration, hyperparameter selection, evaluation metrics, inference procedure, and accompanying code.

### A.1 NSD DATASET AND ARCHITECTURE

The Natural Scenes Dataset (NSD) is a 7 Tesla functional magnetic resonance imaging (fMRI) dataset with responses from 8 subjects viewing natural images sampled from Microsoft Common Objects in Context (MS-COCO) (Allen et al., 2022; Lin et al., 2014). Each subject completed 30–40 one-hour scanning sessions, with 750 trials per session. We follow the one-hour target-adaptation protocol, using one target-subject session and frozen pretrained models.

MindEye2 maps subject-specific voxel responses to a shared latent space with a ridge regression layer, followed by a shared multilayer perceptron backbone that produces semantic features and image tokens (Scotti et al., 2024). Its retrieval branch aligns predicted brain tokens with Contrastive Language–Image Pre-training image tokens through a bidirectional contrastive objective. A diffusion prior predicts high-level image representations, and a low-level reconstruction branch provides coarse visual structure. MindBank retains these decoding branches and adds its Source Bank, a collection of frozen source-subject representations that provides semantic supervision during target adaptation. Reliability Calibration (RC) weights this supervision, Source Disagreement Augmentation (SDA) perturbs target features according to source variation, and Category-aware Semantic Alignment (CSA) organizes the adapted representations by semantic category.

### A.2 TRAINING CONFIGURATION

All target-adaptation runs use one NSD session, a batch size of 10, and 150 epochs. We optimize the target projection and rank-8 low-rank adaptation (LoRA) adapters with a scaling factor of 8 (Hu et al., 2022). Optimization uses AdamW with a maximum learning rate of $3 \times 10^{-4}$ and a OneCycle schedule (Loshchilov & Hutter, 2019; Smith & Topin, 2019). The retrieval, diffusion-prior, and low-level reconstruction loss scales are 1, 30, and 0.5, respectively. Following MindEye2, the retrieval objective uses the bidirectional MixCo (BiMixCo) loss during the first one-third of training and the soft Contrastive Language–Image Pre-training (SoftCLIP) loss thereafter.

### A.3 COMPUTING ENVIRONMENT

Each target-adaptation run is trained on one NVIDIA A100 graphics processing unit with 80 gigabytes (GB) of memory. The experiments use Ubuntu 22.04.5 long-term support (LTS) on a host with an Intel Xeon Gold 6348 processor. The software environment includes Python 3.11.15, PyTorch 2.1.0 with CUDA 12.1 and the CUDA Deep Neural Network library (cuDNN) 8.9.2, torchvision 0.16.0, Transformers 4.37.2, Accelerate 0.24.1, xFormers 0.0.22.post7, NumPy 1.26.4, and pandas 2.2.0.

### A.4 HYPERPARAMETER SELECTION

We retain the optimization and reconstruction settings of MindEye2 and MindTuner (Scotti et al., 2024; Gong et al., 2025). On NSD Subject 2 under the one-hour protocol, the reported one-factor-at-a-time sensitivity study evaluates the main bank weight $\lambda_1 \in \{0, 0.05, 0.10\}$, auxiliary bank weight $\lambda_2 \in \{0, 0.02, 0.05\}$, CSA weight $\lambda_{\mathrm{CSA}} \in \{0.30, 0.50\}$, and base interpolation strength $\alpha_{\mathrm{base}} \in \{0.05, 0.10, 0.15, 0.20\}$. For Subject 7, we additionally evaluate $\alpha_{\mathrm{base}} \in \{0.10, 0.20, 0.30, 0.40\}$. The base interpolation strength controls the overall weight of Source Bank injection. The cross-subject representation analysis shows that the mean target–source similarity is lower when Subjects 1 and 2 serve as targets than when Subjects 5 and 7 serve as targets. This lower agreement indicates lower confidence in Source Bank guidance, so we use $\alpha_{\mathrm{base}} = 0.10$ for Subjects 1 and 2 to assign a smaller injection weight. The higher agreement for Subjects 5 and 7 supports $\alpha_{\mathrm{base}} = 0.30$. Across all four subjects, performance is not highly sensitive to this value; Table 5 illustrates the stable results

Table 5: Ablation of the interpolation strength on Natural Scenes Dataset Subjects 2 and 7 under the one-hour adaptation protocol. Channel-adaptive interpolation uses the base strength $\alpha_{\text{base}}$, while Fixed uses a uniform interpolation coefficient. For Subject 7, all other training settings are held constant.

