MINDREFINE: LANGUAGE-INFORMED SEMANTICS AND PERCEPTUAL REFINEMENT FOR CROSS-SUBJECT BRAIN VISUAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject functional magnetic resonance imaging (fMRI) visual decoding remains challenging when only limited paired data are available for a target subject. Existing methods mainly align brain responses with pretrained visual representations, leaving semantic structure only implicitly constrained, and rely on reconstruction pipelines that lack sample-specific perceptual correction. We propose MindRefine, a framework that improves limited-data fMRI visual decoding through representation learning and reconstruction refinement. Language-informed semantic regularization uses structured descriptions generated by a vision-language model (VLM) exclusively from training images to provide gated multi-channel semantic supervision and semantic-neighborhood distillation. This introduces explicit semantic constraints without using language information at inference. Feature-space reconstruction refinement keeps the trained decoder frozen, estimates sample-specific perceptual and semantic targets from a paired training reference bank, and refines each reconstruction through stage-wise feature-space optimization. Extensive experiments on the Natural Scenes Dataset in the limited-data regime demonstrate that MindRefine improves both low-level reconstruction fidelity and high-level perceptual consistency while preserving competitive retrieval performance. Additional experiments under the full-data setting further confirm the effectiveness and robustness of the framework. The code is available at `https://anonymous.4open.science/r/MindRefine`.

1 INTRODUCTION

Visual decoding from functional magnetic resonance imaging (fMRI) aims to recover perceived visual content from brain activity. It supports research on human visual representations and non-invasive brain-computer interfaces (Doerig et al., 2025; Lévy et al., 2026). Recent advances include hierarchical semantic enhancement and viewpoint-aware generative reconstruction (Wei et al., 2026a; Zhang et al., 2026). Cross-subject alignment further enables knowledge transfer and reduces reliance on subject-specific models (Bao et al., 2025; Li et al., 2026; Liu et al., 2026). However, fMRI acquisition is costly, signals are noisy, and neural responses vary substantially across subjects. These challenges are amplified when only one session of paired fMRI-image data is available for target-subject adaptation, making reliable reconstruction particularly difficult (Scotti et al., 2024).

Recent fMRI decoding methods increasingly map brain responses into pretrained visual spaces, such as global CLIP embeddings or image-token sequences, before generative reconstruction (Gong et al., 2025; Xu et al., 2025; Wei et al., 2025). These representations provide rich visual and semantic information for cross-modal retrieval and image reconstruction (Gong et al., 2024). However, their semantic structure is learned mainly through implicit visual-token alignment. With few target-subject pairs, this alignment may not reliably capture object identity, scene context, visual attributes, and relations among entities (Dai et al., 2025). Consequently, the brain representations may remain semantically incomplete or unstable. Recent language-aligned fMRI models further suggest that linguistic representations can provide structured semantic priors (Wei et al., 2026b). This motivates explicit language-derived constraints during training without language information at inference.

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.

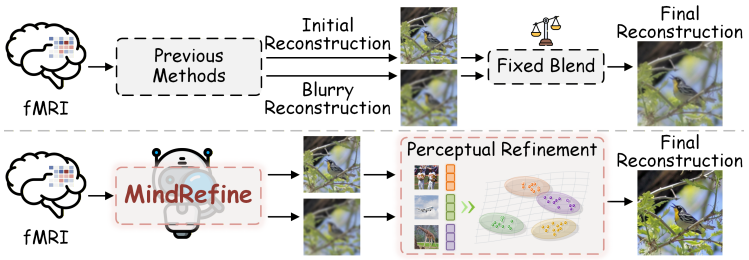

Figure 1: Previous methods use fixed image-space blending, which cannot adapt to sample-specific perceptual and semantic errors, whereas MindRefine performs sample-specific perceptual refinement using paired training references.

Improving brain-side representations alone does not guarantee optimal reconstruction. Visual decoders combine complementary representations. NeuroFlow learns a semantic latent space, while BrainGS integrates visual, semantic, and color information (Mai et al., 2026; Jing et al., 2026). High-level representations capture object identity and scene content but may introduce structural or semantic errors (Scotti et al., 2023), whereas low-level predictions better preserve color and spatial structure but are blurry and ambiguous (Scotti et al., 2024). Transferable decoders combine high-level perception with pixel-wise reconstruction (Liu et al., 2025). However, their conditioning or image-space combination rules are fixed across samples. As illustrated in Figure 1, such strategies do not estimate sample-specific perceptual and semantic feature targets, leaving errors in appearance, layout, or content. This motivates a reconstruction-side mechanism that estimates feature targets from paired training data and refines each output in perceptual and semantic feature spaces.

To this end, we propose MindRefine, a framework that improves limited-data cross-subject fMRI decoding from both the representation and reconstruction sides. Language-informed semantic regularization (LISR) uses a vision-language model (VLM) during training to generate structured descriptions exclusively for the training images. These descriptions cover caption, scene, object, attribute, and relation information. A sample-dependent gate converts the semantic channels into adaptive supervision, while semantic-neighborhood distillation preserves relations among training stimuli in the brain-side representation. Feature-space reconstruction refinement (FSR) improves perceptual reconstruction at inference, with the trained decoder frozen and a subject-specific reference bank constructed from paired training fMRI responses and images. FSR estimates sample-specific perceptual and semantic feature targets from the adapted brain-side representation of each test fMRI response. It then refines the initial reconstruction toward these targets through conservative stage-wise optimization. Our contributions are summarized as follows:

- Language-informed semantic regularization converts VLM-derived descriptions of training images into gated multi-channel semantic supervision and semantic-neighborhood distillation. It explicitly regularizes brain-side visual representations under limited target-subject data and uses no language information at inference.
- We propose feature-space reconstruction refinement to improve each test reconstruction using sample-specific perceptual and semantic targets estimated from a paired training reference bank. The method keeps the trained decoder frozen and optimizes the reconstruction toward these targets through stage-wise refinement.
- Extensive experiments on NSD under the limited-data setting demonstrate the effectiveness of MindRefine in improving decoding performance. Additional results under the full-data setting further confirm its robustness.

2 Related Work

Cross-Subject fMRI Visual Decoding. Recent fMRI decoders map brain responses into pretrained visual or generative spaces for reconstruction. CLIP and latent diffusion models provide semantic targets and strong image priors, improving generation over direct pixel regression (Radford et al., 2021; Rombach et al., 2022; Takagi & Nishimoto, 2023; Chen et al., 2023; Scotti et al., 2023). Cross-subject methods further transfer shared representations across individuals. MindEye2

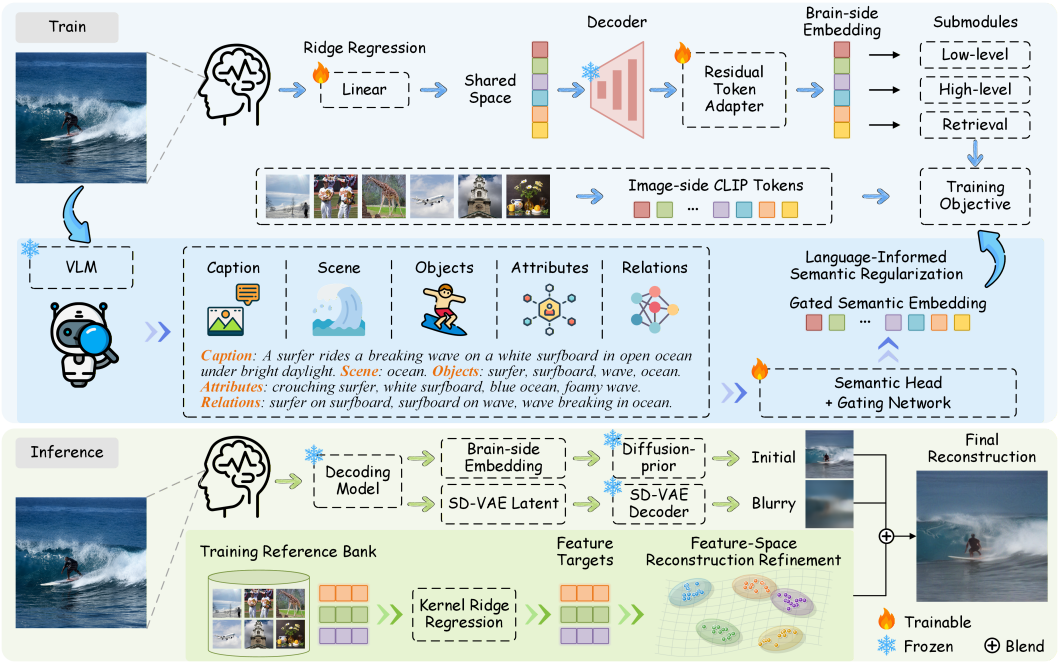

Figure 2: Overview of the proposed MindRefine framework. Language-informed semantic regularization aligns brain-side representations with VLM-generated structured semantics during training. Feature-space reconstruction refinement optimizes the initial reconstruction with respect to sample-specific feature targets estimated from the training reference bank at inference.

uses a shared-subject model with limited target adaptation (Scotti et al., 2024), while MindBridge aligns subject-specific fMRI distributions in a shared latent space (Wang et al., 2024). Subsequent studies explore adapters, subject-stimulus alignment, ROI-guided pretraining, source-free adaptation, and zero-shot or subject-agnostic decoding (Liu et al., 2025; Li et al., 2026; Liu et al., 2026; Xu et al., 2026; Wang et al., 2025a; Nan et al., 2026; Lu et al., 2026). Despite improved transferability, these methods mainly constrain brain representations through visual-token or latent-space alignment. MindRefine instead uses VLM-derived descriptions of training images to explicitly regularize brain-side representations in the limited-data setting.

Hierarchical Visual Priors for Reconstruction. High-quality reconstruction requires both semantic consistency and low-level fidelity. Existing methods combine CLIP or diffusion priors with low-level branches that preserve layout, color, and structure (Takagi & Nishimoto, 2023; Chen et al., 2023; Scotti et al., 2023; 2024). Recent methods model richer visual factors: Brain-IT captures interactions among voxel clusters (Beliy et al., 2026), NEURONS separates concept, scene, segmentation, and blurry reconstruction (Wang et al., 2025b), and SemVideo uses hierarchical semantic guidance (Yang et al., 2026). Other approaches decode interpretable concepts or structured object-centric semantics before generation (Cammarota et al., 2026; Huang et al., 2026). However, these methods mainly rely on fixed branches, predefined conditioning, or intermediate semantic predictions. In contrast, MindRefine estimates sample-specific perceptual and semantic targets from paired training references and refines reconstruction in feature space.

3 Method

3.1 Problem Formulation and Overview

We consider cross-subject fMRI visual decoding under a limited-data setting. Given a limited number of paired fMRI responses and images from a target subject, the goal is to reconstruct the perceived image from an unseen fMRI response. Let $\mathcal{D}_{\mathrm{tr}}^{s} = \{(x_i^s, I_i)\}_{i=1}^{N_s}$ denote the training set of

subject s, where x_i^s is the preprocessed fMRI response and I_i is the visual stimulus. The model receives an fMRI response x^s and predicts a reconstruction $\hat{I}$ during inference.

As illustrated in Figure 2, MindRefine consists of training-stage semantic regularization and inference-stage reconstruction refinement. During training, VLM-derived constraints regularize adapted brain-side visual tokens; during inference, paired training references provide sample-specific targets to refine the initial reconstruction in perceptual and semantic spaces.

3.2 Language-Informed Semantic Regularization

Limited fMRI data provide sparse coverage of natural-image semantics, making image-token supervision alone insufficient to capture stable semantic structure. We introduce language-informed semantic regularization to use structured VLM descriptions of training images as auxiliary supervision for brain-side visual representations. The semantic head and gating network are used only to construct training-time semantic objectives.

