WALLBREAKER: BREAKING THE VISION-IMAGERY BARRIER THROUGH CORTICAL HIERARCHY-DRIVEN NEURAL DECODING

Anonymous authors
Paper under double-blind review

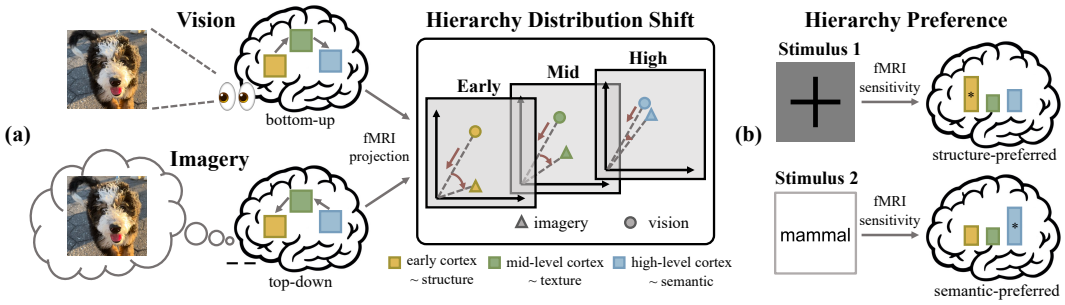


Figure 1: **Hierarchical variability in neural representations. (a) Distribution-level hierarchy shift.** Vision and imagery rely on different brain processing mechanisms; vision is primarily bottom-up, while imagery is more top-down. This induces systematic distribution shifts in fMRI representations across early, mid-level, and high-level visual cortex. **(b) Stimulus-level hierarchy preference.** Neural responses show stimulus-dependent sensitivity across hierarchy levels, reflecting varying reliance on structural versus semantic representations.

ABSTRACT

Decoding mental imagery from fMRI is fundamentally challenging: models trained on visual perception degrade sharply when applied to imagery. We argue that this gap is not merely a signal-to-noise problem, but stems from a *hierarchy-structured representation shift* in the visual cortex. Vision is primarily driven by bottom-up sensory input, whereas imagery is shaped by top-down internal generation, systematically redistributing information across cortical levels. This induces two forms of hierarchy-dependent variability: a distribution-level shift in the balance of cortical information between vision and imagery, and a stimulus-level preference where different stimuli rely to different degrees on structural versus semantic information. We propose WALLBREAKER, a framework that explicitly models cortical hierarchy-dependent variability during neural decoding. We first introduce hierarchy-driven representation perturbation, which simulates realistic hierarchy shifts during training by redistributing and perturbing representations across cortical levels. We further develop hierarchy-conditioned reconstruction, which infers a stimulus-wise hierarchy preference to dynamically balance semantic guidance and structural anchoring during reconstruction. WALLBREAKER improves key imagery reconstruction metrics while maintaining competitive vision reconstruction performance, effectively narrowing the vision-to-imagery gap. More broadly, our results highlight cortical hierarchy as a fundamental inductive bias for bridging vision and imagery in neural decoding.

1 INTRODUCTION

Understanding and decoding human brain activity is a central challenge at the intersection of neuroscience and artificial intelligence [44, 1]. Functional magnetic resonance imaging (fMRI) provides a non-invasive measurement of large-scale neural activity, making it possible to study how visual

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.

information is encoded in the cortex, and to reconstruct images from brain signals [1, 4]. Recent progress in deep generative modeling, especially diffusion-based reconstruction, has substantially improved the quality of fMRI-to-image decoding, by mapping neural activity into pretrained representation spaces, and ultimately translating them into visually plausible images [33, 34, 12, 2].

Despite these advances, most existing work focuses on *visual perception* decoding, where subjects directly observe external stimuli, and fMRI responses are used for image reconstruction. In contrast, *mental imagery* decoding remains substantially more challenging and less developed, as the target representation is internally generated rather than externally driven. Recent studies [14, 15] have shown that reconstruction pipelines trained on vision data degrade markedly when applied to imagery, indicating that the vision-to-imagery gap is a central bottleneck for neural decoding.

We hypothesize that this gap is not sufficiently explained as a generic decrease in signal-to-noise ratio [30, 32], but can be better modeled as a structured representation shift across the cortical hierarchy [9, 26]. Early visual regions mainly encode structural information, mid-level regions capture appearance properties, while higher-level regions are associated with semantic representations [10, 7]. Visual perception is primarily driven by *bottom-up* sensory input, with strong engagement of early visual cortex (hierarchy levels) and progressive abstraction toward higher levels. In contrast, mental imagery is driven by *top-down* feedback from higher hierarchy levels and reactivation of visual representations without external input [8, 9, 26]. Consequently, the relative contribution of early, mid-level, and high-level cortical signals varies systematically between vision and imagery. This hierarchical variability leads to two distinct challenges. First, the transfer from vision to imagery introduces a **distribution-level hierarchy shift** [31] (Figure 1(a)): the balance of information across cortical levels is no longer stable between training and test time. Second, even within the same domain, different stimuli depend on different levels of the visual hierarchy. Some reconstructions rely more heavily on low-level structure, while others depend more on high-level semantic content. This gives rise to a **stimulus-level hierarchy preference** (Figure 1(b)), implying that a fixed reconstruction strategy is unlikely to be optimal for all inputs.

These challenges suggest that robust imagery decoding requires more than a stronger generative backbone. A successful decoder must both *generalize* across hierarchy-structured distribution shifts and *adapt* its reconstruction strategy to the hierarchy preference of each stimulus. We propose WALLBREAKER, a unified framework that addresses both aspects explicitly. Our contributions are:

- We formulate vision-to-imagery decoding as a *hierarchy-structured transfer problem*, where relative contributions of early, mid, and high-level cortical signals vary across domains and stimuli.
- We propose *hierarchy-driven representation perturbation*, a structured training strategy that models both inter-hierarchy redistribution and intra-hierarchy representational drift. Our perturbed representations exhibit substantially reduced vision-to-imagery distance (Figure 5(a)).
- We propose *hierarchy-conditioned reconstruction*, which estimates stimulus-wise hierarchy preference, to adaptively balance semantic guidance and structural anchoring during generation. Our predicted preferences align consistently with different stimulus types (Figure 5(b)).
- We show that explicitly modeling cortical hierarchy improves both mental imagery and visual perception reconstruction on the NSD and NSD-Imagery datasets.

2 RELATED WORK

fMRI-to-Image Reconstruction. Early visual reconstruction studies demonstrated the feasibility of recovering visual content from fMRI, initially from simple visual patterns and later from natural images [22, 23]. Subsequent methods [11, 35] further advanced this line of research by decoding hierarchical visual representations from brain activity, typically through linear mappings from fMRI signals to feature spaces of pretrained models [27, 29]. Although these methods provided important evidence that visual information can be decoded from neural responses, their reconstructions were still constrained by limited spatial fidelity, weak semantic expressiveness, and restricted generalization to complex natural scenes. Recent work has substantially improved reconstruction quality by leveraging pretrained vision-language representations and diffusion-based generative models. Methods based on latent diffusion [38, 25, 21, 47], masked brain modeling [5], multimodal contrastive alignment [17, 33, 34], and brain surface [45, 46] have enabled higher-resolution and more semantically plausible reconstructions from fMRI. These approaches benefit from mapping brain signals

into powerful pretrained embedding spaces and conditioning strong image generators on the predicted representations [38, 25, 33, 34].

However, most existing methods [11, 41, 13] still rely on relatively fixed mappings from neural activity to intermediate visual representations or final reconstructions. As a result, they rarely explicitly model hierarchy-dependent variability across cortical levels, nor do they adapt the reconstruction process to stimulus-specific hierarchical structures. Explicit reconstruction policies that dynamically account for cortical hierarchy and stimulus-dependent hierarchy therefore, remain comparatively underexplored.