| Interpolation | Strength | Subject | Low-Level | | | | High-Level | | | | Retrieval | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | PixCorr↑ | SSIM↑ | Alex(2)↑ | Alex(5)↑ | Incep↑ | CLIP↑ | Eff.↓ | SwAV↓ | Image↑ | Brain↑ |
| Channel-Adaptive | 0.10 | Subj. 2 | 0.233 | 0.425 | 89.98% | 95.14% | 85.59% | 83.63% | 0.785 | 0.445 | 92.47% | 92.37% |
| Channel-Adaptive | 0.05 | Subj. 2 | 0.233 | 0.425 | 89.71% | 94.91% | 85.74% | 83.65% | 0.788 | 0.446 | 91.47% | 92.38% |
| Channel-Adaptive | 0.15 | Subj. 2 | 0.236 | 0.425 | 89.46% | 94.59% | 85.72% | 83.74% | 0.781 | 0.444 | 92.67% | 91.69% |
| Channel-Adaptive | 0.20 | Subj. 2 | 0.233 | 0.425 | 89.11% | 94.68% | 84.80% | 83.07% | 0.794 | 0.449 | 92.02% | 92.13% |
| Fixed | 0.10 | Subj. 2 | 0.229 | 0.425 | 89.69% | 94.65% | 85.34% | 83.39% | 0.790 | 0.447 | 92.04% | 92.31% |
| Channel-Adaptive | 0.10 | Subj. 7 | 0.217 | 0.396 | 85.22% | 91.82% | 82.63% | 80.50% | 0.811 | 0.464 | 72.58% | 70.07% |
| Channel-Adaptive | 0.20 | Subj. 7 | 0.219 | 0.398 | 85.45% | 91.92% | 82.36% | 80.01% | 0.811 | 0.464 | 71.79% | 70.09% |
| Channel-Adaptive | 0.30 | Subj. 7 | 0.215 | 0.398 | 86.30% | 92.42% | 82.56% | 80.45% | 0.806 | 0.463 | 71.87% | 69.30% |
| Channel-Adaptive | 0.40 | Subj. 7 | 0.214 | 0.400 | 85.63% | 91.87% | 82.27% | 80.22% | 0.812 | 0.468 | 70.04% | 67.64% |

over the tested ranges for Subjects 2 and 7. The selected values provide balanced performance across low-level reconstruction, high-level semantic similarity, and bidirectional retrieval. The final loss weights are $\lambda_1 = 0.05$, $\lambda_2 = 0.02$, and $\lambda_{\text{CSA}} = 0.30$. The alignment floor $\tau = 0.5$, augmentation strength $\gamma = 0.5$, and semantic-consistency threshold $\eta = 0.9$ are fixed.

### A.5 Explanation of Metrics

**Low-level metrics.** Pixelwise correlation (PixCorr) measures the linear correspondence of reconstructed and ground-truth pixels. The structural similarity index (SSIM) (Wang et al., 2004) evaluates luminance, contrast, and structure. Alex(2) and Alex(5) measure the correlation between AlexNet activations from the second and fifth convolutional layers (Krizhevsky et al., 2012).

**High-level metrics.** To quantify semantic alignment, we follow prior decoding studies and compare deep features from Inception-V3, CLIP Vision Transformer (ViT)-L/14, EfficientNet-B1, and Swapping Assignments between Views (SwAV) with a 50-layer residual network (ResNet50). Inception-V3 (Incep) and Contrastive Language–Image Pre-training (CLIP) scores use two-way identification: a reconstruction is compared with its paired image and a distractor, and the score is the percentage of correct discriminations (Szegedy et al., 2016; Radford et al., 2021). EfficientNet (Eff.) and SwAV evaluate the average correlation distance between the reconstruction and image embeddings; lower values indicate closer semantic features (Caron et al., 2020).

**Retrieval metrics.** Retrieval evaluates how precisely the latent representations preserve cross-modal associations. Image retrieval selects the image embedding with the highest cosine similarity to an fMRI embedding. Brain retrieval performs the reciprocal search, selecting the fMRI embedding with the highest similarity to an image embedding. A trial is correct when the paired item is retrieved.

### A.6 Inference Details

The Source Bank and Reliability Calibration, bidirectional semantic consistency loss, category-aware semantic alignment loss, and source disagreement augmentation operate during training. During inference, MindBank directly maps a new subject's fMRI response to the aligned latent space without additional optimization. The high-level and low-level outputs are combined with a 3:1 ratio for final image reconstruction. The number of inference parameters is essentially identical to that of existing methods.

### A.7 Code of MindBank

For reproducibility, the supplementary code package includes the MindBank training and inference programs, and implementations of all loss terms.

# B METHOD DETAILS

The following subsections expand the four source-guided components introduced above: the Source Bank, RC, SDA, and CSA. They provide the corresponding algorithms and calibration equations.

## B.1 DETAILS OF THE SOURCE BANK

Algorithm 1 expands the Source Bank computation for one mini-batch. Each source ridge layer and the source backbone are frozen. The four pathways $SS$, $ST$, $TS$, and $TT$ provide the consensus and cross-backbone views used by the bank objective. The source consensus $SS$ is interpolated with the target representation $TT$ using the source contribution $\alpha$.