A structured annotation $a_i = (c_i, s_i, o_i, u_i, r_i)$ is generated for each training image I_i using a VLM, where c_i is a global caption, s_i denotes scene information, o_i denotes object descriptions, u_i denotes visual attributes, and r_i denotes relations among entities. These fields are converted into channel-level text embeddings $T_i = (t_i^1, \ldots, t_i^K)$, where $t_i^k \in \mathbb{R}^{d_t}$. In our implementation, $K = 5$ corresponds to caption, scene, object, attribute, and relation channels. Multiple entries within a field are concatenated using a fixed textual template and encoded into one channel-level embedding.

The decoding model maps fMRI response x_i^s to brain-side visual tokens, followed by a residual token adapter that produces $Z_i \in \mathbb{R}^{L\times d}$. The corresponding image-side CLIP tokens $V_i \in \mathbb{R}^{L\times d}$ serve as targets in the shared visual-token space. We mean-pool Z_i as $m_i = \mathrm{Pool}(Z_i)$, then use a lightweight semantic head $h_\theta(\cdot)$ to project m_i into the text-embedding space. A gating network $\psi_\theta(\cdot)$ predicts channel importance and forms the gated semantic embedding:

$$\pi_i = \mathrm{softmax}(\psi_\theta(m_i)), \quad \bar{t}_i = \mathrm{Norm}\left(\sum_{k=1}^{K} \pi_i^k t_i^k\right), \tag{1}$$

where $\pi_i \in \mathbb{R}^K$. The resulting gate allows different images to rely on different semantic cues.

The semantic projection of the pooled brain representation is aligned with the structured semantic channels to provide direct semantic supervision:

$$\mathcal{L}_{\mathrm{sem}} = \frac{1}{B}\sum_{i\in\mathcal{B}}\sum_{k=1}^{K} \pi_i^k \left[1 - \cos\left(h_\theta(m_i), t_i^k\right)\right], \tag{2}$$

where $\mathcal{B}$ denotes the local mini-batch and $B = |\mathcal{B}|$. To discourage degenerate channel selection, the gate is further regularized by $\mathcal{L}_{\mathrm{gate}} = B^{-1}\sum_{i\in\mathcal{B}} \mathrm{KL}(\pi_i \| p_0)$. The prior follows the channel order caption, scene, object, attribute, and relation and is set to $p_0 = (0.30, 0.25, 0.20, 0.15, 0.10)$. It is weakly imposed to stabilize the gate while retaining sample-dependent channel predictions.

Direct semantic alignment does not explicitly preserve semantic neighborhoods among stimuli. We therefore obtain normalized flattened representations $z_i = \mathrm{Norm}(\mathrm{vec}(Z_i))$ and $v_i = \mathrm{Norm}(\mathrm{vec}(V_i))$. Candidate similarities are computed over the global mini-batch $\mathcal{G}$ gathered across devices. For each local anchor $i \in \mathcal{B}$, its paired instance is excluded, yielding the candidate set $\mathcal{C}_i = \mathcal{G} \setminus \{i\}$. For representations a and b, the candidate distribution over $j \in \mathcal{C}_i$ is

$$\mathcal{S}_{ij}(a, b; \tau) = \frac{\exp\left(\cos(a_i, b_j)/\tau\right)}{\sum_{l\in\mathcal{C}_i} \exp\left(\cos(a_i, b_l)/\tau\right)}. \tag{3}$$

The detached semantic target distribution is $q_{ij} = \mathcal{S}_{ij}(\bar{t}, \bar{t}; \tau_{\mathrm{sem}})$, whereas the model-induced visual distribution is $p_{ij} = \mathcal{S}_{ij}(z, v; \tau_{\mathrm{vis}})$. The resulting soft semantic distillation loss is

$$\mathcal{L}_{\mathrm{soft}} = -\frac{1}{B}\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{C}_i} \mathrm{sg}(q_{ij}) \log p_{ij}, \tag{4}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Excluding the paired sample from $\mathcal{C}_i$ focuses the loss on preserving semantic neighborhoods rather than instance-level CLIP alignment. The complete semantic regularization is $\mathcal{L}_{\mathrm{LISR}} = \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{soft}}\mathcal{L}_{\mathrm{soft}} + \lambda_{\mathrm{gate}}\mathcal{L}_{\mathrm{gate}}$.

3.3 FEATURE-SPACE RECONSTRUCTION REFINEMENT

Fixed image-space blending applies the same rule to all samples and cannot optimize sample-specific perceptual or semantic alignment. We therefore introduce feature-space reconstruction refinement, a post-hoc module that keeps the trained decoder frozen. FSR estimates target feature representations from a paired training reference bank and refines the pre-refinement reconstruction in feature space while preserving consistency with its input.

The frozen decoding pipeline first produces fixed-fusion and training-calibrated adaptive-fusion reconstructions, which are combined to form the pre-refinement reconstruction $\hat{I}_i^{(0)}$. A lightweight low-rank token adapter is fitted on the training samples and then frozen. Applying this adapter to the brain-side visual tokens yields the normalized embedding z_i used for feature-target estimation. Based on the training samples, we construct the subject-specific reference bank:

$$\mathcal{R}_{\mathrm{tr}}^{s} = \left\{ \left(z_j, \{ f_j^m \}_{m=1}^{M} \right) \right\}_{j=1}^{N_s}, \tag{5}$$

where z_j is the normalized embedding obtained with the frozen low-rank token adapter for training sample j, and $f_j^m = \Phi_m(I_j)$ is the feature extracted from its paired image. The frozen extractor $\Phi_m(\cdot)$ includes encoder-specific preprocessing, feature extraction, optional pooling or flattening, and L_2 normalization. All reference features are extracted exclusively from paired training images.

For each feature space m, we estimate a target representation using kernel ridge regression over the frozen brain-side embeddings. Let $F_{\mathrm{tr}}^m = [f_1^m, \ldots, f_{N_s}^m]^\top$ denote the stacked training features, with $[\mathbf{K}_{\mathrm{tr}}]_{jl} = k(z_j, z_l) = z_j^\top z_l$ defining the kernel matrix and $\mathbf{k}_i = [k(z_i, z_1), \ldots, k(z_i, z_{N_s})]^\top$ the kernel vector for test sample i. The L_2 normalization of the brain embeddings makes the linear kernel equivalent to cosine similarity. The resulting feature target is

$$\tilde{f}_i^m = \mathrm{Norm}\left[\mathbf{k}_i^\top \left(\mathbf{K}_{\mathrm{tr}} + \lambda_{\mathrm{KRR}} \mathbf{I}_{N_s} \right)^{-1} F_{\mathrm{tr}}^m \right], \tag{6}$$

where λ_{KRR} is the ridge coefficient and $\mathbf{I}_{N_s}$ is the $N_s \times N_s$ identity matrix.

The refinement is applied stage-wise. Let r denote the refinement stage, with input reconstruction $\hat{I}_i^{(r-1)}$. The first stage starts from the pre-refinement reconstruction $\hat{I}_i^{(0)}$, and later stages use the blended output of the preceding stage. At stage r, an unconstrained image parameter u is initialized as $u_i^{(r,0)} = \mathrm{logit}(\mathrm{Clip}_\epsilon(\hat{I}_i^{(r-1)}))$, where $\mathrm{Clip}_\epsilon(\cdot)$ restricts pixel values to $[\epsilon, 1-\epsilon]$. For the stage-specific feature set $\mathcal{M}_r$, the refinement objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{ref}}^{(r)}(u) = \lambda_{\mathrm{base}}^{(r)} \, \mathrm{MSE}\left(\sigma(u), \hat{I}_i^{(r-1)} \right) + \lambda_{\mathrm{tv}}^{(r)} \, \mathrm{TV}\left(\sigma(u) \right) \\ + \sum_{m \in \mathcal{M}_r} \lambda_m^{(r)} \left[1 - \cos\left(\Phi_m(\sigma(u)), \tilde{f}_i^m \right) \right], \end{aligned} \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function. The MSE term anchors the optimized image to the stage input, the total-variation term suppresses high-frequency artifacts, and the feature terms move the image toward the training-derived targets estimated from the reference bank.

Let $u_i^{(r)\star} = \arg\min_u \mathcal{L}_{\mathrm{ref}}^{(r)}(u)$ and $\hat{I}_{i,\mathrm{ref}}^{(r)} = \sigma(u_i^{(r)\star})$. The stage output is conservatively blended as:

$$\hat{I}_i^{(r)} = \mathrm{Clip}\left[(1 - \alpha_r) \hat{I}_i^{(r-1)} + \alpha_r \hat{I}_{i,\mathrm{ref}}^{(r)} \right], \tag{8}$$

where $\alpha_r \in [0, 1]$ is the stage-specific blending coefficient and $\mathrm{Clip}(\cdot)$ restricts pixel values to $[0, 1]$. This formulation ensures that each stage refines its current input rather than repeatedly blending with the original reconstruction.

The first stage refines the pre-refinement reconstruction using a compact feature set to correct major perceptual deviations, and the second stage further refines the first-stage blended output using an expanded feature set. The final reconstruction is $\hat{I}_i^{\mathrm{out}} = \hat{I}_i^{(2)}$.

3.4 TRAINING OBJECTIVE

MindRefine is trained by combining the base reconstruction objective with language-informed semantic regularization and auxiliary stability terms. For each paired sample (x_i^s, I_i), the decoding

model predicts adapted brain-side visual tokens Z_i and generation-branch representations, with the image-side CLIP tokens denoted by V_i.

The base reconstruction objective (Scotti et al., 2024) combines retrieval-oriented visual-token alignment, diffusion-prior supervision, and low-level blurry reconstruction:

$$\mathcal{L}_{\text{base}}^{(e)} = \lambda_{\text{align}}\mathcal{L}_{\text{align}}^{(e)} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}} + \lambda_{\text{low}}\mathcal{L}_{\text{low}}, \tag{9}$$

where $\mathcal{L}_{\text{align}}^{(e)}$ follows the BiMixCo-to-SoftCLIP schedule, $\mathcal{L}_{\text{prior}}$ is the diffusion-prior loss, and $\mathcal{L}_{\text{low}}$ supervises the blurry reconstruction branch. The residual token adapter is regularized by

$$\mathcal{L}_{\text{adapt}} = \frac{1}{|\mathcal{D}_\Delta|}\sum_{\Delta\in\mathcal{D}_\Delta}\frac{1}{|\Delta|}\|\Delta\|_F^2, \tag{10}$$

where $\mathcal{D}_\Delta$ contains the residual updates produced for the backbone and CLIP-token representations in the current mini-batch. We also use a training-only stimulus-structure objective $\mathcal{L}_{\text{stim}}$ to encourage instance discrimination, prototype alignment, relation consistency, and perturbation robustness within the training set (Li et al., 2026). The complete training objective at epoch e is

$$\mathcal{L}_{\text{train}}^{(e)} = \mathcal{L}_{\text{base}}^{(e)} + \mathcal{L}_{\text{LISR}} + \lambda_{\text{adapt}}\mathcal{L}_{\text{adapt}} + \gamma_e\mathcal{L}_{\text{stim}}. \tag{11}$$

During the BiMixCo phase, $\mathcal{L}_{\text{LISR}} = \mathcal{L}_{\text{adapt}} = 0$; both regularization terms are enabled afterward, while γ_e gradually increases from zero.

4 Experiment

4.1 Experimental Setup

Dataset. We evaluate MindRefine on the Natural Scenes Dataset (NSD) (Allen et al., 2022), a large-scale 7T fMRI dataset widely used for visual brain decoding. NSD records brain responses while subjects viewed natural images sampled from MSCOCO (Lin et al., 2014). Following prior cross-subject decoding settings, we use subjects 1, 2, 5, and 7 for evaluation (Scotti et al., 2024).