Vision-to-Imagery Transfer. Compared with visual perception decoding, which has been extensively studied in fMRI-to-image reconstruction, mental imagery decoding remains considerably more challenging. The NSD-Imagery benchmark [14] highlights a substantial performance gap between vision and imagery decoding, indicating that models trained on perception data do not directly generalize to internally generated representations. Recent studies [15, 24, 36] have begun to narrow this gap by improving cross-decoding robustness through multimodal conditioning, domain adaptation, and latent alignment. These efforts suggest that the vision-to-imagery gap is not merely caused by increased noise, but reflects a structured mismatch between perception-evoked and imagery-evoked neural representations. However, existing methods mainly address this mismatch as a global robustness or representation-alignment problem, without explicitly modeling where the shift arises across the cortical hierarchy or how it varies with stimulus content.

3 WALLBREAKER: CORTICAL HIERARCHY-DRIVEN NEURAL DECODING

The central idea of our framework is that imagery decoding requires solving two complementary problems. First, robustness to *distribution-level hierarchy shift* is essential, since the relative contributions of early, mid-level, and high-level cortical signals differ between vision and imagery. Second, adaptation to *stimulus-level hierarchy preference* also plays a critical role, since different stimuli rely on the cortical hierarchy to varying degrees and may favor different balances between semantic guidance and structural preservation during reconstruction.

As illustrated in Figure 2, our full pipeline follows a teacher-student design. A reference model trained on clean vision data provides stable target representations and sensitivity-based supervision. A student model is trained with hierarchy-driven perturbations to produce robust multimodal reconstruction embeddings. During inference, a hierarchy preference model predicts the extent to which the current input depends on different cortical levels, and this preference is used to modulate semantic and structural pathways during image generation.

3.1 PROBLEM FORMULATION: HIERARCHY-STRUCTURED TRANSFER

Let $x \in \mathbb{R}^V$ denote the fMRI signal and y the target representation. Standard decoding methods learn a mapping $f_\theta : x \mapsto y$ under the assumption that training and test data follow a consistent distribution. However, this assumption breaks down in vision-to-imagery transfer: The visual cortex is hierarchically organized, with early, mid-level, and high-level regions encoding progressively more abstract information. Vision is dominated by bottom-up input, while imagery relies more on top-down feedback, leading to shifts in the contribution of different cortical levels. We formulate the input as

$$x = [x_e, x_m, x_h], \quad x \sim p(x \mid \pi), \quad \pi \in \Delta^3, \tag{1}$$

where x_e, x_m, and x_h correspond to the early, mid-level, and high-level visual cortex regions [10, 28]. We use $\pi(x)$ as an operational estimate of hierarchy dependence, which measures the sensitivity of the predicted reconstruction representation to each level.

Under this formulation, vision-to-imagery transfer corresponds to a shift in the distribution over π: imagery does not simply weaken the signal uniformly; rather, it changes how information is distributed across the cortical hierarchy. In addition, π varies across stimuli, reflecting whether image reconstruction depends more strongly on lower-level structures or higher-level semantics.

This formulation motivates our method design. We seek a framework that is robust to changes in hierarchy composition between visual perception and mental imagery, while adapting its reconstruction strategy to the hierarchy preference of each stimulus.

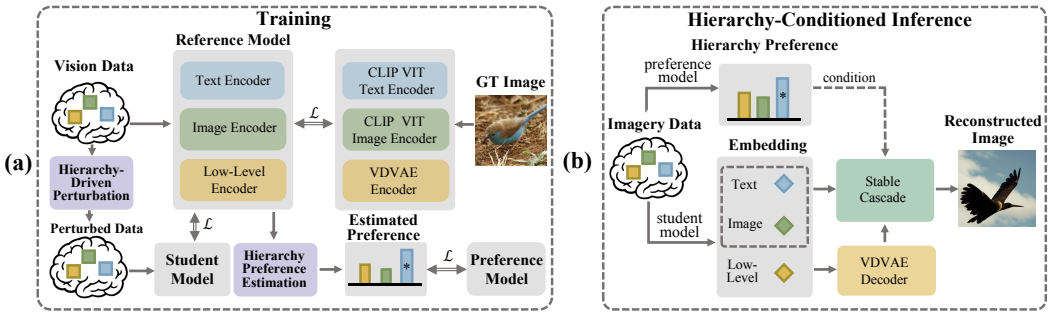


Figure 2: **Hierarchy-conditioned reconstruction pipeline. (a) Model training.** Both the reference and student encoders share the same architecture, consisting of three ridge regression models that map fMRI signals into CLIP text, CLIP image, and VDVAE-based low-level representations, providing complementary semantic and structural supervision. The reference encoder is trained on clean inputs to provide stable targets, while the student encoder is trained on hierarchy-perturbed inputs, where hierarchy-driven perturbation introduces structured variability across cortical levels. In parallel, the hierarchy preference model is trained using sensitivity-derived targets to estimate stimulus-wise hierarchy preference for adaptive reconstruction. **(b) Hierarchy-conditioned inference.** Given fMRI signals, the preference model predicts a stimulus-wise hierarchy preference, while the encoder produces multimodal representations. Semantic embeddings are fed into the Stable Cascade generator, whereas low-level representation provides a structural prior. The preference adaptively modulates the balance between semantic and structural guidance during reconstruction.

3.2 Hierarchy-Driven Representation Perturbation

We introduce hierarchy-driven representation perturbation to model distribution shift at both the inter-hierarchy and intra-hierarchy levels. As shown in Figure 3, our perturbation strategy models two complementary variabilities: **(1) Hierarchy-driven degradation:** inter-level redistribution that modifies the relative contribution of different cortical levels. **(2) Subspace perturbation:** intra-level representational drift that alters the internal geometry within each hierarchy.

Hierarchy-Driven Degradation. We first model changes in hierarchy composition by stochastically degrading different cortical hierarchy levels of the input $x = [x_e, x_m, x_h]$. For each stimulus, we independently sample degradation coefficients at each hierarchy level:

$$s = (s_e, s_m, s_h), \quad s_l \sim p_l(\cdot),$$

where each $p_l(\cdot)$ is a pre-defined, biased distribution over degradation levels for hierarchy level $l \in \{e, m, h\}$. Lower-level representations are assigned a higher probability of stronger degradation, while higher-level representations are more likely to receive mild attenuation. The degraded representation is then

$$\tilde{x} = [s_e x_e,\ s_m x_m,\ s_h x_h].$$

Unlike global noise injection, this operation explicitly changes the balance of information across hierarchy levels , which helps the model generalize more effectively under hierarchy-structured distribution shifts.

Subspace Perturbation. Hierarchy-driven degradation models *which* cortical level contributes more or less strongly, yet it does not capture changes in the internal structure of the representation *within* each level. To account for this, we further introduce subspace perturbation. Specifically, for each hierarchy level $l \in \{e, m, h\}$, let $x_l \in \mathbb{R}^{d_l}$ denote the corresponding subvector. We estimate a low-rank subspace using clean training data. Let $U_l \in \mathbb{R}^{d_l \times k_l}$ be the top-k_l PCA basis estimated from clean training data for hierarchy level l. We define the dominant voxel-space component as $z_l = U_l U_l^\top x_l$ and the corresponding residual as $r_l = x_l - z_l$. We then perturb the dominant subspace using the two following transformations.

- *Subspace Rotation.* We apply a near-identity orthogonal transformation, $\tilde{z}_l = R_l \circ z_l$, where the magnitude of rotation depends on the hierarchy level. Stronger rotations are applied to lower-level representations, which are more sensitive to fine-grained reorganization, while higher-level representations are perturbed more conservatively.