**Algorithm 1** Source Bank construction.

**Require:** Target fMRI batch $v$; target modules $R_t, G_t$; frozen source ridge layers $\{R_s\}_{s=1}^{K}$, frozen source backbone $G_{\mathrm{src}}$, and shared linear layer $J$; $\alpha$

**Ensure:** Fused target feature $\widetilde{TT}$ and bank paths

$x_t \leftarrow R_t(v)$
$TT \leftarrow J(G_t(x_t))$, $TS \leftarrow J(G_{\mathrm{src}}(x_t))$
**for** each source subject $s$ **do**
  $x_s \leftarrow R_s(\mathrm{Align}(v, \dim(R_s)))$
  $SS^s \leftarrow J(G_{\mathrm{src}}(x_s))$, $ST^s \leftarrow J(G_t(x_s))$
**end for**
$SS \leftarrow K^{-1}\sum_s SS^s$, $ST \leftarrow K^{-1}\sum_s ST^s$
$\widetilde{TT} \leftarrow (1-\alpha)TT + \alpha SS$
**return** $\widetilde{TT}, \{SS, ST, TS, TT\}$

## B.2 DETAILS OF RELIABILITY CALIBRATION

Reliability Calibration assigns sample- and channel-level weights to the Source Bank contributions.

At the sample level, it assigns greater weights to samples with more consistent source representations. Specifically, the semantic features are flattened, and the agreement between each source representation $SS^s$ and the bank consensus $SS$ is measured as

$$u = \frac{1}{K}\sum_{s=1}^{K}\cos(SS^s, SS). \tag{15}$$

$u$ represents the agreement score of the Source Bank and directly serves as the sample-level reliability weight. For a mini-batch $\mathcal{B}$, the calibrated bank objective is

$$\mathcal{L}_{\mathrm{bsc}} = \frac{\sum_{b\in\mathcal{B}} u_b \ell_{\mathrm{bsc},b}}{\max\left(\sum_{b\in\mathcal{B}} u_b, \epsilon\right)}, \tag{16}$$

where $\ell_{\mathrm{bsc},b}$ is the per-sample BSC loss, $b$ indexes the mini-batch, and $\epsilon$ is a small positive constant. The same normalized weighting is applied to the main and auxiliary terms. Thus, samples with more coherent source representations provide stronger bank supervision.

At the channel level, Reliability Calibration adjusts the source contribution according to the agreement and target–source alignment of each channel. For channel $c$, the target–source alignment $q_c$ is

$$q_c = \frac{\cos(TT_c, SS_c) + 1}{2}, \tag{17}$$

and $u_c$ is the agreement score for channel $c$ computed in the same way as Equation 15. These two factors are combined into the channel reliability

$$\rho_c = u_c\left(\tau + (1-\tau)q_c\right), \tag{18}$$

where $\tau = 0.5$ is the alignment floor. The interpolation coefficient is then calculated as

$$\alpha_c = \alpha_{\mathrm{base}}\frac{\rho_c}{C^{-1}\sum_{j=1}^{C}\rho_j}, \tag{19}$$

where $\alpha_{\text{base}}$ controls the overall source contribution. The calibrated target representation is

$$\widetilde{TT}_c = (1 - \alpha_c)TT_c + \alpha_c SS_c. \tag{20}$$

This design incorporates more source semantics into channels with high source agreement and strong target–source alignment while preserving target-specific information in less reliable channels.

**Algorithm 2** Reliability Calibration (RC).

**Require:** Mini-batch $\mathcal{B}$; for each sample, source representations $\{SS^s\}_{s=1}^K$, source consensus $SS$, target representation $TT$, and BSC loss $\ell_{\text{bsc}}$; $\alpha_{\text{base}}, \tau, \epsilon$

**Ensure:** Calibrated representations $\widetilde{TT}$ and bank objective $\mathcal{L}_{\text{bsc}}$

**for** each sample in $\mathcal{B}$ **do**
  $u \leftarrow K^{-1} \sum_s \cos(SS^s, SS)$
  **for** each feature channel $c$ **do**
    $q_c \leftarrow (\cos(TT_c, SS_c) + 1)/2$
    $u_c \leftarrow K^{-1} \sum_s \cos(SS_c^s, SS_c)$
    $\rho_c \leftarrow u_c[\tau + (1 - \tau)q_c]$
  **end for**
  $\overline{\rho} \leftarrow C^{-1} \sum_c \rho_c$
  **for** each feature channel $c$ **do**
    $\alpha_c \leftarrow \alpha_{\text{base}} \rho_c / \overline{\rho}$
    $\widetilde{TT}_c \leftarrow (1 - \alpha_c)TT_c + \alpha_c SS_c$
  **end for**
  Store $u$, $\ell_{\text{bsc}}$, and $\widetilde{TT}$
**end for**
$\mathcal{L}_{\text{bsc}} \leftarrow \sum_{\mathcal{B}} u \ell_{\text{bsc}} / \max(\sum_{\mathcal{B}} u, \epsilon)$
**return** $\{\widetilde{TT}\}, \mathcal{L}_{\text{bsc}}$