Implementation Details. MindRefine is implemented in PyTorch and trained for 150 epochs on four NVIDIA TITAN RTX GPUs. Qwen3-VL-Plus (Bai et al., 2025) is used as the VLM backbone. All hyperparameters are fixed across subjects. For $\mathcal{L}_{\text{LISR}}$, we set $[\lambda_{\text{sem}}, \lambda_{\text{soft}}, \lambda_{\text{gate}}] = [0.015, 0.008, 0.001]$, and the residual-adapter weight is set to $\lambda_{\text{adapt}} = 0.001$. We set the kernel ridge coefficient to $\lambda_{\text{KRR}} = 0.05$ and use the same blending coefficient $\alpha_r = 0.10$ for both refinement stages. The ramp coefficient γ_e is activated after the BiMixCo phase and increases linearly from 0 to 1 over 45 epochs. The one-session setting is trained with batch size 4 for approximately 1.5 hours, while the full-data setting is trained with batch size 20 for approximately 29 hours.

Metrics. We report PixCorr, SSIM, AlexNet(2), and AlexNet(5) (Krizhevsky et al., 2012) for low-level reconstruction quality, and InceptionV3 (Szegedy et al., 2016), CLIP (Radford et al., 2021), EfficientNet-B (Tan & Le, 2019), and SwAV (Caron et al., 2020) for high-level consistency. Image and Brain top-1 retrieval use 300 candidates and 30 repeats (Scotti et al., 2024).

4.2 Experimental Results

We compare MindRefine with representative fMRI-to-image decoding methods under the 1-session setting, including MindEye2 (Scotti et al., 2024), MindAligner (Dai et al., 2025), MindTuner (Gong et al., 2025), Duala (Li et al., 2026), and NeuroFlow (Mai et al., 2026).

As reported in Table 1, MindRefine establishes an overall advantage in reconstruction quality across low-level and high-level evaluations. The gains remain consistent across different subjects, demonstrating improvements in structural fidelity, perceptual quality, and semantic consistency. This pattern supports the complementary roles of LISR in representation learning and FSR in reconstruction optimization. Its competitive retrieval performance confirms that the improved visual quality preserves brain-image alignment. Figure 3 presents qualitative comparisons. MindRefine more consistently preserves the global composition, object identity, and spatial structure of the perceived images compared with baseline methods. Its reconstructions also exhibit fewer semantic deviations and structural distortions, while retaining clearer boundaries and more faithful visual details.

Table 1: Quantitative comparison on Subjects 1, 2, 5, and 7 under the 1-session setting, with average results computed over the four subjects. The **best** and second-best results are highlighted. ↑ and ↓ indicate that higher and lower values are better, respectively.

Method	Subject	Low-Level				High-Level				Retrieval	
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
MindEye2 (Scotti et al., 2024)	Avg	0.195	0.419	84.2%	90.6%	81.2%	79.3%	0.810	0.469	79.0%	57.4%
MindAligner (Dai et al., 2025)	Avg	0.206	0.414	85.6%	91.6%	83.0%	81.2%	0.802	0.463	78.9%	75.3%
MindTuner (Gong et al., 2025)	Avg	0.224	0.420	87.8%	93.6%	84.8%	83.5%	0.781	0.440	83.0%	76.0%
Duala (Li et al., 2026)	Avg	0.230	0.416	87.9%	93.5%	85.4%	83.5%	0.781	0.445	**84.5%**	81.1%
NeuroFlow (Mai et al., 2026)	Avg	0.209	**0.425**	83.3%	88.6%	82.6%	79.8%	0.816	0.489	81.5%	–
MindRefine (Ours)	Avg	**0.239**	0.419	**88.3%**	**94.1%**	**87.8%**	**83.6%**	**0.758**	**0.416**	84.2%	**81.8%**
MindEye2 (Scotti et al., 2024)	Subj1	0.235	0.428	88.0%	93.3%	83.6%	80.8%	0.798	0.459	94.0%	77.6%
MindAligner (Dai et al., 2025)	Subj1	0.226	0.415	88.2%	93.3%	83.5%	81.8%	0.800	0.459	90.9%	86.9%
MindTuner (Gong et al., 2025)	Subj1	0.262	0.422	90.6%	94.9%	85.8%	84.6%	0.774	0.433	94.2%	87.4%
Duala (Li et al., 2026)	Subj1	0.253	0.418	90.8%	95.3%	86.6%	85.1%	0.768	0.433	**94.8%**	91.2%
NeuroFlow (Mai et al., 2026)	Subj1	0.233	**0.437**	87.1%	90.1%	82.9%	79.2%	0.811	0.490	89.8%	–
MindRefine (Ours)	Subj1	**0.276**	0.425	**91.3%**	**95.7%**	**89.1%**	**85.3%**	**0.742**	**0.401**	93.8%	**92.3%**
MindEye2 (Scotti et al., 2024)	Subj2	0.200	**0.433**	85.0%	92.1%	81.9%	79.4%	0.807	0.467	90.5%	67.2%
MindAligner (Dai et al., 2025)	Subj2	0.218	0.426	88.1%	93.3%	84.1%	82.5%	0.791	0.452	90.0%	85.6%
MindTuner (Gong et al., 2025)	Subj2	0.225	0.425	89.1%	95.1%	84.8%	83.7%	0.781	0.440	93.7%	82.8%
Duala (Li et al., 2026)	Subj2	0.236	0.428	89.2%	94.4%	85.4%	83.7%	0.778	0.443	**94.6%**	91.5%
NeuroFlow (Mai et al., 2026)	Subj2	0.219	0.424	86.2%	89.8%	83.8%	78.2%	0.812	0.490	88.1%	–
MindRefine (Ours)	Subj2	**0.250**	0.428	**90.1%**	**95.8%**	**88.9%**	**84.5%**	**0.745**	**0.406**	93.3%	**92.4%**
MindEye2 (Scotti et al., 2024)	Subj5	0.175	0.405	83.1%	91.0%	84.3%	82.5%	0.781	0.444	66.9%	47.0%
MindAligner (Dai et al., 2025)	Subj5	0.197	0.409	84.7%	91.6%	84.6%	82.8%	0.784	0.454	70.6%	66.0%
MindTuner (Gong et al., 2025)	Subj5	0.208	0.415	86.8%	**93.7%**	87.7%	**85.9%**	**0.750**	0.422	72.2%	68.1%
Duala (Li et al., 2026)	Subj5	**0.214**	0.413	86.5%	93.1%	86.4%	84.5%	0.772	0.439	76.6%	**71.2%**
NeuroFlow (Mai et al., 2026)	Subj5	0.195	0.414	81.3%	88.9%	83.8%	82.9%	0.804	0.473	**88.1%**	–
MindRefine (Ours)	Subj5	**0.214**	**0.416**	**86.9%**	93.5%	**87.9%**	84.9%	0.755	**0.416**	75.4%	70.5%
MindEye2 (Scotti et al., 2024)	Subj7	0.170	0.408	80.7%	85.9%	74.9%	74.3%	0.854	0.504	64.4%	37.8%
MindAligner (Dai et al., 2025)	Subj7	0.183	0.407	81.5%	88.3%	79.9%	77.8%	0.834	0.487	64.2%	62.6%
MindTuner (Gong et al., 2025)	Subj7	0.202	0.417	84.5%	90.8%	80.7%	79.6%	0.817	0.465	71.9%	65.5%
Duala (Li et al., 2026)	Subj7	0.215	0.405	**85.2%**	91.2%	83.3%	**80.7%**	0.805	0.463	71.9%	70.5%
NeuroFlow (Mai et al., 2026)	Subj7	0.189	**0.425**	78.7%	85.6%	80.0%	78.8%	0.836	0.502	59.8%	–
MindRefine (Ours)	Subj7	**0.217**	0.408	84.9%	**91.5%**	**85.2%**	79.8%	**0.788**	**0.439**	**74.4%**	**72.0%**



Figure 3: Qualitative comparison of image reconstructions under the 1-session setting.

4.3 Ablation Study

We evaluate four ablated variants to assess the contribution of each component. The w/o LISR and w/o $\mathcal{L}_{\text{stim}}$ variants remove the corresponding training-time regularizers. The w/o FSR variant disables feature-space reconstruction refinement and directly evaluates the pre-refinement fused output. CLIP-only Refinement retains the refinement pipeline but limits the feature constraint to the CLIP space. As shown in Table 2, MindRefine achieves the strongest overall reconstruction per-

Table 2: Average ablation results under the 1-session setting.

Ablation	Low-Level				High-Level				Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
w/o LISR	0.229	0.413	87.5%	93.9%	86.8%	**83.9%**	0.765	0.418	83.4%	**82.4%**
w/o $\mathcal{L}_{\text{stim}}$	0.232	0.409	87.7%	93.9%	87.2%	83.1%	0.764	0.417	83.9%	82.2%
w/o FSR	0.228	0.407	87.2%	93.6%	87.2%	81.4%	0.797	0.468	84.0%	81.6%
CLIP-only Refinement	0.230	0.411	87.5%	93.7%	87.3%	83.2%	0.793	0.462	84.0%	81.3%
MindRefine (Ours)	**0.239**	**0.419**	**88.3%**	**94.1%**	**87.8%**	83.6%	**0.758**	**0.416**	**84.2%**	81.8%

formance across the compared variants. Removing LISR or $\mathcal{L}_{\text{stim}}$ reduces low-level and high-level performance, indicating that training-time semantic and stimulus-structure regularization improves the learned brain-side representation. Disabling FSR or using only CLIP features for refinement also degrades reconstruction quality, showing that FSR benefits from complementary perceptual and semantic feature spaces. Retrieval performance remains relatively stable across variants, consistent with the reconstruction-oriented design of the proposed components.

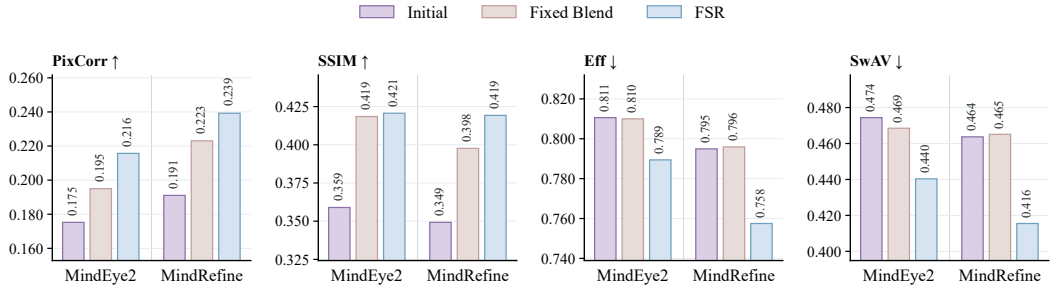

Figure 4: Average initial, fixed-blend, and FSR reconstructions for MindEye2 and MindRefine.

4.4 Analysis of Perceptual Refinement

We analyze Feature-Space Reconstruction Refinement by comparing three variants: the initial reconstruction, fixed image-space blending, and the final FSR output. Figure 4 reports averaged results on four metrics. FSR improves over both initial reconstruction and fixed blending, with consistent gains in low-level similarity and distance-based perceptual metrics. Similar improvements on MindEye2 suggest that FSR can transfer to another decoder. These results support our motivation that fixed image-space blending cannot fully adapt to sample-specific feature targets, whereas FSR provides refinement through training-reference-guided feature-space optimization.

Table 3: Average quantitative comparison under the 40-session setting.

Method	Low-Level				High-Level				Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
MindEye2 (Scotti et al., 2024)	0.322	**0.431**	**96.1%**	98.6%	95.4%	93.0%	0.619	0.344	99.2%	99.0%
NeuroFlow (Mai et al., 2026)	0.274	0.266	94.4%	98.0%	95.6%	**93.9%**	0.669	0.380	90.3%	–
MindRefine (Ours)	**0.341**	0.430	**96.1%**	**99.0%**	**96.4%**	93.4%	**0.580**	**0.313**	**99.4%**	**99.2%**

4.5 Results under the 40-Session Setting

We further evaluate MindRefine under the 40-session setting against MindEye2 (Scotti et al., 2024) and NeuroFlow (Mai et al., 2026). As shown in Table 3, MindRefine achieves the best or tied-best performance on most reconstruction and retrieval metrics, showing that LISR and FSR remain effective beyond the limited-data setting. Figure 5 further shows that MindRefine better preserves object identity, global layout, and visual structure, consistent with the quantitative results.