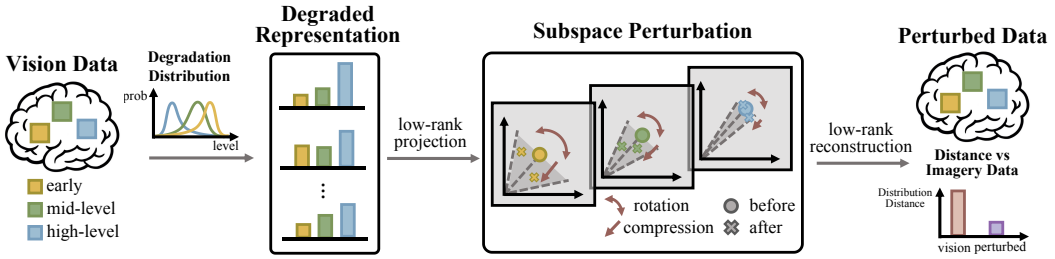


Figure 3: **Hierarchy-driven representation perturbation.** We construct structured perturbations on vision fMRI data to simulate hierarchy-dependent variability. We first apply **hierarchy-driven degradation** by sampling from a biased degradation distribution over cortical levels. Next, the representation is projected into a low-rank subspace, where we perform **subspace perturbation** by applying rotation and compression within each level. The perturbed representation is then mapped back to fMRI space, producing perturbed samples that capture both inter-hierarchy redistribution and intra-hierarchy variation.

- *Subspace Compression.* We further introduce controlled attenuation within the subspace, $\tilde{x}_l = \alpha_l \tilde{z}_l + \beta_l r_l$. Here, $\alpha_l \in [0, 1]$ controls compression of the dominant subspace, and β_l controls the residual contribution. Again, the perturbation strength is made hierarchy-dependent, with stronger compression applied to lower-level representations and weaker compression to higher-level ones.

The final perturbed representation $\tilde{x}$ is obtained by mapping the modified components back to the original voxel space, $\tilde{x} = [\tilde{x}_e, \tilde{x}_m, \tilde{x}_h]$.

Hierarchy-driven degradation and subspace perturbation together provide a structured model of both inter-hierarchy redistribution and intra-hierarchy drift, and expose the encoder to hierarchy variations for better text-time approximation, improving generalization under vision-imagery shift.

3.3 Hierarchy Preference Estimation

While hierarchy-driven perturbation improves robustness, variability across stimuli remains essential for imagery decoding. Different stimuli may rely on the visual hierarchy in different ways: some are reconstructed more effectively when structural signals are emphasized, while others benefit more from semantic guidance. To capture this variability, we estimate a **stimulus-wise hierarchy preference** (Figure 4).

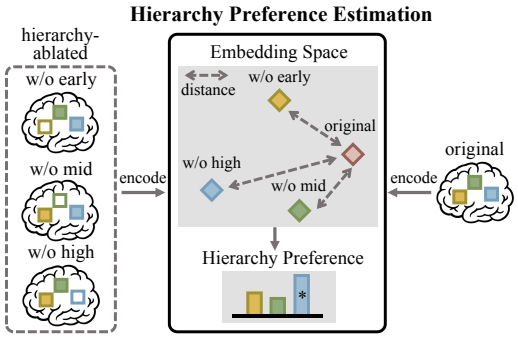


Figure 4: **Hierarchy preference estimation via representation sensitivity.** We construct hierarchy-ablated inputs by removing early-, mid-, or high-level components, and measure the representation changes to estimate sensitivity to each cortical level, which is normalized to obtain a simplex-valued hierarchy preference.

Given an input x, we aim to infer its hierarchy preference π, introduced in Equation (1),

$$\pi(x) = (\pi_e, \pi_m, \pi_h), \quad \pi \in \Delta^3,$$

which explicitly measures how strongly the predicted representation depends on each hierarchy.

We derive this signal with a pretrained reference model f_{θ^*} trained on clean vision data. For each hierarchy level $l \in \{e, m, h\}$, we construct an cortical hierarchy-ablated input $x_{\setminus l} = T_l(x)$. Here, T_l suppresses corresponding hierarchy component by setting the associated voxel subset to zero.

We then compare the reference-model predictions for the original and ablated inputs:

$$\Delta_l(x) = d\big(f_{\theta^*}(x),\ f_{\theta^*}(x_{\setminus l})\big), \quad l \in \{e, m, h\},$$

where $d(\cdot, \cdot)$ measures the discrepancy in representation space (e.g., cosine distance). We further normalize the sensitivity profile$(\Delta_e(x), \Delta_m(x), \Delta_h(x))$ to obtain the final hierarchy preference

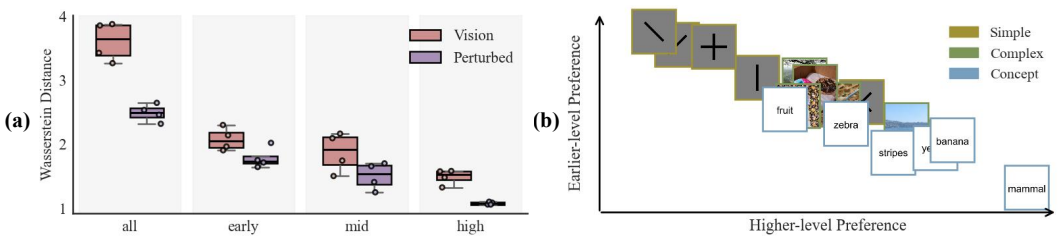


Figure 5: **Hierarchy-driven decoding analysis. (a) Effect of hierarchy-driven perturbation.** We measure the Wasserstein distance between vision and imagery representations using the full fMRI representation and different cortical hierarchy levels. The perturbation consistently reduces the distance to imagery representations, demonstrating improved alignment across all hierarchy levels. **(b) Results from hierarchy preference estimation.** We visualize the predicted hierarchy preferences across stimulus types. Simple stimuli rely more on early-level regions, whereas complex natural images and semantic concepts show greater dependence on higher-level regions, revealing meaningful stimulus-specific hierarchy preferences.

$\pi(x)$. Intuitively, a large change in the predicted representation after ablating a hierarchy group indicates that the corresponding cortical level is important for the current stimulus.

3.4 Hierarchy-Conditioned Reconstruction

We develop a hierarchy-conditioned reconstruction framework that first learns multimodal representations and then adapts the reconstruction process based on a stimulus-wise hierarchy preference.

Learning Hierarchy-Conditioned Representations. The framework consists of a reference model, a student model, and a hierarchy preference estimation model. Both the reference and student models adopt three ridge regression models, as illustrated in Figure 2(a), to map fMRI signals into text, image, and low-level representations, capturing complementary semantic and structural information. The reference model is trained on clean inputs to provide stable supervision, while the student model is trained with hierarchy-driven perturbations to improve robustness under distribution shift. We estimate hierarchy preference for each stimulus , and the preference estimation model is trained to predict this preference directly from the input signals.

Hierarchy-Conditioned Inference. At inference time, the student model produces text, image, and low-level representations. The inferred hierarchy preference is then transformed into control signals that regulate the relative contributions of semantic and structural information (Figure 2(b)). Intuitively, higher-level preference increases the strength of semantic guidance, while stronger early-level preference emphasizes structural anchoring. These signals are mapped to decoder-side coefficients that directly modulate the reconstruction process (see Appendix B for details). Specifically, semantic representations are adjusted to control the strength of high-level guidance, while the low-level pathway is scaled to preserve structural details.

Overall, the hierarchy preference defines a continuous reconstruction policy that dynamically reweights semantic and structural embeddings, enabling a stimulus-wise tradeoff between semantic fidelity and structural accuracy.