### B.3 Details of SDA

Source Disagreement Augmentation (SDA) adjusts a target latent according to the variation observed across source subjects. For each sample, $z_t \in \mathbb{R}^{1 \times D}$ denotes the target latent and $\{z_s\}_{s=1}^K$ denote the source latents. SDA first averages the source latents to obtain the source consensus $\mu_{\text{src}}$ and computes their feature-wise standard deviation $\sigma_{\text{src}}$ across the source-subject dimension. It then aggregates $\sigma_{\text{src}}$ over the $D$ feature dimensions using its root-mean-square magnitude, yielding the scalar source-disagreement scale $\widetilde{\sigma}_{\text{src}}$. The centered target residual $\Delta = z_t - \mu_{\text{src}}$ describes the target direction from the source consensus, and $\sigma$ measures its scale across feature dimensions. The factor $v$ blends the identity scale with the ratio $\widetilde{\sigma}_{\text{src}}/\sigma$, controlled by the augmentation strength $\gamma$. Applying $v$ to $\Delta$ preserves the source consensus and target-residual direction while adjusting its magnitude to reflect cross-source disagreement. The positive constant $\epsilon$ provides numerical stability in the scale estimates. As detailed in Algorithm 3, the resulting perturbation is applied with probability $p = 0.5$ during training, and $z_t$ is retained otherwise.

**Algorithm 3** Source Disagreement Augmentation (SDA).

**Require:** Target latent $z_t \in \mathbb{R}^{1 \times D}$; source latents $\{z_s\}_{s=1}^K$; augmentation strength $\gamma$; probability $p$; $\epsilon > 0$

**Ensure:** Augmented target latent $\widehat{z}_t$

$\mu_{\text{src}} \leftarrow K^{-1} \sum_s z_s$
$\sigma_{\text{src}} \leftarrow \sqrt{\text{Var}_s(z_s) + \epsilon}$
$\widetilde{\sigma}_{\text{src}} \leftarrow \sqrt{D^{-1} \sum_d \sigma_{\text{src},d}^2 + \epsilon}$
$\Delta \leftarrow z_t - \mu_{\text{src}}$, $\sigma \leftarrow \sqrt{\text{Var}_d(\Delta_d) + \epsilon}$
$v \leftarrow (1 - \gamma) + \gamma \widetilde{\sigma}_{\text{src}} / \sigma$
$\widehat{z}_t \leftarrow \mu_{\text{src}} + v\Delta$
With probability $1 - p$, set $\widehat{z}_t \leftarrow z_t$
**return** $\widehat{z}_t$

### B.4 Details of CSA

CSA uses frozen CLIP image features to assign each training image a category label and stores normalized target features in a first-in-first-out memory bank. Algorithm 4 calculates the mean positive similarity and the hardest negative similarity for each anchor. It also admits an SDA-perturbed feature only when its cosine similarity to the original target feature exceeds the threshold $\eta = 0.9$, retaining semantically consistent augmentations.

**Algorithm 4** Category-aware Semantic Alignment (CSA).

**Require:** Target features $z$, SDA features $\widehat{z}$, training images $I$, frozen category assigner $h$, memory bank $\mathcal{M}$, margin $m$, and threshold $\eta$
**Ensure:** CSA loss $\mathcal{L}_{\mathrm{CSA}}$ and updated $\mathcal{M}$

$y \leftarrow h(I)$, $z \leftarrow \mathrm{normalize}(z)$, $\widehat{z} \leftarrow \mathrm{normalize}(\widehat{z})$
Initialize the valid-loss set $\mathcal{V} \leftarrow \varnothing$
**for** each anchor $(z_i, y_i)$ **do**
  $\mathcal{M}_i^+ \leftarrow \{b \in \mathcal{M} : \mathrm{label}(b) = y_i\}$
  $\mathcal{M}_i^- \leftarrow \{b \in \mathcal{M} : \mathrm{label}(b) \neq y_i\}$
  **if** $\mathcal{M}_i^+ \neq \varnothing$ and $\mathcal{M}_i^- \neq \varnothing$ **then**
    $s_i^+ \leftarrow |\mathcal{M}_i^+|^{-1} \sum_{b \in \mathcal{M}_i^+} z_i^\top b$
    $s_i^- \leftarrow \max_{b \in \mathcal{M}_i^-} z_i^\top b$
    $\ell_i \leftarrow \max(0, m + s_i^- - s_i^+)$
    Add $\ell_i$ to $\mathcal{V}$
  **end if**
  **if** $\cos(z_i, \widehat{z}_i) > \eta$ **then**
    Enqueue $(\widehat{z}_i, y_i)$
  **end if**
**end for**
$\mathcal{L}_{\mathrm{CSA}} \leftarrow |\mathcal{V}|^{-1} \sum_{\ell \in \mathcal{V}} \ell$
Enqueue $(z_i, y_i)$ for all $i$
**return** $\mathcal{L}_{\mathrm{CSA}}, \mathcal{M}$

## C Additional Results

This section presents ablations of the interpolation strength, a comparison with reproduced MindTuner results, cross-subject representation similarities, and subject-specific qualitative analyses.