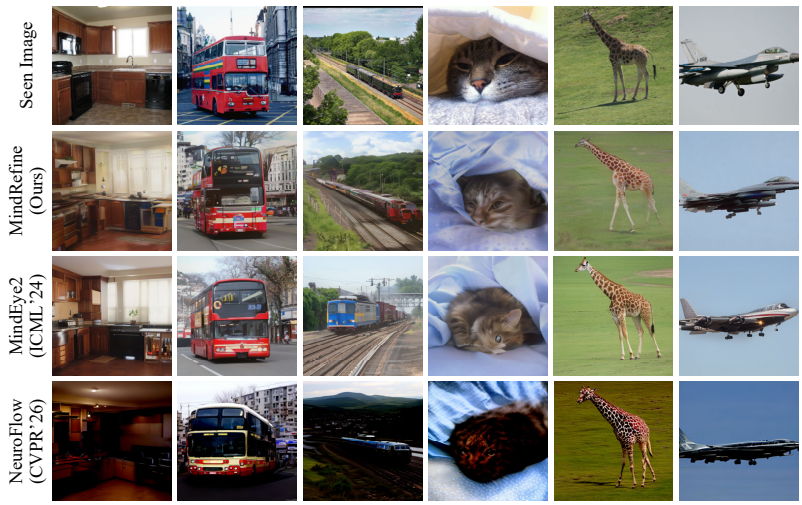

Figure 5: Qualitative comparison of image reconstructions under the 40-session setting.

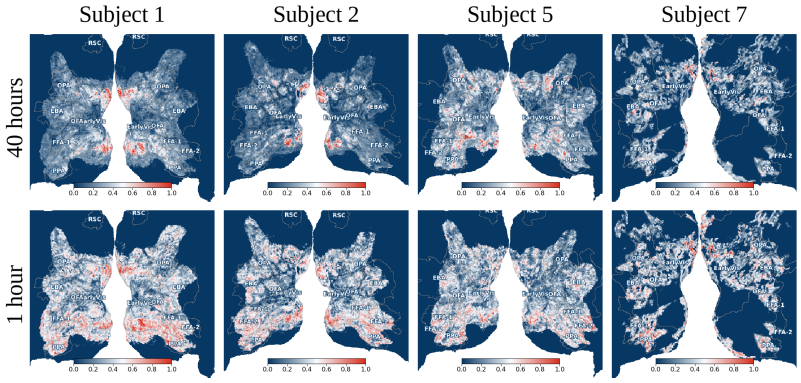

Figure 6: Voxel-wise ridge relevance under 1-session and 40-session settings.

4.6 Neuroscience Interpretability of MindRefine

Figure 6 visualizes voxel-wise ridge relevance on the NSD cortical flat map under the 40-session and 1-session settings (Allen et al., 2022). The 40-session model shows structured relevance over canonical visual regions, including EarlyVis, OFA, OPA, EBA, FFA, PPA, and RSC (Contier et al., 2024). With only one session, MindRefine preserves cortical organization, although with lower regional contrast. These patterns indicate that MindRefine preserves interpretable cortical organization under limited fMRI data.

5 Conclusion

We presented MindRefine, a cross-subject fMRI visual decoding framework that combines language-informed semantic regularization with feature-space reconstruction refinement. MindRefine leverages VLM-derived descriptions of training images to regularize brain-side representations and estimates sample-specific feature targets from paired training references to refine reconstructions with the decoder frozen. Experiments on NSD demonstrate improved reconstruction quality under the limited-data setting, while results under the full-data setting further validate the effectiveness and robustness of the proposed design.

Limitations and Future Work. MindRefine still requires paired target-subject fMRI data, and its semantic regularization depends on VLM annotation quality. Future work will improve annotation control and extend the framework to broader subjects and dynamic stimuli.

AI USE STATEMENT

Generative AI tools were used in this work for research support and manuscript preparation. As part of the proposed method, Qwen3-VL-Plus was used to generate semantic annotations from training images, including captions, scenes, objects, attributes, and relations. These annotations were used exclusively as training-time semantic supervision and were not used during inference. General-purpose large language models were also used to provide feedback on research methodology and experimental design, assist with the interpretation and presentation of experimental results, and support manuscript drafting and revision. In addition, generative AI tools were used for literature search and summarization, language editing, and improving the organization of the paper.

Generative AI tools were not used to formulate or prove mathematical claims, clean or reformat the experimental dataset, or perform qualitative or thematic data analysis; survey- and interview-related tasks are not applicable. All AI-assisted suggestions and generated content were reviewed by the authors before inclusion. All experimental results and quantitative values were produced by the implemented experimental pipeline rather than generated by AI. The authors verified the final manuscript, method descriptions, analyses, and reported results, and take full responsibility for the paper, including text, claims, and artifacts produced with the assistance of generative AI.

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.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14194–14202, 2025.

Roman Beliy, Amit Zalcher, Jonathan Kogman, navve wasserman, and michal Irani. Brain-IT: Image reconstruction from fMRI via brain-interaction transformer. In *The Fourteenth International Conference on Learning Representations*, 2026.

Sara Cammarota, Matteo Ferrante, and Nicola Toschi. Bridging brains and concepts: Interpretable visual decoding from fmri with semantic bottlenecks. *Advances in Neural Information Processing Systems*, 38:86820–86850, 2026.

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. Curran Associates, Inc., 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 (CVPR)*, pp. 22710–22720, June 2023.

Oliver Contier, Chris I Baker, and Martin N Hebart. Distributed representations of behaviour-derived object dimensions in the human visual system. *Nature Human Behaviour*, 8(11):2179–2193, 2024.

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, pp. 12214–12228. PMLR, 2025.

Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7(8):1220–1234, 2025.

Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, Ke Liu, and Yu Zhang. Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 51655–51683. Curran Associates, Inc., 2024.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14247–14255, 2025.

Zheng Huang, Enpei Zhang, Weikang Qiu, Yinghao Cai, Carl Yang, Elynn Chen, Xiang Zhang, Rex Ying, Dawei Zhou, and Yujun Yan. Seeing through the brain: New insights from decoding visual stimuli with fMRI. In *The Fourteenth International Conference on Learning Representations*, 2026.

Haodong Jing, Dongyao Jiang, Jixin Wang, Junhao Jia, Yanshu Li, Yongqiang Ma, and Nanning Zheng. More natural, more real: Object-aware gaussian splatting for 3d visual decoding from human brain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19033–19044, June 2026.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Jarod Lévy, Mingfang Zhang, Svetlana Pinet, Jérémy Rapin, Hubert Banville, Stéphane d'Ascoli, and Jean-Rémi King. Noninvasive decoding of typed sentences from human brain activity. *Nature Neuroscience*, pp. 1–7, 2026.

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, June 2026.

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 *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Yulong Liu, Yongqiang Ma, Guibo Zhu, Haodong Jing, and Nanning Zheng. See through their minds: Learning transferable brain decoding models from cross-subject fMRI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5730–5738, 2025.

Yulong Liu, Hua Xu, Yiyang Cai, Chunyang Jiang, Sirui Han, and Yike Guo. Modeling the brain's grammar: Roi-guided fmri pretraining for transferable and interpretable vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6900–6909, June 2026.

Jingyu Lu, Haonan Wang, Qixiang ZHANG, and Xiaomeng Li. A cognitive process-inspired architecture for subject-agnostic brain visual decoding. In *The Fourteenth International Conference on Learning Representations*, 2026.

Weijian Mai, Mu Nan, Yu Zhu, Jiahang Cao, Rui Zhang, Yuqin Dai, Chunfeng Song, Andrew Luo, and Jiamin Wu. Neuroflow: Toward unified visual encoding and decoding from neural activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12966–12976, 2026.

Mu Nan, Muquan Yu, Weijian Mai, Jacob S. Prince, Hossein Adeli, Rui Zhang, Jiahang Cao, Benjamin Becker, John A. Pyles, Margaret M. Henderson, Chunfeng Song, Nikolaus Kriegeskorte, Michael J. Tarr, Xiaoqing Hu, and Andrew F. Luo. Meta-learning in-context enables training-free cross subject brain decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3604–3616, June 2026.

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, pp. 8748–8763. PMLR, 18–24 Jul 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 (CVPR)*, pp. 10684–10695, June 2022.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, ethan cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth 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. Curran Associates, Inc., 2023.

Paul S Scotti, Mihir Tripathy, Cesare 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. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 44038–44059, 2024.

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 (CVPR)*, June 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 (CVPR)*, pp. 14453–14463, June 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, pp. 6105–6114. PMLR, 09–15 Jun 2019.

Haonan Wang, Jingyu Lu, Hongrui Li, and Xiaomeng Li. Zebra: Towards zero-shot cross-subject generalization for universal brain visual decoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 1030–1051. Curran Associates, Inc., 2025a.

Haonan Wang, Qixiang Zhang, Lehan Wang, Xuanqi Huang, and Xiaomeng Li. Neurons: Emulating the human visual cortex improves fidelity and interpretability in fmri-to-video reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18367–18376, October 2025b.

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 (CVPR)*, pp. 11333–11342, June 2024.

Yujie Wei, Chenglong Ma, Jianxiong Gao, Chenhui Wang, Shiwei Zhang, Biao Gong, Shuai Tan, Hangjie Yuan, and Hongming Shan. Bridging brain and semantics: A hierarchical framework for semantically enhanced fmri-to-video reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28211–28223, June 2026a.

Yuxiang Wei, Yanteng Zhang, Xi Xiao, Tianyang Wang, Xiao Wang, and Vince D. Calhoun. Morebrain: Routed mixture of experts for interpretable and generalizable cross-subject fmri visual decoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 55870–55902. Curran Associates, Inc., 2025.

Yuxiang Wei, Yanteng Zhang, Xi Xiao, Chengxuan Qian, Tianyang Wang, and Vince D. Calhoun. fmri-lm: Towards a universal foundation model for language-aligned fmri understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6931–6940, June 2026b.

Ganxi Xu, Jinyi Long, and Jia Zhang. Probability distribution alignment and low-rank weight decomposition for source-free domain adaptive brain decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27233–27241, 2026.

Yangyang Xu, Bangzhen Liu, Wenqi Shao, Yong Du, Shengfeng He, and Tingting Zhu. Cross-subject mind decoding from inaccurate representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15066–15075, October 2025.

Minghan Yang, Lan Yang, Ke Li, Honggang Zhang, Kaiyue Pang, and Yi-Zhe Song. Semvideo: Reconstructs what you watch from brain activity via hierarchical semantic guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13658–13669, June 2026.

Xun Zhang, Weihao Xia, Yulong Liu, Bo Yang, Alessandro Bozzon, and Pan Wang. Single-stage fmri-to-3d reconstruction via viewpoint-aware embedding and hierarchical guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18037–18045, 2026.

A Detailed Component Losses

This appendix expands the component losses summarized in the training objective. For each paired sample (x_i^s, I_i), the decoding model predicts adapted brain-side visual tokens Z_i and generation-branch representations, and the corresponding image-side CLIP tokens are denoted by V_i.

Retrieval-oriented alignment. The alignment objective follows the base decoder schedule (Scotti et al., 2024). Let e_{mix} denote the end of the BiMixCo phase, and let $\mathbf{Z}_{\mathcal{B}}$ and $\mathbf{V}_{\mathcal{B}}$ be the ordered brain-side and image-side token batches. The alignment loss is

$$\mathcal{L}_{\mathrm{align}}^{(e)} = \begin{cases} \mathcal{L}_{\mathrm{BiMixCo}}\left(\mathbf{Z}_{\mathcal{B}}, \mathbf{V}_{\mathcal{B}}\right), & e < e_{\mathrm{mix}}, \\ \mathcal{L}_{\mathrm{SoftCLIP}}\left(\mathbf{Z}_{\mathcal{B}}, \mathbf{V}_{\mathcal{B}}\right), & e \geq e_{\mathrm{mix}}. \end{cases} \tag{12}$$

The BiMixCo objective uses mixed fMRI responses and mixed target assignments during the early phase, whereas SoftCLIP subsequently matches the brain-to-image similarity distribution to the image-to-image similarity distribution.