4 Experiments

4.1 Experimental Settings

Datasets. The Natural Scenes Dataset (**NSD**) [1] provides large-scale fMRI-image pairs from subjects viewing natural images from COCO [18]. Its imagery extension, **NSD-Imagery** [14], further contains three types of stimuli: *simple* (geometric shapes), *complex* (natural scenes), and *concept* (semantic concepts); and includes two evaluation settings: a *vision* setting, where subjects observe stimuli, and a *mental imagery* setting, where subjects are instructed to imagine the corresponding visual content. As the standard–and effectively the only–benchmark, NSD-Imagery is one of the very few public datasets providing paired visual perception and mental imagery fMRI data with

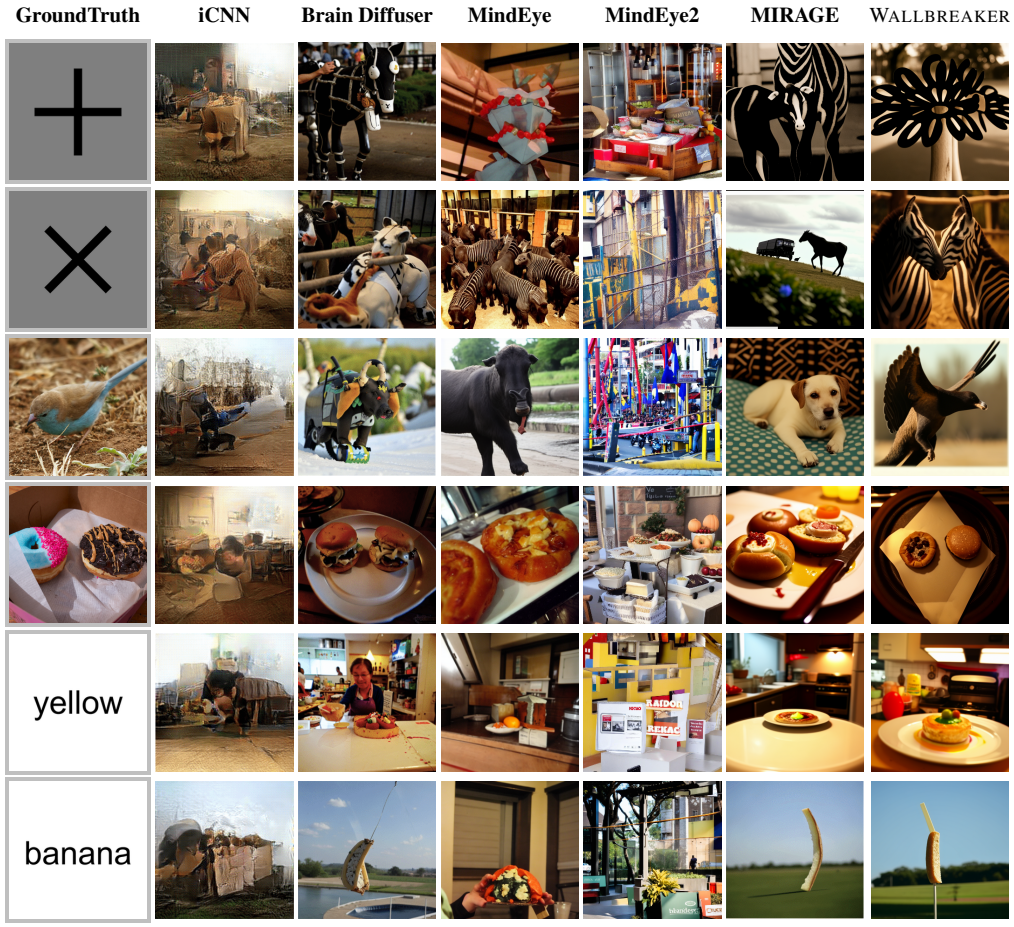


Figure 6: **Qualitative comparison on mental imagery reconstruction.** The first two rows correspond to *simple* stimuli (geometric shapes), the middle two rows correspond to *complex* stimuli (natural images), and the last two rows correspond to *concept* stimuli (semantic concepts).

natural-image stimuli. To enable the evaluation of vision-to-imagery generalization, we train our models exclusively on **NSD**, and evaluate them on **NSD-Imagery**.

Evaluation Metrics. We follow the standard protocol [15] and evaluate low-level fidelity, high-level semantics, and brain alignment. Low-level quality is measured by PixCorr, SSIM [42], and AlexNet accuracy (Alex(2), Alex(5)) [16]. High-level alignment is evaluated by Inception accuracy (Incep) [37], CLIP similarity (CLIP) [27], EfficientNet distance (Eff) [39], and SwAV features [3]. Brain alignment is assessed via correlations in early visual (Early Vis.), higher visual (Higher Vis.), and the overall visual cortex. Other experimental details, including architectures, hyperparameters, data processing, evaluation procedures, and computing costs, are provided in Appendices C and H.

4.2 HIERARCHY-DRIVEN DECODING ANALYSIS

To illustrate the effectiveness of WALLBREAKER's hierarchy modeling in imagery decoding, we first analyze core components: *hierarchy-driven perturbation* and *hierarchy preference estimation*.

Hierarchy-Driven perturbation. We analyze whether hierarchy-driven perturbation effectively reduces the distribution gap between vision and imagery representations. We measure the Wasserstein distance [40] between vision and imagery representations before and after applying perturbations, computed on both the full set of voxels and different cortical hierarchy levels (early, mid, and high). As shown in Figure 5(a), the distribution gap varies systematically across regions: early visual cortex exhibits the largest discrepancy, followed by mid-level areas, while higher-level regions show the smallest gap. After applying hierarchy-driven perturbation, the distance is consistently reduced

Method	Low-Level				High-Level				Brain Correlation		
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Early Vis.↑	Higher Vis.↑	Visual Cortex↑
Mental Imagery Reconstructions											
iCNN [35]	0.108	0.340	0.506	0.553	0.494	0.417	0.994	**0.560**	0.113	0.062	0.081
Brain Diffuser [25]	0.064	0.401	0.521	0.584	0.527	0.541	0.935	0.585	0.133	0.127	0.141
MindEye [33]	0.086	0.349	0.596	0.610	0.520	0.547	0.948	0.564	0.180	0.135	0.155
BrainRAM [43]	0.056	0.372	0.518	0.563	0.525	0.537	0.927	0.577	0.139	**0.145**	0.112
MindBridge [41]	0.030	0.200	0.441	0.525	0.456	0.497	0.980	0.627	0.079	0.079	0.075
NeuroPictor [13]	0.022	0.305	0.432	0.449	0.442	0.464	0.994	0.612	0.084	0.054	0.140
MindEye2 [34]	0.036	**0.414**	0.476	0.554	0.460	0.508	0.966	0.591	0.069	0.055	0.061
MIRAGE [15]	0.104	0.398	0.639	**0.625**	0.523	0.575	0.914	0.575	**0.204**	0.142	0.168
WALLBREAKER	**0.109**	0.401	**0.648**	0.620	**0.560**	**0.583**	**0.912**	0.572	0.193	0.143	**0.169**
Vision Reconstructions											
iCNN [35]	0.224	0.385	0.717	**0.814**	0.612	0.490	0.926	0.524	**0.442**	0.246	0.338
Brain Diffuser [25]	0.107	0.455	0.603	0.728	0.610	0.583	0.908	0.555	0.247	0.229	0.255
MindEye [33]	0.218	0.412	0.736	0.808	0.624	0.653	0.881	**0.510**	0.374	0.253	0.311
BrainRAM [43]	0.097	0.409	0.631	0.680	0.585	0.598	0.894	0.530	0.215	0.226	0.175
MindBridge [41]	0.117	0.352	0.585	0.708	0.588	0.645	0.915	0.565	0.245	0.227	0.232
NeuroPictor [13]	0.055	0.364	0.623	0.664	0.499	0.535	0.949	0.571	0.272	0.192	0.363
MindEye2 [34]	0.161	**0.480**	0.701	0.775	0.627	0.659	0.886	0.512	0.352	0.237	0.290
MIRAGE [15]	0.221	0.442	0.790	0.766	0.698	**0.667**	0.879	0.546	0.363	**0.262**	0.316
WALLBREAKER	**0.243**	0.441	**0.832**	0.799	**0.695**	0.657	**0.874**	0.539	0.410	0.237	**0.372**

Table 1: **Quantitative comparison on mental imagery and vision reconstruction.** Bold and underlined values indicate the best and second-best results, respectively, for each metric.