### C.1 Ablation Results of $\alpha_{\mathrm{base}}$

To investigate how $\alpha_{\mathrm{base}}$ affects reconstruction and retrieval, Table 5 evaluates channel-adaptive interpolation on Subjects 2 and 7 as representatives of the lower- and higher-alignment groups. Tables 7–10 show target-to-average-source similarities of 0.664 and 0.729 for Subjects 1 and 2, compared with 0.840 and 0.850 for Subjects 5 and 7, motivating strengths of 0.10 and 0.30 for the two groups. On Subject 2, PixCorr and SSIM vary by at most 0.003, and channel-adaptive interpolation at 0.10 improves most metrics over fixed interpolation. On Subject 7, 0.30 improves Alex(2), Alex(5), EfficientNet, and SwAV, while smaller strengths favor retrieval. Overall, $\alpha_{\mathrm{base}}$ reflects target–source alignment but remains robust across reconstruction, semantic similarity, and retrieval metrics for both representative subjects.

### C.2 t-SNE Analysis

To examine how MindBank changes the semantic organization of target representations, Figure 6 compares MindEye2 and MindBank using t-distributed stochastic neighbor embedding (t-SNE). The seven image categories form more compact within-category groups with clearer between-category separation under MindBank. Brain Retrieval increases from 77.63% to 93.72%, while Cluster Ratio increases from 2.237 to 3.160. Retrieval measures cross-modal correspondence, whereas Cluster

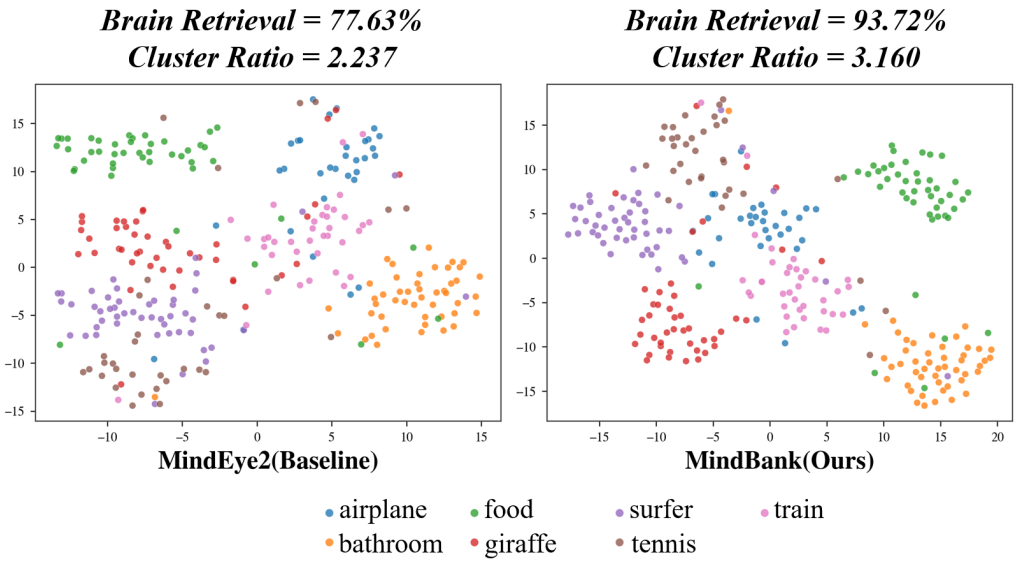



Figure 6: t-SNE comparison of target representations learned by MindEye2 and MindBank. Each point is a held-out fMRI–image pair, and colors denote image categories. MindBank yields more compact and better-separated category clusters. Brain Retrieval and Cluster Ratio, defined as the mean distance between category centroids divided by the mean within-category radius, summarize the corresponding representation quality.