High-level generative supervision. The diffusion-prior branch is supervised by

$$\mathcal{L}_{\mathrm{prior}} = \mathcal{L}_{\mathrm{DP}}(H_i, V_i), \tag{13}$$

where H_i is the brain-side conditioning representation and $\mathcal{L}_{\mathrm{DP}}$ follows the base diffusion-prior loss (Scotti et al., 2024).

Low-level blurry reconstruction. The blurry reconstruction branch predicts a latent code ℓ_i in the Stable Diffusion variational autoencoder space, while the paired training image provides the target latent ℓ_i^*. The low-level objective is

$$\mathcal{L}_{\mathrm{low}} = \mathrm{MAE}(\ell_i, \ell_i^*) + \beta_{\mathrm{cont}}\mathcal{L}_{\mathrm{cont}}, \tag{14}$$

where $\mathrm{MAE}(\cdot,\cdot)$ denotes element-wise mean absolute error, and $\mathcal{L}_{\mathrm{cont}}$ aligns blurry-branch transformer features with image features under augmentation.

Residual-adapter regularization. The residual token adapter is regularized by

$$\mathcal{L}_{\mathrm{adapt}} = \frac{1}{|\mathcal{D}_\Delta|} \sum_{\Delta \in \mathcal{D}_\Delta} \frac{1}{|\Delta|} \|\Delta\|_F^2, \tag{15}$$

where $\mathcal{D}_\Delta$ contains the residual updates produced for the backbone and CLIP-token representations in the current mini-batch, and $|\Delta|$ is the number of elements in residual tensor Δ.

Stimulus-structure regularization. We use a training-only stimulus-structure objective

$$\mathcal{L}_{\mathrm{stim}} = \lambda_{\mathrm{exact}}\mathcal{L}_{\mathrm{exact}} + \lambda_{\mathrm{sa}}\mathcal{L}_{\mathrm{sa}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}} + \lambda_{\mathrm{sdp}}\mathcal{L}_{\mathrm{sdp}}. \tag{16}$$

The four terms respectively encourage instance discrimination over sampled training-bank candidates, margin-based alignment to class prototypes, prototype-level relation consistency, and robustness to small feature perturbations within the training set (Li et al., 2026).

Complete objective. The base reconstruction objective is

$$\mathcal{L}_{\mathrm{base}}^{(e)} = \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}^{(e)} + \lambda_{\mathrm{prior}}\mathcal{L}_{\mathrm{prior}} + \lambda_{\mathrm{low}}\mathcal{L}_{\mathrm{low}}. \tag{17}$$

Algorithm 1 MindRefine Training and Reconstruction

Input: Training pairs $\mathcal{D}_{\mathrm{tr}}^{s} = \{(x_i^s, I_i)\}_{i=1}^{N_s}$, test fMRI response x^s, VLM annotator, frozen feature extractors $\{\Phi_m\}_{m=1}^{M}$
Parameter: Loss weights, λ_{KRR}, refinement stages $\{\mathcal{M}_r, \alpha_r\}_{r=1}^{2}$
Output: Final reconstruction $\hat{I}^{\mathrm{out}}$

Generate structured annotations a_i for each training image I_i using the VLM, and encode them into channel-level text embeddings $T_i = (t_i^1, \ldots, t_i^K)$.
Initialize the decoding model, residual token adapter, semantic head h_θ, and gating network ψ_θ.
for each training epoch e **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}_{\mathrm{tr}}^{s}$ **do**
    Predict adapted brain-side visual tokens Z_i and generation-branch representations from x_i^s, and obtain image-side CLIP tokens V_i from I_i.
    Compute $\mathcal{L}_{\mathrm{base}}^{(e)}$ and $\mathcal{L}_{\mathrm{stim}}$.
    if the BiMixCo phase is active **then**
      Set $\mathcal{L}_{\mathrm{LISR}} = \mathcal{L}_{\mathrm{adapt}} = 0$.
    else
      Compute $\mathcal{L}_{\mathrm{LISR}}$ and $\mathcal{L}_{\mathrm{adapt}}$.
    end if
    Update trainable parameters by minimizing

$$\mathcal{L}_{\mathrm{train}}^{(e)} = \mathcal{L}_{\mathrm{base}}^{(e)} + \mathcal{L}_{\mathrm{LISR}} + \lambda_{\mathrm{adapt}}\mathcal{L}_{\mathrm{adapt}} + \gamma_e \mathcal{L}_{\mathrm{stim}}.$$

  end for
end for
Freeze the trained decoder and residual token adapter; discard the training-only semantic head and gating network.
Fit a lightweight low-rank token adapter on the training samples and freeze it.
Construct the reference bank $\mathcal{R}_{\mathrm{tr}}^{s} = \{(z_j, \{f_j^m\}_{m=1}^{M})\}_{j=1}^{N_s}$ using the token-adapted training embeddings z_j.
Decode x^s to obtain the fixed-fusion and training-calibrated adaptive-fusion reconstructions, combine them into the pre-refinement reconstruction $\hat{I}^{(0)}$, and obtain the token-adapted brain-side embedding z.
for each feature space m **do**
  Estimate the sample-specific target feature $\tilde{f}^m$ by kernel ridge regression over $\mathcal{R}_{\mathrm{tr}}^{s}$.
end for
for refinement stage $r = 1, 2$ **do**
  Initialize $u^{(r,0)} = \mathrm{logit}(\mathrm{Clip}_\epsilon(\hat{I}^{(r-1)}))$.
  Optimize u with $\mathcal{L}_{\mathrm{ref}}^{(r)}(u)$ over feature set $\mathcal{M}_r$.
  Set $\hat{I}_{\mathrm{ref}}^{(r)} = \sigma(u^{(r)\star})$ and blend

$$\hat{I}^{(r)} = \mathrm{Clip}\left[(1-\alpha_r)\hat{I}^{(r-1)} + \alpha_r \hat{I}_{\mathrm{ref}}^{(r)}\right].$$

end for
return $\hat{I}^{\mathrm{out}} = \hat{I}^{(2)}$

Together with the language-informed semantic regularization term $\mathcal{L}_{\mathrm{LISR}}$, the complete training objective is

$$\mathcal{L}_{\mathrm{train}}^{(e)} = \mathcal{L}_{\mathrm{base}}^{(e)} + \mathcal{L}_{\mathrm{LISR}} + \lambda_{\mathrm{adapt}}\mathcal{L}_{\mathrm{adapt}} + \gamma_e \mathcal{L}_{\mathrm{stim}}. \tag{18}$$

During the BiMixCo phase, $\mathcal{L}_{\mathrm{LISR}} = \mathcal{L}_{\mathrm{adapt}} = 0$; both regularization terms are enabled afterward, while γ_e gradually increases from zero.

The component losses above define the training stage of MindRefine, while the complete procedure also includes the inference-stage reference-bank construction and feature-space refinement. Algorithm 1 summarizes the full workflow, including VLM-based semantic annotation, loss-based decoder training, and stage-wise reconstruction refinement with the trained decoder frozen.

B Dataset Details

We use the Natural Scenes Dataset (NSD) (Allen et al., 2022), a large-scale 7T fMRI dataset collected while participants viewed natural images from MSCOCO (Lin et al., 2014). NSD contains whole-brain high-resolution fMRI responses from eight subjects over 30–40 scan sessions. Following MindEye2 (Scotti et al., 2024), we evaluate Subjects 1, 2, 5, and 7, which completed all 40 sessions. The shared set of 1,000 images viewed by all evaluated subjects is used as the held-out test set, and the remaining subject-specific image trials are used for training.

Following the preprocessing protocol of MindEye2 (Scotti et al., 2024), fMRI responses are normalized single-trial beta estimates from GLMSingle. We use preprocessed flattened voxels from the NSD "nsdgeneral" region in 1.8-mm native volume space, which covers visually responsive posterior cortex. Voxel-wise normalization is performed using statistics computed from the training set. For evaluation, repeated responses to the shared test images are averaged, while training samples are treated as single-trial fMRI–image pairs.

We consider two training-data regimes. In the limited-data setting, the target-subject model is trained using only the first session, corresponding to approximately one hour of fMRI data. In the full-data setting, the model is trained using all 40 sessions. This setup follows prior NSD-based cross-subject decoding protocols and keeps the held-out test images fixed across settings.

C Full Hyperparameter Settings

All hyperparameters are fixed across subjects. The loss weights for $\mathcal{L}_{\mathrm{LISR}}$ are set to $[\lambda_{\mathrm{sem}}, \lambda_{\mathrm{soft}}, \lambda_{\mathrm{gate}}] = [0.015, 0.008, 0.001]$. The loss weights for $\mathcal{L}_{\mathrm{stim}}$ are set to $[\lambda_{\mathrm{exact}}, \lambda_{\mathrm{sa}}, \lambda_{\mathrm{rel}}, \lambda_{\mathrm{sdp}}] = [0.08, 0.04, 0.02, 0.01]$. The weights in the base reconstruction objective are set to $[\lambda_{\mathrm{align}}, \lambda_{\mathrm{prior}}, \lambda_{\mathrm{low}}] = [1.2, 35, 0.75]$, and the residual-adapter weight is set to $\lambda_{\mathrm{adapt}} = 0.001$. The ramp coefficient γ_e is activated after the BiMixCo phase and increases linearly from 0 to 1 over 45 epochs.

For Feature-Space Reconstruction Refinement, the fixed-fusion and training-calibrated adaptive-fusion reconstructions are combined with weights 0.6 and 0.4, respectively. The low-rank token adapter used for feature-target estimation is fitted on training samples and frozen during refinement. The kernel ridge coefficient is $\lambda_{\mathrm{KRR}} = 0.05$. We use two refinement stages. The first stage uses CLIP, EfficientNet-B1, and SwAV feature spaces with feature weights [1.0, 0.25, 1.0], base-consistency weight 6.0, total-variation weight 0.02, 40 optimization steps, learning rate 0.04, and blending coefficient $\alpha_1 = 0.10$. The second stage uses AlexNet(2), AlexNet(5), CLIP, EfficientNet-B1, and SwAV feature spaces with feature weights [1.20, 0.45, 0.20, 0.05, 0.20], base-consistency weight 10.0, total-variation weight 0.02, 24 optimization steps, learning rate 0.03, and blending coefficient $\alpha_2 = 0.10$.

D VLM Annotation Prompt

Figure 7 shows the VLM prompt used to generate structured annotations from training images, with Qwen3-VL-Plus (Bai et al., 2025) serving as the VLM backbone. The prompt requires concise JSON outputs with fields for caption, objects, attributes, relations, scene, style, and uncertain content. It also constrains the model to describe only visible image content and leave unreliable fields empty. The style and uncertain fields are used only for annotation quality control and are not encoded as semantic supervision channels.

E Computing Infrastructure

All MindRefine experiments are conducted on a Linux workstation running Ubuntu 22.04.4 LTS with kernel 6.5.0. The machine is equipped with two Intel Xeon Gold 5118 CPUs at 2.30 GHz, providing 48 logical CPU cores in total, 376 GiB system memory, and four NVIDIA TITAN RTX GPUs. Each GPU has 24 GB memory. The NVIDIA driver version is 535.171.04, with CUDA 12.2 reported by `nvidia-smi`. The PyTorch build uses CUDA 12.4.

You are annotating one natural image for an fMRI-to-image reconstruction study. Return only valid JSON. Do not use markdown.

Use concise English. Be factual and only describe visible content.