Components		Low-Level				High-Level				Brain Correlation		
Hier. Deg.	Sub. Pert.	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Early Vis.↑	Higher Vis.↑	Visual Cortex↑
Mental Imagery Reconstruction												
		0.099	0.398	0.637	0.613	0.534	0.555	0.913	0.575	0.194	0.140	0.164
✓		0.104	**0.403**	0.643	0.609	0.547	0.583	0.913	0.573	**0.206**	0.141	0.168
	✓	0.105	0.398	0.632	0.616	0.523	**0.592**	0.912	**0.571**	0.198	0.138	0.162
✓	✓	**0.109**	0.401	**0.648**	**0.620**	**0.560**	0.583	**0.912**	0.572	0.193	**0.143**	**0.169**
Vision Reconstruction												
		0.219	0.441	0.794	0.770	0.686	**0.665**	0.884	0.543	0.364	0.224	0.312
✓		0.236	**0.442**	0.811	0.778	**0.699**	0.632	0.875	0.544	0.396	0.234	0.314
	✓	0.232	**0.442**	0.801	0.775	0.680	0.637	0.879	0.543	0.387	**0.240**	0.314
✓	✓	**0.243**	0.441	**0.832**	**0.799**	0.695	0.657	**0.874**	**0.539**	**0.410**	0.237	**0.372**

Table 2: **Ablation of hierarchy-driven perturbation under mental imagery and vision settings.** "Hier. Deg." denotes hierarchy degradation, and "Sub. Pert." denotes subspace perturbation.

across all regions. This indicates that our perturbation strategy aligns vision representations with imagery distributions, particularly by mitigating the mismatch in early and mid-level representations.

Hierarchy Preference Estimation. In Figure 5(b), we further examine whether the predicted hierarchy preference captures meaningful stimulus-dependent structure. Simple stimuli (e.g., geometric shapes) favor earlier-level representations and low-level structure, whereas complex natural images and semantic concepts favor higher-level representations and semantic information. Held-out validation and its association with reconstruction gains are reported in Appendix E.

4.3 VISION AND IMAGERY RECONSTRUCTION RESULTS

We compare WALLBREAKER with recent state-of-the-art approaches on both mental imagery and visual perception reconstruction in Table 1. Our method achieves strong performance across low-level fidelity, semantic alignment, and brain correlation metrics, while reducing the gap between vision and imagery decoding. On mental imagery, our method achieves the best performance on several key metrics, including PixCorr, Inception, CLIP similarity, and overall visual cortex correlation, indicating improved semantic consistency under top-down-driven conditions. On visual perception, our method also attains competitive or best results across most metrics, including PixCorr and AlexNet accuracy, while maintaining strong semantic alignment and visual cortex correlation. Detailed participant-wise results and paired uncertainty estimates are provided in Appendix D.

Setting	Low-Level				High-Level				Brain Correlation		
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Early Vis.↑	Higher Vis.↑	Visual Cortex↑
Mental Imagery Reconstruction											
No Control	0.099	0.398	0.637	0.613	0.534	0.555	0.913	0.575	0.194	0.140	0.164
Fixed Control	0.108	0.399	0.629	0.600	0.565	**0.596**	0.915	0.573	**0.204**	0.132	0.159
Preference Only	0.106	0.400	0.623	0.605	**0.570**	0.577	0.916	0.573	0.189	0.135	0.156
Hier.-Cond. Control	**0.109**	**0.401**	**0.648**	**0.620**	0.560	0.583	**0.912**	**0.572**	0.193	**0.143**	**0.169**
Vision Reconstruction											
No Control	0.219	0.441	0.794	0.770	0.686	**0.665**	0.884	0.543	0.364	0.224	0.312
Fixed Control	0.239	**0.445**	0.809	0.776	0.674	0.625	0.884	0.542	0.398	0.232	0.313
Preference Only	0.225	0.437	0.798	0.768	0.689	0.648	0.881	0.540	0.355	**0.256**	0.310
Hier.-Cond. Control	**0.243**	0.441	**0.832**	**0.799**	**0.695**	0.657	**0.874**	**0.539**	**0.410**	0.237	**0.372**

Table 3: **Ablation of hierarchy-conditioned reconstruction strategies under mental imagery and vision settings.** "No Control" disables decode-time modulation; "Fixed Control" uses constant coefficients; "Preference Only" uses $\pi(x)$ without adaptive mapping; and "Hier.-Cond. Control" applies the full hierarchy-conditioned control.

We further provide visual comparisons of imagery reconstruction results across stimulus types in Figure 6, including *simple* geometric shapes, *complex* natural images, and *concept* stimuli. For simple stimuli, our method better preserves low-level geometric structures. For complex stimuli, our reconstructions exhibit improved semantic and structural consistency; for example, when the target stimulus is a bird, competing methods may generate other animals, while our method consistently reconstructs a bird. For concept stimuli, our method also produces images that better match the intended semantics, such as generating a banana for the concept "banana." These results demonstrate that our approach achieves both stronger structural and semantic alignment across stimulus types.

4.4 Ablation Studies

Hierarchy-Driven Perturbation. Table 2 shows that both hierarchy degradation and subspace perturbation contribute to performance. On mental imagery, hierarchy degradation mainly improves structural fidelity (SSIM, Early Vis.), while subspace perturbation benefits semantic alignment (CLIP, SwAV). Combining both yields the best overall results across most metrics. A similar trend holds for vision reconstruction, where the full model achieves the strongest and most balanced performance, indicating that the two components are complementary.

Hierarchy-Conditioned Control. Table 3 reports the effect of different control strategies. Using either fixed control or preference alone improves performance over no control, indicating the importance of incorporating hierarchy information at inference. However, both variants exhibit trade-offs, with gains in either structural or semantic metrics but not both. In contrast, the full hierarchy-conditioned control consistently achieves the best or most balanced performance across metrics, especially on PixCorr, AlexNet accuracy, and visual cortex correlation. This demonstrates that combining stimulus-wise preference with adaptive control is crucial for effective reconstruction. The hyperparameter and grouping sensitivity analyses in Appendix F further confirm that the default configuration provides balanced performance without tuning.

5 Conclusion

We presented WallBreaker, a framework for robust and adaptive fMRI-to-image decoding that explicitly models cortical hierarchy. We formulated the gap between visual perception and mental imagery as a hierarchy-structured representation shift, coupled with stimulus-dependent hierarchy preference. To address these challenges, we introduced hierarchy-driven representation perturbation to improve robustness under domain shift, together with hierarchy-conditioned reconstruction to adapt decoding according to stimulus-wise hierarchy preference. Experiments on NSD and NSD-Imagery demonstrate that our approach improves key reconstruction metrics on mental imagery while maintaining strong performance on visual perception reconstruction. Ablation studies further show that both perturbation and control contribute substantially and complement one another. More broadly, our results highlight the importance of explicitly modeling cortical hierarchy for bridging vision-to-imagery transfer and enabling robust, adaptive neural decoding.

REFERENCES

[1] 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.

[2] Roman Beliy, Amit Zalcher, Jonathan Kogman, navve wasserman, and michal Irani. Brain-it: Image reconstruction from fmri via brain-interaction transformer. In *International Conference on Learning Representations*, volume 2026, pp. 124487–124521, 2026.

[3] 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.

[4] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific Data*, 6(1):49, 2019.

[5] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22710–22720. IEEE, 2023.

[6] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

[7] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[8] Nadine Dijkstra, Peter Zeidman, Sasha Ondobaka, Marcel AJ van Gerven, and Karl Friston. Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific Reports*, 7(1):5677, 2017.

[9] Nadine Dijkstra, Sander E Bosch, and Marcel AJ Van Gerven. Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences*, 23(5):423–434, 2019.

[10] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

[11] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017.

[12] 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 *International Conference on Learning Representations*, volume 2026, pp. 88145–88169, 2026.