Table 6: Quantitative comparison with our reproduction of MindTuner. We reproduce MindTuner on Natural Scenes Dataset Subjects 1, 2, 5, and 7 under the same one-hour adaptation protocol used for MindBank. Bold indicates the best performance in each subject block.

| Method | Venue | Subject | Low-Level | | | | High-Level | | | | Retrieval | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | PixCorr↑ | SSIM↑ | Alex(2)↑ | Alex(5)↑ | Incep↑ | CLIP↑ | Eff.↓ | SwAV↓ | Image↑ | Brain↑ |
| MindTuner | AAAI'25 | Avg. | 0.218 | 0.410 | 86.7% | 92.7% | 83.8% | 81.4% | 0.804 | 0.456 | 83.8% | 80.5% |
| **MindBank (Ours)** | – | Avg. | **0.228** | **0.412** | **88.33%** | **94.12%** | **85.62%** | **83.64%** | **0.784** | **0.445** | **83.83%** | **81.64%** |
| MindTuner | AAAI'25 | Subj. 1 | 0.241 | 0.415 | 89.56% | 94.46% | 84.83% | 83.12% | 0.788 | 0.444 | **94.77%** | 92.29% |
| **MindBank (Ours)** | – | Subj. 1 | **0.259** | **0.417** | **91.04%** | **95.40%** | **87.22%** | **85.87%** | **0.774** | **0.432** | 94.23% | **93.72%** |
| MindTuner | AAAI'25 | Subj. 2 | 0.231 | 0.423 | 88.78% | 94.56% | 84.73% | 81.98% | 0.794 | 0.447 | **92.91%** | 90.76% |
| **MindBank (Ours)** | – | Subj. 2 | **0.233** | **0.425** | **89.98%** | **95.14%** | **85.59%** | **83.63%** | **0.785** | **0.445** | 92.47% | **92.37%** |
| MindTuner | AAAI'25 | Subj. 5 | 0.194 | 0.408 | 84.72% | 91.56% | 83.89% | 81.21% | 0.808 | 0.459 | 75.63% | 70.40% |
| **MindBank (Ours)** | – | Subj. 5 | **0.203** | **0.409** | **86.01%** | **93.53%** | **87.11%** | **84.60%** | **0.770** | **0.438** | **76.76%** | **71.19%** |
| MindTuner | AAAI'25 | Subj. 7 | 0.204 | 0.394 | 83.75% | 90.29% | 81.57% | 79.16% | 0.828 | 0.475 | **71.92%** | 68.69% |
| **MindBank (Ours)** | – | Subj. 7 | **0.215** | **0.398** | **86.30%** | **92.42%** | **82.56%** | **80.45%** | **0.806** | **0.463** | 71.87% | **69.30%** |

Ratio summarizes category structure within the representation space. Together, these results show that MindBank learns a more discriminative and semantically organized target representation for cross-subject visual decoding with limited target data.

### C.3 Reproducing MindTuner for Fair Comparison

To compare MindBank and MindTuner under a controlled evaluation setting, we reproduce MindTuner (Gong et al., 2025) using the same one-hour target-adaptation protocol and evaluation pipeline as MindBank. All MindTuner entries in Table 6 are obtained from this reproduction. Averaged across the four subjects, MindBank improves all ten metrics: Alex(2), Alex(5), Inception, and CLIP increase by 1.63, 1.42, 1.82, and 2.24 percentage points, while the EfficientNet and SwAV distances decrease by 0.020 and 0.011. Brain retrieval increases from 80.50% to 81.64%, and image retrieval increases from 83.80% to 83.83%. At the subject level, MindBank improves brain retrieval by 1.43, 1.61, 0.79, and 0.61 percentage points for Subjects 1, 2, 5, and 7, respectively, while image retrieval

Table 7: Mean pairwise cosine similarities with Subject 1 as the target. The target–Avg entry is shown in bold.

| | Subj. 1 | Subj. 2 | Subj. 3 | Subj. 4 | Subj. 5 | Subj. 6 | Subj. 7 | Subj. 8 | Avg |
|---|---|---|---|---|---|---|---|---|---|
| Subj. 1 | 1.000 | 0.530 | 0.510 | 0.490 | 0.505 | 0.500 | 0.515 | 0.502 | **0.664** |
| Subj. 2 | 0.530 | 1.000 | 0.536 | 0.521 | 0.491 | 0.513 | 0.523 | 0.520 | 0.767 |
| Subj. 3 | 0.510 | 0.536 | 1.000 | 0.517 | 0.497 | 0.479 | 0.511 | 0.499 | 0.758 |
| Subj. 4 | 0.490 | 0.521 | 0.517 | 1.000 | 0.544 | 0.487 | 0.531 | 0.529 | 0.770 |
| Subj. 5 | 0.505 | 0.491 | 0.497 | 0.544 | 1.000 | 0.500 | 0.541 | 0.525 | 0.765 |
| Subj. 6 | 0.500 | 0.513 | 0.479 | 0.487 | 0.500 | 1.000 | 0.477 | 0.491 | 0.742 |
| Subj. 7 | 0.515 | 0.523 | 0.511 | 0.531 | 0.541 | 0.477 | 1.000 | 0.517 | 0.764 |
| Subj. 8 | 0.502 | 0.520 | 0.499 | 0.529 | 0.525 | 0.491 | 0.517 | 1.000 | 0.764 |
| Avg | **0.664** | 0.767 | 0.758 | 0.770 | 0.765 | 0.742 | 0.764 | 0.764 | 1.000 |

improves by 1.13 percentage points for Subject 5 and remains close for the other subjects. These results show that Source Bank supervision consistently strengthens semantic reconstruction and brain retrieval under the unified adaptation protocol across all four evaluated target subjects.