Required schema:

```
{
  "caption": "one concise sentence, <= 25 words",
  "objects": ["salient visible object/category nouns, singular if possible"],
  "attributes": ["short attribute phrases, e.g. brown dog, wooden table"],
  "relations": ["short visual relations, e.g. dog on couch"],
  "scene": "short scene category or environment",
  "style": "photo/illustration/etc.",
  "uncertain": ["items that are ambiguous or hard to identify"]
}
```

Rules:

- objects should contain 3 to 10 salient items when possible.
- relations should describe spatial/action relations visible in the image.
- If a field has no reliable content, use an empty list or an empty string.

Figure 7: VLM prompt for structured annotation.

Table 4: Comparison of trainable parameters during subject-specific fine-tuning. Ridge regression parameters are reported for Subject 1. Numbers for previous methods follow their reported settings. For NeuroFlow, E&D denotes its unified encoding-decoding model for visual encoding and decoding, so its parameters are not decomposed into ridge, MLP/adapter, and prior parts. MindRefine is counted from our training configuration.

Method	Model Parts			Total
	Ridge	MLP / Adapter	Prior	
MindEye2	64.41M	1903M	260M	2.2G
MindAligner	139.23M	0M	0M	139.23M
MindTuner	64.41M	12.3M	0M	76.71M
Duala	64.41M	4.68M	0M	69.09M
NeuroFlow	Unified E&D model			660M
MindRefine	**64.41M**	**2.97M**	**0M**	**67.38M**

MindRefine is implemented with Python 3.10.19 and PyTorch 2.5.1. Distributed training uses HuggingFace Accelerate 1.13.0 and DeepSpeed 0.18.9 with FP16 mixed precision. The main software dependencies include torchvision 0.20.1, diffusers 0.23.0, transformers 5.8.0, NumPy 2.2.5, pandas 2.3.3, Pillow 11.3.0, torchmetrics 1.3.0, OpenAI CLIP 1.0, sentence-transformers 5.5.0, einops 0.8.2, webdataset 1.0.2, h5py 3.16.0, and kornia 0.7.1.

For the one-session setting, each subject is trained for 150 epochs with batch size 4 on four GPUs, requiring approximately 1.5 hours. For the full-data setting, each subject is trained for 150 epochs with batch size 20 on four GPUs, requiring approximately 29 hours. All reported MindRefine training, reconstruction, and evaluation runs use the same computing environment unless otherwise specified. Each reported model configuration is trained once with a fixed random seed of 42, while retrieval metrics are computed with 300 candidates and 30 repeats using the same seed.

Table 5: Ablation results on Subjects 1, 2, 5, and 7 under the 1-session setting, with average results computed over the four subjects. The **best** and second-best results are highlighted. ↑ and ↓ indicate that higher and lower values are better, respectively.

Ablation	Subject	Low-Level				High-Level				Retrieval	
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
w/o LISR	Avg	0.229	0.413	87.5%	93.9%	86.8%	**83.9%**	0.765	0.418	83.4%	**82.4%**
w/o $\mathcal{L}_{\text{stim}}$	Avg	0.232	0.409	87.7%	93.9%	87.2%	83.1%	0.764	0.417	83.9%	82.2%
w/o FSR	Avg	0.228	0.407	87.2%	93.6%	87.2%	81.4%	0.797	0.468	84.0%	81.6%
CLIP-only Refinement	Avg	0.230	0.411	87.5%	93.7%	87.3%	83.2%	0.793	0.462	84.0%	81.3%
MindRefine (Ours)	Avg	**0.239**	**0.419**	**88.3%**	**94.1%**	**87.8%**	83.6%	**0.758**	**0.416**	**84.2%**	81.8%
w/o LISR	Subj1	0.266	0.422	91.0%	95.3%	87.9%	**85.5%**	0.754	0.406	91.7%	91.2%
w/o $\mathcal{L}_{\text{stim}}$	Subj1	0.263	0.421	91.1%	95.5%	88.7%	84.8%	0.747	**0.401**	**93.8%**	**93.8%**
w/o FSR	Subj1	0.264	0.412	91.1%	95.3%	88.4%	83.0%	0.781	0.453	**93.8%**	92.3%
CLIP-only Refinement	Subj1	0.265	0.423	91.0%	95.3%	88.6%	85.1%	0.780	0.446	**93.8%**	92.2%
MindRefine (Ours)	Subj1	**0.276**	**0.425**	**91.3%**	**95.7%**	**89.1%**	85.3%	**0.742**	**0.401**	**93.8%**	92.3%
w/o LISR	Subj2	0.235	0.425	89.6%	95.7%	87.9%	**84.7%**	0.755	0.411	92.8%	**92.6%**
w/o $\mathcal{L}_{\text{stim}}$	Subj2	**0.251**	0.421	90.0%	95.7%	88.5%	84.3%	0.750	0.408	**93.3%**	92.2%
w/o FSR	Subj2	0.232	0.418	87.9%	95.5%	88.5%	83.3%	0.785	0.460	92.5%	92.4%
CLIP-only Refinement	Subj2	0.241	**0.429**	89.8%	95.6%	88.6%	84.5%	0.784	0.453	92.5%	90.4%
MindRefine (Ours)	Subj2	0.250	0.428	**90.1%**	**95.8%**	**88.9%**	84.5%	**0.745**	**0.406**	**93.3%**	92.4%
w/o LISR	Subj5	0.212	0.409	85.8%	93.2%	87.8%	84.7%	0.757	**0.412**	74.9%	70.1%
w/o $\mathcal{L}_{\text{stim}}$	Subj5	0.201	0.399	85.9%	93.0%	86.7%	83.1%	0.769	0.418	75.2%	**70.5%**
w/o FSR	Subj5	0.212	0.412	86.0%	92.8%	87.1%	80.3%	0.800	0.473	**75.4%**	**70.5%**
CLIP-only Refinement	Subj5	0.202	0.410	85.6%	92.8%	87.1%	82.7%	0.789	0.467	**75.4%**	**70.5%**
MindRefine (Ours)	Subj5	**0.214**	**0.416**	**86.9%**	**93.5%**	**87.9%**	**84.9%**	**0.755**	0.416	**75.4%**	**70.5%**
w/o LISR	Subj7	0.202	0.394	83.7%	91.3%	83.5%	**80.5%**	0.795	0.441	74.2%	**75.6%**
w/o $\mathcal{L}_{\text{stim}}$	Subj7	0.213	0.394	83.8%	**91.5%**	84.7%	80.1%	0.791	**0.439**	73.3%	72.4%
w/o FSR	Subj7	0.203	0.385	83.8%	90.9%	84.6%	79.0%	0.821	0.487	**74.4%**	71.0%
CLIP-only Refinement	Subj7	0.213	0.383	83.6%	90.9%	84.8%	80.4%	0.820	0.482	**74.4%**	72.0%
MindRefine (Ours)	Subj7	**0.217**	**0.408**	**84.9%**	**91.5%**	**85.2%**	79.8%	**0.788**	**0.439**	**74.4%**	72.0%

F Parameter Efficiency Analysis

Table 4 compares the number of trainable parameters during subject-specific fine-tuning. Following prior work, the ridge regression parameters are reported for Subject 1. MindRefine freezes the pretrained decoder backbone and diffusion prior, and only updates the subject-specific ridge layer together with lightweight adaptation and semantic modules. The additional trainable modules include the residual token adapter, semantic gating network, and semantic projection head, totaling 2.97M parameters. As a result, MindRefine requires only 67.38M trainable parameters, fewer than the compared baselines under this trainable-parameter accounting. NeuroFlow reports 660M trainable parameters for its unified encoding-and-decoding model, which follows a different architectural decomposition but remains substantially larger than MindRefine. This indicates that MindRefine improves parameter efficiency without updating the large backbone or diffusion prior.

G Baseline Methods

We compare MindRefine with representative fMRI-to-image decoding methods that cover shared-subject adaptation, explicit cross-subject alignment, semantic correction, dual-level alignment, and unified neural–visual modeling. All baselines are evaluated on NSD subjects commonly used in prior cross-subject decoding studies, and the results are organized under the same metric groups as our evaluation.

MindEye2. MindEye2 (Scotti et al., 2024) is a shared-subject fMRI-to-image reconstruction framework designed to reduce the amount of target-subject data required for adaptation. It first pretrains a model across multiple source subjects and then fine-tunes it with limited data from the target subject. MindEye2 maps flattened fMRI voxels into a shared-subject latent space through subject-specific ridge regression, predicts CLIP image embeddings with a shared nonlinear decoder,

Table 6: Quantitative comparison on Subjects 1, 2, 5, and 7 under the 40-session setting, with average results computed over the four subjects. The **best** and second-best results are highlighted. ↑ and ↓ indicate that higher and lower values are better, respectively.

Method	Subject	Low-Level				High-Level				Retrieval	
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
MindEye2 (Scotti et al., 2024)	Avg	0.322	**0.431**	**96.1%**	98.6%	95.4%	93.0%	0.619	0.344	99.2%	99.0%
NeuroFlow (Mai et al., 2026)	Avg	0.274	0.266	94.4%	98.0%	95.6%	**93.9%**	0.669	0.380	90.3%	–
MindRefine (Ours)	Avg	**0.341**	0.430	**96.1%**	**99.0%**	**96.4%**	93.4%	**0.580**	**0.313**	**99.4%**	**99.2%**
MindEye2 (Scotti et al., 2024)	Subj1	0.374	**0.439**	97.8%	99.1%	96.2%	93.6%	0.609	0.338	99.7%	**100.0%**
NeuroFlow (Mai et al., 2026)	Subj1	0.305	0.265	96.1%	98.6%	96.2%	**94.6%**	0.661	0.371	95.5%	–
MindRefine (Ours)	Subj1	**0.392**	0.438	**97.9%**	**99.5%**	**97.1%**	93.8%	**0.575**	**0.309**	**99.9%**	**100.0%**
MindEye2 (Scotti et al., 2024)	Subj2	0.328	**0.430**	97.0%	98.8%	94.9%	91.7%	0.631	0.347	99.6%	99.6%
NeuroFlow (Mai et al., 2026)	Subj2	0.278	0.264	95.2%	98.4%	95.5%	**93.2%**	0.670	0.377	92.2%	–
MindRefine (Ours)	Subj2	**0.352**	0.429	**97.3%**	**99.3%**	**96.3%**	92.8%	**0.580**	**0.310**	**99.7%**	**99.9%**
MindEye2 (Scotti et al., 2024)	Subj5	0.301	**0.432**	**95.3%**	98.5%	96.5%	94.3%	0.600	0.335	98.6%	98.1%
NeuroFlow (Mai et al., 2026)	Subj5	0.262	0.266	93.6%	98.1%	96.2%	**95.1%**	0.656	0.376	90.7%	–
MindRefine (Ours)	Subj5	**0.317**	0.431	94.9%	**99.0%**	**97.0%**	94.9%	**0.559**	**0.304**	**98.7%**	**98.2%**
MindEye2 (Scotti et al., 2024)	Subj7	0.283	**0.423**	94.3%	98.0%	94.1%	92.4%	0.638	0.357	98.9%	98.2%
NeuroFlow (Mai et al., 2026)	Subj7	0.252	0.269	92.6%	97.1%	94.6%	**92.9%**	0.688	0.396	82.9%	–
MindRefine (Ours)	Subj7	**0.304**	0.422	**94.3%**	**98.4%**	**95.1%**	92.2%	**0.607**	**0.330**	**99.4%**	**98.6%**

and reconstructs images through an unCLIP-based generative model. It also contains retrieval and low-level reconstruction submodules, making it a strong baseline for both 1-session and 40-session comparisons.

MindAligner. MindAligner (Dai et al., 2025) addresses limited-data cross-subject visual decoding through explicit brain functional alignment. Its core component is a Brain Transfer Matrix that projects fMRI signals from a new subject into the functional space of a known subject, enabling the use of pretrained decoding models. To learn this transfer reliably, MindAligner introduces a brain functional alignment module with multi-level alignment losses, aiming to capture fine-grained cross-subject correspondences under different visual stimuli.