[13] Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, pp. 56–73. Springer, 2024.

[14] Reese Kneeland, Paul S Scotti, Ghislain St-Yves, Jesse Breedlove, Kendrick Kay, and Thomas Naselaris. Nsd-imagery: A benchmark dataset for extending fmri vision decoding methods to mental imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28852–28862. IEEE, 2025.

[15] Reese Kneeland, Cesar Kadir Torrico Villanueva, Tong Chen, Jordyn Ojeda, Shubh Khanna, Jonathan Xu, Paul S Scotti, and Thomas Naselaris. Mirage: Robust multi-modal architectures translate fmri-to-image models from vision to mental imagery. *PLoS computational biology*, 22(5):e1014263, 2026.

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

[17] 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.

[18] 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.

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26286–26296. IEEE, 2024.

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

[22] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.

[23] Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

[24] Alexander Olza, David Soto, and Roberto Santana. Domain adaptation-enhanced searchlight: Enabling brain decoding from visual perception to mental imagery. *CoRR*, 2024.

[25] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

[26] Joel Pearson. The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, 20(10):624–634, 2019.

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

[28] Edmund T Rolls. Two what, two where, visual cortical streams in humans. *Neuroscience & Biobehavioral Reviews*, 160:105650, 2024.

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10674–10685. ieee, 2022.

[30] Tiasha Saha Roy, Jesse Breedlove, Ghislain St-Yves, Kendrick Kay, and Thomas Naselaris. Comparison of signal to noise in vision and imagery for qualitatively different kinds of stimuli. *Journal of Vision*, 23(9):5961–5961, 2023.

[31] Tiasha Saha Roy, Jesse Breedlove, Ghislain St-Yves, Kendrick Kay, and Thomas Naselaris. A transformation from vision to imagery in the human brain. *bioRxiv*, 2025.

[32] T Saha Roy, Jesse Breedlove, Ghislain St-Yves, Kendrick Kay, and Thomas Naselaris. Mental imagery: Weak vision or compressed vision. In *2023 Conference on Cognitive Computational Neuroscience*, volume 10, pp. 2023–1693, 2023.

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

[34] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *International Conference on Machine Learning*, 2024.

[35] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15(1):e1006633, 2019.

[36] Fabrizio Spera, Tommaso Boccato, Michal Olak, Sara Cammarota, Matteo Ciferri, Michelangelo Tronti, Nicola Toschi, and Matteo Ferrante. Seeing the imagined: a latent functional alignment in visual imagery decoding from fmri data. *arXiv preprint arXiv:2604.15374*, 2026.

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

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

[39] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PmLR, 2019.

[40] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

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

[42] 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.

[43] Dian Xie, Peiang Zhao, Jiarui Zhang, Kangqi Wei, Xiaobao Ni, and Jiong Xia. Brainram: Cross-modality retrieval-augmented image reconstruction from human brain activity. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3994–4003, 2024.

[44] Shihao Yang, Xiying Huang, Danilo Bernardo, Jun-En Ding, Andrew Michael, Jingmei Yang, Patrick Kwan, Ashish Raj, and Feng Liu. Foundation and large-scale ai models in neuroscience: A comprehensive review. *arXiv preprint arXiv:2510.16658*, 2025.

[45] Sijin Yu, Zijiao Chen, Wenxuan Wu, Shengxian Chen, Zhongliang Liu, Jingxin Nie, Xiaofen Xing, Xiangmin Xu, and Xin Zhang. From flat to round: Redefining brain decoding with surface-based fmri and cortex structure. In *IEEE/CVF International Conference on Computer Vision Workshops*, pp. 6248–6257. IEEE, 2025.

[46] Sijin Yu, Zijiao Chen, Zhenyu Yang, Zihao Tan, Jiakun Xu, Zhongliang Liu, Shengxian Chen, Wenxuan Wu, Xiangmin Xu, and Xin Zhang. Neurips: Neuro-anatomical inductive priors for sphere-based brain decoding. In *International Conference on Machine Learning*, 2026.

[47] Xu Zhang, Ruijie Quan, Wenguan Wang, and Yi Yang. Moving beyond diffusion: Hierarchy-to-hierarchy autoregression for fmri-to-image reconstruction. In *International Conference on Learning Representations*, volume 2026, pp. 68490–68515, 2026.

Appendix

A Limitations and Future Work 13

B Details for Hierarchy-Conditioned Reconstruction 14

C Implementation Details 14

D Subject-wise Results and Paired Uncertainty 15

E Hierarchy-Preference Validation 17

F Hyperparameter and Grouping Sensitivity 17

G Multi-subject and LOSO Generalization 18

H Computational Cost 19

AI USE STATEMENT

We used large language models (LLMs) to assist with writing and editing, including improving clarity, grammar, and conciseness. We also used LLMs for literature retrieval and discovery, such as identifying potentially missing related work. All technical content and results are original contributions of the authors, and the authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Experimental details including architectures, hyperparameters, data processing, and evaluation procedures are detailed in Appendix C. Further discussion and comparison on computational cost are provided in Appendix H. The source code for reproducing the reported experiments is submitted together with the manuscript.

A LIMITATIONS AND FUTURE WORK

Data-driven hierarchy grouping. Although our method substantially improves mental imagery reconstruction, imagery decoding remains considerably more challenging than visual perception decoding, and there is still a noticeable performance gap between the two settings. In particular, reconstruction quality can degrade for difficult cases involving complex background textures, fine-grained spatial details, or highly compositional scenes. In addition, the current hierarchy modeling is based on predefined cortical groupings and a simplified hierarchy decomposition, which may not fully capture the complex and dynamic interactions across visual regions. Future work could explore more fine-grained or data-driven hierarchy representations, as well as adaptive hierarchy structures that better reflect subject-specific neural organization and stimulus-dependent cortical dynamics.

Cross-subject generalization. Due to the substantial and fundamental differences across subjects in cortical anatomy, voxel organization, and functional responses, current SOTA models for fMRI-based reconstruction (e.g., MIRAGE [15], MindEye [33, 34], MindDiffuser [21]) are all designed to be subject-specific, i.e., trained and evaluated on individual participants. In Appendix G, we initially explored WALLBREAKER's generalizability in multi-subject training and unseen-subject generalizability. Although beyond the scope of our work, WALLBREAKER still demonstrated superior performance to competing methods. That being said, the consistently visible performance drop in all methods compared to their performance in subject-specific reconstruction indicates that cross-subject generalization, although desirable, remains a particularly challenging task.

Large-scale and independent evaluation. A population-level assessment would further strengthen the evidence for our method's superiority and its robustness. However, such a larger-scale evaluation

is not currently feasible because of limited data availability: to the best of our knowledge, NSD-Imagery is one of the very few datasets providing paired visual perception and mental imagery fMRI data with natural-image stimuli. And it remains the standard–and effectively the only–benchmark used by state-of-the-art studies.

B Details for Hierarchy-Conditioned Reconstruction

We provide additional implementation details for model training and hierarchy-conditioned reconstruction.

Model Training. The framework consists of a reference model f_{θ^*}, a student model f_θ, and a hierarchy preference model g_ϕ. Following [15], both f_{θ^*} and f_θ use three ridge regression models to map fMRI signals into low-level, image-level, and text-level representations:

$$f_\theta(x) = \big(y_{\text{low}},\ y_{\text{image}},\ y_{\text{text}}\big),$$

where y_{low} captures structural information and $y_{\text{image}}, y_{\text{text}}$ encode high-level semantics.

For the low-level pathway, we use a pretrained Very Deep Variational Autoencoder (VDVAE) [6] to extract structural features from the first 31 latent layers, producing a 1×91168 feature vector. During inference, the predicted features are decoded through the VDVAE decoder to provide structural guidance for reconstruction.

For image-level semantics, we use the CLS token from CLIP ViT-L/14 [27], resulting in a compact 1×768 image representation. Compared to high-dimensional spatial ViT embeddings, this representation provides more stable semantic guidance for imagery decoding.