### C.4 Cross-Subject Representation Similarity

To determine whether the averaged source representation provides a reliable cross-subject reference, we analyze 1,000 unique held-out images for each target subject. For every image, the same target functional magnetic resonance imaging (fMRI) input is passed through the target path and seven frozen source-subject paths after matching their ridge input dimensions. The three fMRI repeats are processed separately; their $256 \times 1664$ representations after the shared linear layer $J$ and before the Contrastive Language–Image Pre-training projection are averaged and flattened to 425,984 dimensions.

We then compute image-wise cosine similarities between every pair of the eight path representations. Avg is formed for each image as the element-wise mean of the seven source representations excluding the target. Each table entry averages the corresponding cosine similarity over 1,000 images and is rounded to three decimal places. The matrices are therefore symmetric, with self-similarities of 1 on the diagonal. Repeating this procedure for each target makes Avg and all off-diagonal entries specific to the corresponding table, yielding matched-image estimates of target-specific cross-subject alignment.

Tables 7–10 show target–Avg similarities of 0.664, 0.729, 0.840, and 0.850 for Subjects 1, 2, 5, and 7, respectively. Each value exceeds the target similarity to the most similar individual source by 0.134, 0.136, 0.068, and 0.072, respectively. Although the absolute similarity varies across target subjects, source averaging provides a consistent alignment advantage in all four cases. The results support the source consensus as a shared structure for target adaptation across subjects under limited target data in the one-hour adaptation setting for all evaluated target subjects.

### C.5 Cross-Subject Reconstruction

To assess whether MindBank preserves visual semantics across target subjects, Figure 7 compares reconstructions of the same stimulus images for all four evaluated subjects under the one-hour adaptation protocol. MindBank recovers the main scene categories and semantic content, including indoor layouts, animals, vehicles, and outdoor scenes, while subject-specific variations are concentrated in texture, viewpoint, and other fine spatial details. These examples show that MindBank consistently preserves category-level semantics across the four subjects while accommodating subject-dependent differences in low-level appearance and spatial detail under limited target data.

Table 8: Mean pairwise cosine similarities with Subject 2 as the target. The target–Avg entry is shown in bold.

| | Subj. 1 | Subj. 2 | Subj. 3 | Subj. 4 | Subj. 5 | Subj. 6 | Subj. 7 | Subj. 8 | Avg |
|---|---|---|---|---|---|---|---|---|---|
| Subj. 1 | 1.000 | 0.582 | 0.550 | 0.551 | 0.570 | 0.577 | 0.537 | 0.528 | 0.784 |
| Subj. 2 | 0.582 | 1.000 | 0.568 | 0.575 | 0.567 | 0.578 | 0.593 | 0.550 | **0.729** |
| Subj. 3 | 0.550 | 0.568 | 1.000 | 0.557 | 0.548 | 0.576 | 0.556 | 0.548 | 0.789 |
| Subj. 4 | 0.551 | 0.575 | 0.557 | 1.000 | 0.583 | 0.557 | 0.566 | 0.562 | 0.792 |
| Subj. 5 | 0.570 | 0.567 | 0.548 | 0.583 | 1.000 | 0.555 | 0.566 | 0.541 | 0.790 |
| Subj. 6 | 0.577 | 0.578 | 0.576 | 0.557 | 0.555 | 1.000 | 0.545 | 0.524 | 0.790 |
| Subj. 7 | 0.537 | 0.593 | 0.556 | 0.566 | 0.566 | 0.545 | 1.000 | 0.541 | 0.779 |
| Subj. 8 | 0.528 | 0.550 | 0.548 | 0.562 | 0.541 | 0.524 | 0.541 | 1.000 | 0.772 |
| Avg | 0.784 | **0.729** | 0.789 | 0.792 | 0.790 | 0.790 | 0.779 | 0.772 | 1.000 |

Table 9: Mean pairwise cosine similarities with Subject 5 as the target. The target–Avg entry is shown in bold.