MindTuner. MindTuner (Gong et al., 2025) focuses on cross-subject visual decoding with limited target-subject data by modeling subject-specific visual fingerprints and applying semantic correction. It uses lightweight LoRA-based adaptation modules to learn individual visual-response characteristics and introduces an fMRI-to-text alignment paradigm through the image modality as an intermediate bridge. The predicted semantic information is then used to improve fMRI-to-image reconstruction, making MindTuner a relevant baseline for semantic enhancement under the 1-session setting.

Duala. Duala (Li et al., 2026) is a dual-level alignment framework for cross-subject fMRI decoding. At the stimulus level, it introduces semantic alignment and relational consistency to preserve intra-class similarity and inter-class separability during adaptation. At the subject level, it uses distribution-based feature perturbation to model both global and subject-specific variations, improving adaptation to individual neural representations with limited target-subject data. Because Duala explicitly targets both stimulus consistency and subject alignment, it is a key comparison for limited-data cross-subject decoding.

NeuroFlow. NeuroFlow (Mai et al., 2026) aims to unify visual encoding and decoding from neural activity within a single flow-based framework. It introduces a variational neural backbone to model neural variability and build a compact semantically structured latent space, together with cross-modal flow matching to learn a reversible mapping between visual and neural latent distributions. Unlike methods that separately train encoding and decoding models, NeuroFlow formulates the two directions as a shared bidirectional modeling problem. We include NeuroFlow as a recent strong baseline in both the 1-session and 40-session settings.



Figure 8: Qualitative comparison of image reconstructions under the 1-session setting.

H Additional Qualitative Results

Figure 8 provides additional qualitative comparisons under the 1-session setting. MindRefine consistently preserves object identity, global composition, and spatial structure across diverse scenes.

I Detailed Subject-Wise Ablation Analysis

Table 5 reports the complete subject-wise ablation results under the 1-session setting. The averaged results are consistent with those reported in the main paper: MindRefine achieves the best performance on seven of the eight low-level and high-level reconstruction metrics, with the second-best result on CLIP similarity. A similar pattern is observed across all four subjects, where the full model achieves the best or tied-best results on most low-level and high-level metrics. This consistency indicates that the overall gains extend across subjects with different neural response patterns.

The training-time ablations verify the contributions of LISR and $\mathcal{L}_{\text{stim}}$. Removing LISR decreases the average PixCorr from 0.239 to 0.229, AlexNet(2) from 88.3% to 87.5%, and Inception from 87.8% to 86.8%. Similar reductions appear across the subject-wise reconstruction results. Although w/o LISR occasionally obtains higher CLIP similarity or brain retrieval accuracy, these isolated gains do not extend to the broader set of reconstruction metrics. This suggests that LISR complements visual-token alignment with structured semantic constraints, improving the overall organization of brain-side representations rather than optimizing only their similarity to the CLIP space. Removing $\mathcal{L}_{\text{stim}}$ also reduces every averaged reconstruction metric. Its effect is particularly visible in PixCorr, SSIM, AlexNet features, and perceptual distance metrics, supporting the role of stimulus-structure regularization in stabilizing representation learning under limited data.

The inference-time ablations show a clearer effect on perceptual reconstruction quality. Without FSR, the average CLIP score decreases from 83.6% to 81.4%, while Eff and SwAV deteriorate from 0.758 and 0.416 to 0.797 and 0.468, respectively. The degradation in Eff and SwAV is consistently observed for every subject, indicating that feature-space refinement improves perceptual similarity beyond the initial fused reconstruction. Low-level metrics also decline without FSR, showing that the refinement does not improve semantic consistency at the expense of image structure.

CLIP-only Refinement partially recovers the performance of the full refinement pipeline but remains inferior to MindRefine on all averaged reconstruction metrics. In particular, its Eff and SwAV scores remain close to those of w/o FSR, despite achieving stronger CLIP similarity. Subject-wise results show that CLIP-only Refinement may match or exceed the full model on an individual metric, but it does not provide the same balanced improvements across low-level structure and high-level perception. These results confirm that CLIP features alone provide an incomplete refinement target and that combining complementary perceptual and semantic feature spaces is important for FSR.

Retrieval performance varies less consistently across the ablations. The full model achieves the highest average image retrieval score, whereas w/o LISR and w/o $\mathcal{L}_{\mathrm{stim}}$ obtain slightly higher average brain retrieval scores. This difference is consistent with the reconstruction-oriented role of the proposed components: LISR, $\mathcal{L}_{\mathrm{stim}}$, and FSR primarily improve image fidelity and perceptual consistency, while their effects on retrieval depend on the subject and retrieval direction. Overall, the subject-wise results support the complementary contributions of training-time representation regularization and inference-time feature-space refinement.

J Detailed Results under the 40-Session Setting

Table 6 reports the complete subject-wise comparison under the 40-session setting. On average, MindRefine achieves the best or tied-best result on six of the eight low-level and high-level reconstruction metrics, with the second-best performance on SSIM and CLIP. It also obtains the highest image and brain retrieval accuracy, showing that its advantages remain evident when substantially more subject-specific fMRI data are available.

The subject-wise results exhibit a consistent pattern. MindRefine achieves the best PixCorr, AlexNet(5), Inception, Eff, and SwAV results across all four subjects, while attaining the best or tied-best AlexNet(2) performance for most subjects. The improvements across both similarity- and distance-based metrics indicate that MindRefine enhances spatial structure, perceptual fidelity, and high-level visual consistency rather than favoring a single evaluation criterion. Although MindEye2 remains slightly stronger on SSIM and NeuroFlow achieves higher CLIP similarity in several cases, MindRefine provides a more balanced reconstruction profile across the full set of metrics.

This performance is consistent with the complementary design of LISR and FSR. LISR introduces structured semantic constraints into brain-side representation learning, which helps organize visual content beyond implicit token alignment. FSR further estimates sample-specific targets in multiple feature spaces and optimizes the reconstruction with the decoder frozen. The strong Inception, Eff, and SwAV results, despite the second-best average CLIP score, suggest that the refinement does not overfit to a single embedding space and instead improves broader perceptual and semantic consistency.

K Parameter Sensitivity Analysis

We analyze the sensitivity of MindRefine to two groups of hyperparameters: the semantic regularization weight λ_{sem} and the feature-space reconstruction refinement parameters. All sensitivity experiments use the same random seed of 42 and the same evaluation protocol as the main experiments. Each configuration is trained once; therefore, confidence intervals across independent training runs are not reported. Retrieval metrics are computed over 30 repeated candidate samplings, consistent with the main evaluation protocol.

For λ_{sem}, we evaluate four scaling factors relative to the default weight: $0\times$, $0.5\times$, $1\times$, and $2\times$. As shown in Figure 9, the default scale $1\times$ provides a balanced profile across low-level reconstruction quality, high-level perceptual consistency, and retrieval performance. Removing semantic regularization weakens the overall profile, while overly increasing its strength does not consistently improve all metrics. This suggests that language-informed semantic regularization is beneficial but should be balanced with the original reconstruction objectives.

We further study the inference-time refinement parameters, including the blending coefficient α and the base-consistency scale. The blending coefficient controls how strongly the refined reconstruction contributes to the final output, while the base-consistency scale adjusts the strength of the consistency constraint between the optimized reconstruction and the stage input. As shown in

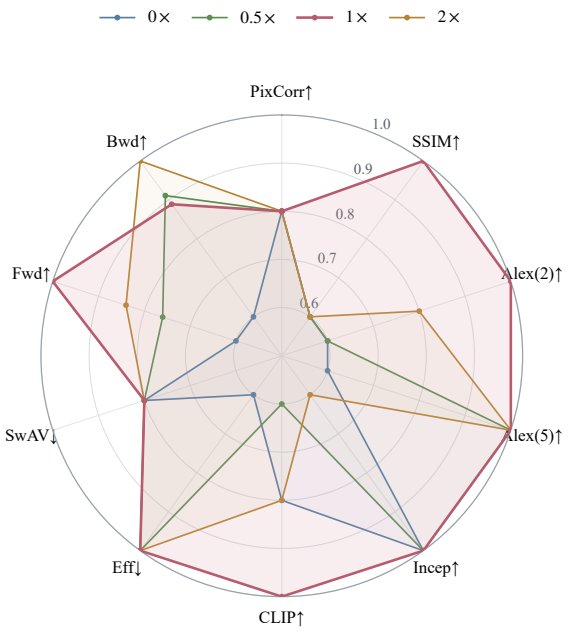

Figure 9: Parameter sensitivity analysis of the semantic regularization weight λ_{sem}. Each axis denotes one evaluation metric. For visualization, metric values are first converted to direction-consistent scores, where higher values indicate better performance. EffNet-B and SwAV are therefore reversed before normalization. Scores are then normalized within each metric to $[0, 1]$, and the radar plot displays the range from 0.6 to 1.0 for clarity. The default setting uses semantic scale $1\times$, corresponding to $\lambda_{\text{sem}} = 0.015$.

Figure 10, the refinement stage shows a clear trade-off across reconstruction metrics. Moderate blending preserves low-level quality while improving high-level perceptual metrics, supporting our default choice of $\alpha = 0.10$ and base-consistency scale $1\times$. These results indicate that FSR provides a robust post-hoc adjustment rather than relying on a narrowly tuned parameter choice.

L Detailed Analysis of Perceptual Refinement

We focus on the eight low-level and high-level reconstruction metrics and omit retrieval metrics because this analysis isolates post-hoc image refinement. The retrieval branch and brain-side representations remain unchanged across the initial, fixed-blend, and FSR variants.

Figure 11 provides the averaged comparison of the initial reconstruction, fixed image-space blending, and FSR across all eight reconstruction metrics. For both MindEye2 and MindRefine, FSR improves most metrics, including PixCorr, SSIM, AlexNet(5), Inception, CLIP, Eff, and SwAV, covering both low-level similarity and high-level perceptual consistency. AlexNet(2) changes only marginally, suggesting that the refinement does not uniformly optimize every feature space but improves the overall reconstruction profile. Compared with fixed blending, which mainly benefits low-level metrics and shows less consistent high-level effects, FSR further enhances semantic and perceptual quality by estimating sample-specific feature targets from paired training references and optimizing the reconstruction in multiple feature spaces.

Applying FSR to MindEye2 produces a similar improvement pattern, demonstrating that the refinement is not specific to the MindRefine training architecture. Its effectiveness across two decoders suggests that training-reference-guided feature optimization can serve as a general post-hoc refinement strategy. Overall, the averaged results confirm that FSR provides more balanced perceptual correction than either the initial reconstruction or fixed image-space blending.