For text-level semantics, we use unpooled CLIP ViT-bigG/14 text embeddings of dimension 77×1280. To improve text supervision quality, we replace the original COCO captions with more descriptive captions generated by LLaVA v1.5-13B [19, 20].

The reference model is trained on clean inputs, while the student model is trained with hierarchy-driven perturbations to improve robustness. To estimate stimulus-wise hierarchy dependence, we measure representation sensitivity under hierarchy-level ablations:

$$\Delta_g(x) = d\big(f_{\theta^*}(x), f_{\theta^*}(x_{\setminus l})\big), \quad l \in \{e, m, h\},$$

and normalize $(\Delta_e, \Delta_m, \Delta_h)$ to obtain the hierarchy preference $\pi(x)$. The model g_ϕ is trained to predict $\pi(x)$ from clean inputs.

Hierarchy-conditioned reconstruction. At inference time, the student model produces conditional representations $(y_{\text{text}}, y_{\text{image}}, y_{\text{low}})$. The inferred hierarchy preference $\pi(x) = (\pi_e, \pi_m, \pi_h)$ is transformed into continuous control signals that regulate semantic and structural contributions.

We first define a semantic score s and a structural score r as projections of $\pi(x)$:

$$s = \mathcal{N}(-\alpha_e\pi_e + \alpha_m\pi_m + \alpha_h\pi_h)\,, \quad r = \mathcal{N}(\beta_e\pi_e + \beta_m\pi_m - \beta_h\pi_h)\,,$$

where $\mathcal{N}(\cdot)$ maps values to $[0, 1]$. To account for mid-level effects, we further define an image-semantic control:

$$i = \text{clip}\big(s + \gamma\left(\pi_m - \tfrac{1}{3}\right),\ 0,\ 1\big)\,,$$

where $\frac{1}{3}$ corresponds to the uniform hierarchy prior.

These signals are mapped to decoder-side coefficients $(\lambda_{\text{text}}, \lambda_{\text{image}}, \lambda_{\text{low}})$, which directly modulate the reconstruction process. In particular, semantic conditions are interpolated with unconditional embeddings, while the structural pathway is scaled to control low-level anchoring.

C Implementation Details

Datasets. We train models for subjects 1, 2, 5, and 7, as only these subjects completed the full 30k trials of the NSD experiment [1]. We apply the provided `nsdgeneral` voxel mask at 1.8 mm resolution to the preprocessed fMRI signals, covering visual regions from early visual cortex to higher-level visual areas. To evaluate imagery decoding, we use the NSD-Imagery benchmark [14],

where the same subjects additionally perform mental imagery trials. The dataset contains three types of stimuli: geometric shapes, natural scenes, and semantic concepts. All scanning and preprocessing procedures follow the original NSD protocol.

Training Details. Both the reference and student models use three ridge regression models to predict low-level, image-level, and text-level representations. The student model is trained with hierarchy-driven perturbations, while the hierarchy preference model is trained using sensitivity-derived targets computed from hierarchy-ablated inputs.

For hierarchy-driven degradation, we sample degradation coefficients from hierarchy-dependent Beta distributions: $s_l \sim \mathrm{Beta}(a_l, b_l), l \in \{e, m, h\}$, where the parameters are chosen to impose different attenuation biases across hierarchy levels. In practice, we use $(a_e, b_e) = (1, 3), (a_m, b_m) = (2, 2), (a_h, b_h) = (5, 1)$, such that lower-level representations are biased toward stronger attenuation, while higher-level representations are more likely to preserve their original signal strength.

For subspace perturbation, we estimate low-rank ROI subspaces using singular value decomposition (SVD) on centered training representations and retain up to 8 dominant components for each hierarchy group. We implement subspace rotation as a near-identity orthogonal perturbation in the low-rank space, with rotation strength set to 0.35.

Hierarchy Grouping. We group visual regions into early-, mid-, and high-level hierarchy components following established cortical hierarchy organization [10, 28]. Early-level regions mainly include primary and secondary visual cortex areas (V1–V3). Mid-level regions include intermediate visual areas associated with shape and object processing, such as hV4, VO, LO, TO, and V3A/B regions. High-level regions consist of category-selective and higher-order association areas, including face-, place-, body-selective cortices, and parietal and attentional regions (e.g., IPS, SPL, and FEF).

Experimental Setup. All experiments are conducted on a single NVIDIA RTX PRO 6000 GPU with CUDA 13.0 and PyTorch 2.10.0.

D Subject-wise Results and Paired Uncertainty

We report subject-wise comparisons and participant-level paired uncertainty estimates for the four NSD-Imagery participants. Because participants are the independent experimental units, aggregate performance alone may obscure inter-participant variability or improvements driven by a single participant. WALLBREAKER and MIRAGE are therefore trained and evaluated separately for each participant following the standard subject-specific protocol.

Table 4 reports the complete subject-wise results. For mental imagery, WALLBREAKER improves PixCorr, AlexNet(2), CLIP, SwAV, and overall visual-cortex correlation for all four participants. The agreement in improvement direction suggests that the aggregate gains are not driven by a single participant. Performance on the remaining metrics is more variable, particularly for the individual early- and higher-visual correlations.

Under visual perception, WALLBREAKER also shows consistent gains in PixCorr, AlexNet(2), AlexNet(5), and overall visual-cortex correlation, while the differences on several high-level metrics are mixed across participants. Across both imagery and vision settings, WALLBREAKER performs better in 37 of the 44 subject–metric comparisons. This count is reported only as a descriptive summary because the reconstruction metrics are correlated and should not be treated as independent statistical observations.

To quantify uncertainty while respecting the paired experimental design, in Table 5, we compute the within-participant difference between WALLBREAKER and MIRAGE and bootstrap participants with replacement while preserving the method pairing. We define the direction-normalized improvement Δ_m as $m_{\text{WALLBREAKER}} - m_{\text{MIRAGE}}$ for higher-is-better metrics and reverse the difference for lower-is-better metrics, such that $\Delta_m > 0$ always favors WALLBREAKER.

For mental imagery, the paired confidence intervals support positive effects on PixCorr, AlexNet(2), CLIP, and SwAV. The mean overall visual-cortex correlation also increases consistently, although the estimated effect is small. In contrast, the intervals for the remaining metrics overlap zero, indicating greater uncertainty in their improvement. For visual perception, the clearest paired gains occur in PixCorr, AlexNet(2), AlexNet(5), EfficientNet distance, early-visual correlation, and overall

Subj.	Method	Low-Level				High-Level				Brain Correlation		
		PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Early Vis.↑	Higher Vis.↑	Visual Cortex↑
Mental Imagery Reconstructions												
1	MIRAGE	0.119	0.402	0.662	0.599	0.622	0.638	0.896	0.566	0.208	0.170	0.157
	WALLBREAKER	**0.120**	**0.404**	**0.671**	**0.620**	**0.651**	**0.646**	**0.895**	**0.565**	**0.209**	**0.174**	**0.158**
2	MIRAGE	0.117	0.416	0.664	0.630	0.584	0.609	0.909	0.565	0.231	0.198	0.200
	WALLBREAKER	**0.118**	0.415	**0.665**	**0.631**	**0.668**	**0.631**	**0.895**	**0.556**	**0.232**	**0.200**	**0.201**
5	MIRAGE	0.103	0.392	0.680	0.701	0.499	0.580	0.919	0.566	0.169	0.071	0.157
	WALLBREAKER	**0.105**	**0.401**	**0.681**	0.658	0.489	**0.581**	**0.917**	**0.565**	0.137	**0.072**	**0.158**
7	MIRAGE	0.077	0.382	0.550	0.570	0.387	0.473	0.932	0.603	0.208	0.129	0.158
	WALLBREAKER	**0.093**	**0.384**	**0.575**	**0.571**	**0.432**	**0.474**	0.941	**0.602**	0.194	0.126	**0.159**
Vision Reconstructions												
1	MIRAGE	0.267	0.452	0.841	0.791	0.689	0.676	0.862	0.538	0.352	0.239	0.320
	WALLBREAKER	**0.289**	**0.453**	**0.866**	**0.834**	**0.690**	**0.677**	**0.861**	**0.522**	**0.399**	0.203	**0.382**
2	MIRAGE	0.215	0.448	0.785	0.754	0.653	0.665	0.883	0.547	0.435	0.364	0.367
	WALLBREAKER	**0.243**	**0.452**	**0.836**	**0.784**	**0.654**	0.602	**0.875**	**0.546**	**0.469**	0.298	**0.434**
5	MIRAGE	0.182	0.429	0.793	0.769	0.709	0.657	0.890	0.553	0.345	0.234	0.309
	WALLBREAKER	**0.206**	0.419	**0.802**	**0.803**	**0.731**	**0.658**	**0.882**	**0.541**	**0.402**	**0.235**	**0.349**
7	MIRAGE	0.220	0.439	0.741	0.750	0.741	0.670	0.881	0.546	0.320	0.211	0.268
	WALLBREAKER	**0.234**	**0.440**	**0.824**	**0.775**	0.705	**0.691**	**0.878**	0.547	**0.370**	**0.212**	**0.323**