| | Subj. 1 | Subj. 2 | Subj. 3 | Subj. 4 | Subj. 5 | Subj. 6 | Subj. 7 | Subj. 8 | Avg |
|---|---|---|---|---|---|---|---|---|---|
| Subj. 1 | 1.000 | 0.770 | 0.781 | 0.765 | 0.738 | 0.775 | 0.753 | 0.740 | 0.895 |
| Subj. 2 | 0.770 | 1.000 | 0.778 | 0.771 | 0.772 | 0.778 | 0.749 | 0.765 | 0.900 |
| Subj. 3 | 0.781 | 0.778 | 1.000 | 0.774 | 0.761 | 0.765 | 0.748 | 0.758 | 0.898 |
| Subj. 4 | 0.765 | 0.771 | 0.774 | 1.000 | 0.760 | 0.750 | 0.759 | 0.771 | 0.894 |
| Subj. 5 | 0.738 | 0.772 | 0.761 | 0.760 | 1.000 | 0.733 | 0.738 | 0.741 | **0.840** |
| Subj. 6 | 0.775 | 0.778 | 0.765 | 0.750 | 0.733 | 1.000 | 0.728 | 0.756 | 0.891 |
| Subj. 7 | 0.753 | 0.749 | 0.748 | 0.759 | 0.738 | 0.728 | 1.000 | 0.726 | 0.874 |
| Subj. 8 | 0.740 | 0.765 | 0.758 | 0.771 | 0.741 | 0.756 | 0.726 | 1.000 | 0.883 |
| Avg | 0.895 | 0.900 | 0.898 | 0.894 | **0.840** | 0.891 | 0.874 | 0.883 | 1.000 |

Table 10: Mean pairwise cosine similarities with Subject 7 as the target. The target–Avg entry is shown in bold.

| | Subj. 1 | Subj. 2 | Subj. 3 | Subj. 4 | Subj. 5 | Subj. 6 | Subj. 7 | Subj. 8 | Avg |
|---|---|---|---|---|---|---|---|---|---|
| Subj. 1 | 1.000 | 0.789 | 0.793 | 0.761 | 0.777 | 0.782 | 0.758 | 0.762 | 0.899 |
| Subj. 2 | 0.789 | 1.000 | 0.792 | 0.789 | 0.784 | 0.794 | 0.778 | 0.775 | 0.908 |
| Subj. 3 | 0.793 | 0.792 | 1.000 | 0.768 | 0.775 | 0.802 | 0.758 | 0.778 | 0.905 |
| Subj. 4 | 0.761 | 0.789 | 0.768 | 1.000 | 0.785 | 0.769 | 0.768 | 0.777 | 0.894 |
| Subj. 5 | 0.777 | 0.784 | 0.775 | 0.785 | 1.000 | 0.777 | 0.778 | 0.786 | 0.900 |
| Subj. 6 | 0.782 | 0.794 | 0.802 | 0.769 | 0.777 | 1.000 | 0.765 | 0.766 | 0.903 |
| Subj. 7 | 0.758 | 0.778 | 0.758 | 0.768 | 0.778 | 0.765 | 1.000 | 0.755 | **0.850** |
| Subj. 8 | 0.762 | 0.775 | 0.778 | 0.777 | 0.786 | 0.766 | 0.755 | 1.000 | 0.894 |
| Avg | 0.899 | 0.908 | 0.905 | 0.894 | 0.900 | 0.903 | **0.850** | 0.894 | 1.000 |

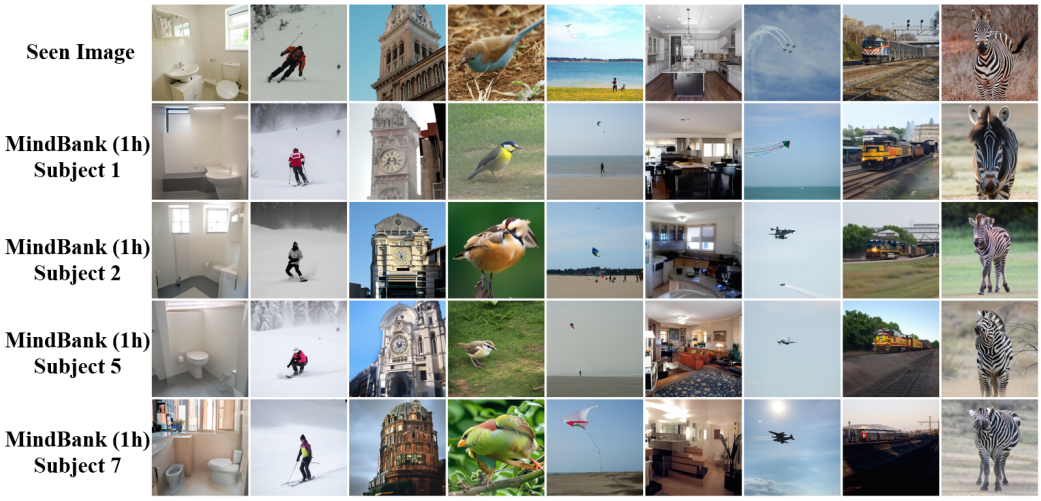



Figure 7: Cross-subject reconstruction examples produced by MindBank under the one-hour target-adaptation protocol. The top row shows the seen images, and the subsequent rows show reconstructions for Natural Scenes Dataset Subjects 1, 2, 5, and 7. Each column corresponds to the same stimulus image.