Figures 12, 13, 14, and 15 report the corresponding subject-wise results. The overall trend remains consistent across subjects, with FSR producing broad gains for both decoders. The magnitude of

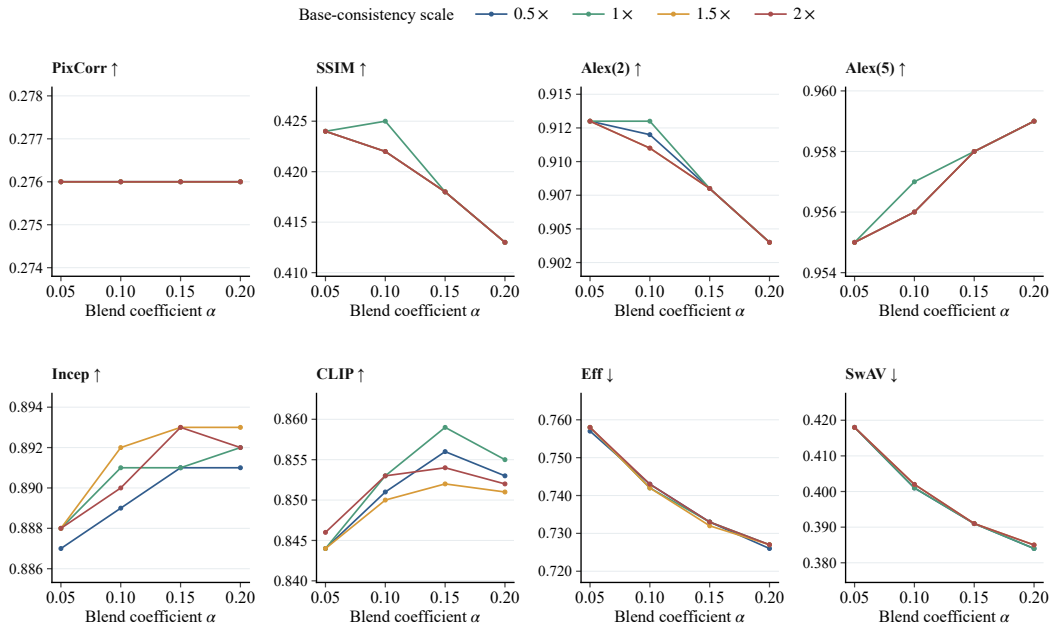

Figure 10: Parameter sensitivity analysis of the feature-space reconstruction refinement stage. We vary the blending coefficient α and the base-consistency scale used by the refinement module. Each panel reports one reconstruction metric, with arrows indicating the preferred direction. Larger α gives more weight to the feature-space refined reconstruction, while the base-consistency scale adjusts the strength of the consistency constraint between the optimized reconstruction and the stage input. The default setting is $\alpha = 0.10$ with base-consistency scale $1\times$, corresponding to $\lambda_{\text{base}}^{(1)} = 6.0$ in the first refinement stage and $\lambda_{\text{base}}^{(2)} = 10.0$ in the second refinement stage.

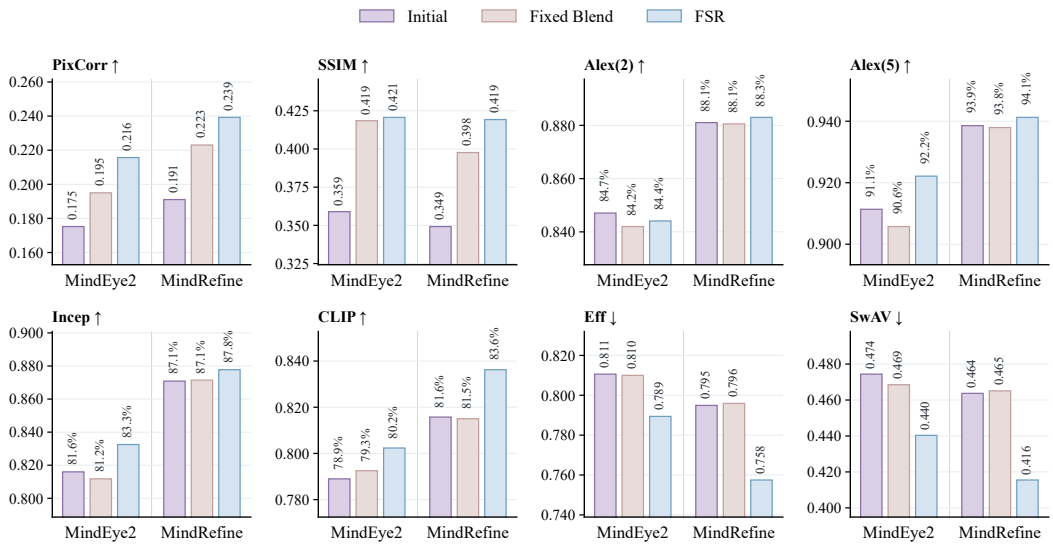

Figure 11: Average comparison of the initial, fixed blend, and FSR reconstructions for MindEye2 and MindRefine across Subjects 1, 2, 5, and 7.

improvement varies across subjects and metrics, reflecting differences in decoding difficulty and initial reconstruction quality. Small decreases are observed on AlexNet(2) for some subjects, but these isolated cases are accompanied by improvements across most remaining low-level and high-level evaluations.

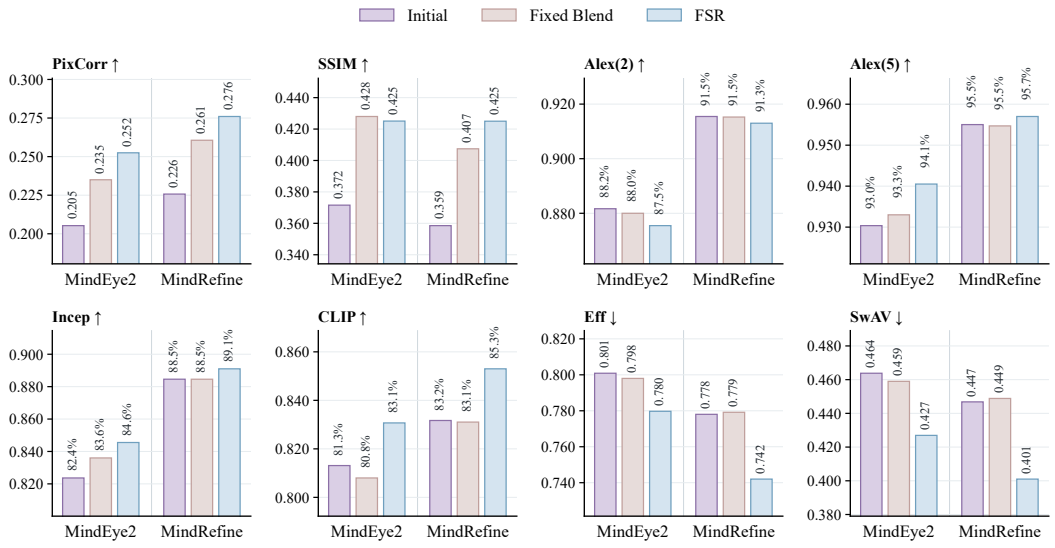

Figure 12: Comparison of the initial, fixed blend, and FSR reconstructions for MindEye2 and MindRefine on Subject 1.

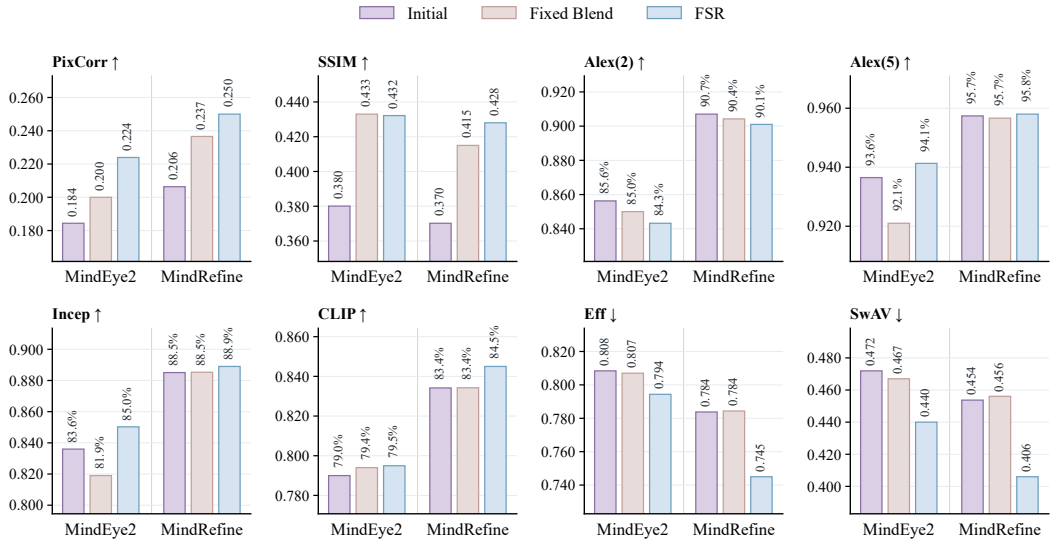

Figure 13: Comparison of the initial, fixed blend, and FSR reconstructions for MindEye2 and MindRefine on Subject 2.

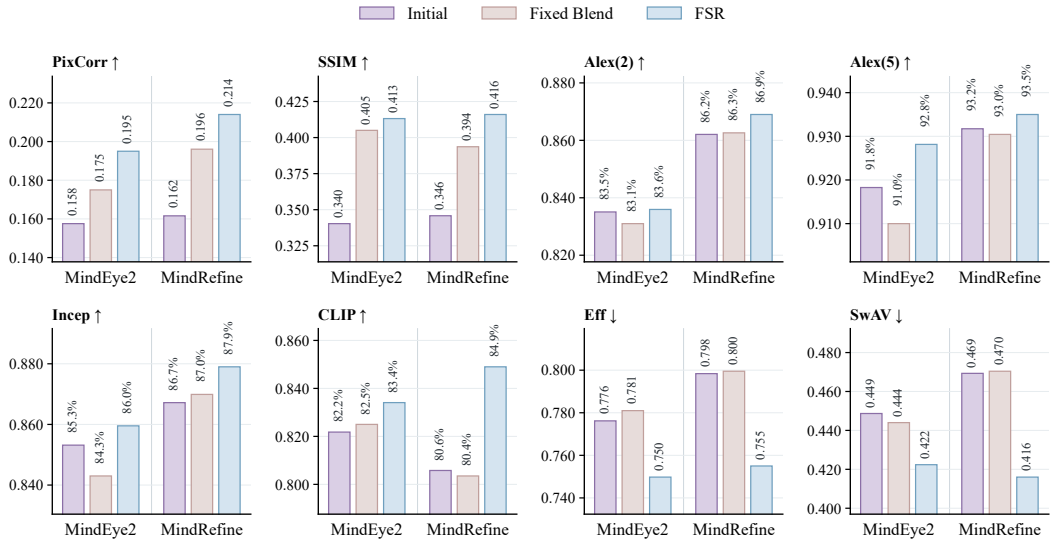

Figure 14: Comparison of the initial, fixed blend, and FSR reconstructions for MindEye2 and MindRefine on Subject 5.

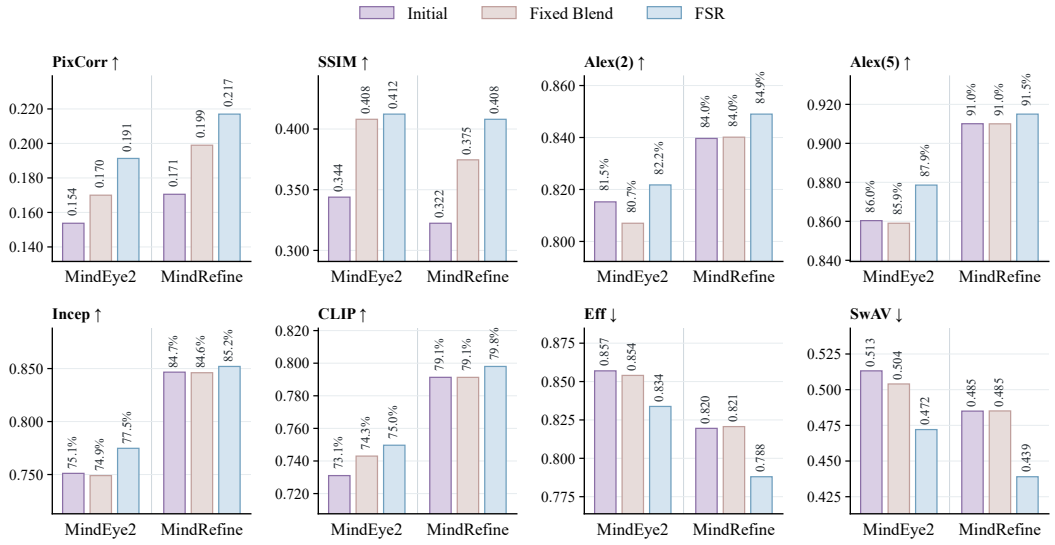

Figure 15: Comparison of the initial, fixed blend, and FSR reconstructions for MindEye2 and MindRefine on Subject 7.