Table 4: **Subject-wise reconstruction results.** WALLBREAKER and MIRAGE are trained and evaluated separately for each participant following the standard subject-specific protocol. Bold values indicate that WALLBREAKER outperforms MIRAGE for the corresponding participant and metric. Higher is better for all metrics except Eff and SwAV.

Metric	Mental Imagery			Vision		
	MIRAGE	WALLBREAKER	Paired 95% CI	MIRAGE	WALLBREAKER	Paired 95% CI
Low-Level						
PixCorr↑	0.104±0.019	**0.109±0.013**	[**+0.001**, **+0.012**]	0.221±0.035	**0.243±0.034**	[**+0.014**, **+0.028**]
SSIM↑	0.398±0.015	**0.401±0.013**	[−0.001, +0.009]	0.442±0.010	0.441±0.016	[−0.010, +0.004]
Alex(2)↑	0.639±0.060	**0.648±0.049**	[**+0.001**, **+0.019**]	0.790±0.041	**0.832±0.027**	[**+0.009**, **+0.083**]
Alex(5)↑	0.625±0.056	0.620±0.036	[−0.043, +0.021]	0.766±0.019	**0.799±0.026**	[**+0.025**, **+0.043**]
High-Level						
Incep↑	0.523±0.104	**0.560±0.117**	[−0.010, +0.084]	0.698±0.037	0.695±0.032	[−0.036, +0.022]
CLIP↑	0.575±0.072	**0.583±0.078**	[**+0.001**, **+0.017**]	0.667±0.008	0.657±0.039	[−0.063, +0.021]
Eff↓	0.914±0.015	**0.912±0.022**	[−0.009, +0.014]	0.879±0.012	**0.874±0.009**	[**+0.001**, **+0.008**]
SwAV↓	0.575±0.019	**0.572±0.020**	[**+0.001**, **+0.007**]	0.546±0.006	**0.539±0.012**	[−0.001, +0.016]
Brain Correlation						
Early Vis.↑	0.204±0.026	0.193±0.040	[−0.032, +0.001]	0.363±0.050	**0.410±0.042**	[**+0.034**, **+0.057**]
Higher Vis.↑	0.142±0.055	**0.143±0.056**	[−0.003, +0.004]	0.262±0.069	0.237±0.043	[−0.066, +0.001]
Visual Cortex↑	0.168±0.021	**0.169±0.021**	[0.000, +0.002]	0.316±0.041	**0.372±0.048**	[**+0.040**, **+0.067**]

Table 5: **Participant-level paired uncertainty.** Results are reported as mean ± standard deviation across the four participants. Paired 95% bootstrap confidence intervals are computed for the direction-normalized improvement Δ_m, such that positive values favor WALLBREAKER for both higher-is-better and lower-is-better metrics. Bold WALLBREAKER values outperform MIRAGE, and bold confidence intervals exclude zero.

visual-cortex correlation. These results support improvements on several complementary structural, perceptual, and brain-alignment measures, rather than a uniform advantage on every metric.

As a complementary nonparametric analysis, we apply exact one-sided paired Wilcoxon signed-rank tests across participants. For mental imagery, PixCorr, AlexNet(2), CLIP, SwAV, and overall visual-cortex correlation each attain $p = 0.0625$, the smallest possible one-sided value with four nonzero paired observations. Given the limited number of participants, we do not interpret these tests as establishing conventional population-level significance. Together, the subject-wise results, paired confidence intervals, and exact tests instead provide evidence that the principal improvements are directionally consistent within the available NSD-Imagery cohort.

E HIERARCHY-PREFERENCE VALIDATION

The hierarchy-conditioned reconstruction module relies on the estimated stimulus-wise hierarchy preference. We therefore conduct two complementary analyses to examine whether the preference estimator (i) generalizes to held-out sensitivity targets and (ii) provides a signal associated with reconstruction improvements.

Held-out hierarchy-preference prediction. As described in Section 3.3, the target hierarchy preference is derived from the sensitivity of the reference model to hierarchy-level ablations. For each hierarchy level $l \in \{e, m, h\}$, we compute

$$\Delta_l(x) = d\big(f_{\theta^*}(x), f_{\theta^*}(x_{\setminus l})\big)\,,$$

where $x_{\setminus l}$ denotes the input with hierarchy level l ablated. The normalized sensitivity profile $(\Delta_e, \Delta_m, \Delta_h)$ defines the target preference $\pi(x)$, and the estimator g_ϕ predicts $\widehat{\pi}(x)$ directly from the complete fMRI input.

Metric	Performance
Pearson correlation ↑	0.5802
Spearman correlation ↑	0.6026
MAE ↓	0.0309
R^2 score ↑	0.5307

Table 6: **Held-out hierarchy-preference prediction.**

We evaluate the agreement between $\widehat{\pi}(x)$ and $\pi(x)$ on samples not used to train the estimator. As shown in Table 6, the estimator achieves a Pearson correlation of 0.5802, a Spearman correlation of 0.6026, an MAE of 0.0309, and an R^2 of 0.5307. These results indicate that the estimator generalizes beyond its training samples and captures a substantial portion of the variation in the sensitivity-derived hierarchy preferences.

Association with reconstruction gains. We further examine whether more accurate preference prediction is associated with larger improvements over the component-matched baseline.

For each evaluation unit, we relate hierarchy-preference prediction accuracy to the corresponding reconstruction gain. Table 7 reports the resulting Pearson correlations.

The strongest associations occur for the brain-correlation metrics, including Early Visual ($r = 0.8305$), Higher Visual ($r = 0.8407$), and overall Visual Cortex ($r = 0.8370$). Preference-prediction accuracy is also positively associated with improvements in PixCorr ($r = 0.6704$), Inception ($r = 0.6422$), and CLIP similarity ($r = 0.7537$).

These results provide complementary evidence for the hierarchy-preference estimator. The held-out analysis shows that it can recover sensitivity-derived preferences for unseen samples, while the association analysis suggests that more accurate preference estimates coincide with larger semantic and brain-alignment gains. This pattern is consistent with the intended role of hierarchy preference in adaptively balancing structural and semantic conditions during reconstruction. Because the association analysis is observational, however, it should not be interpreted as establishing a causal relationship between preference-prediction accuracy and reconstruction quality.

Reconstruction Gain	Pearson r
Low-Level	
ΔPixCorr	**0.6704**
ΔSSIM	−0.1156
ΔAlex(2)	−0.0947
ΔAlex(5)	0.2496
High-Level	
ΔInception	**0.6422**
ΔCLIP	**0.7537**
ΔEff	0.4094
ΔSwAV	0.4513
Brain Correlation	
ΔEarly Vis.	**0.8305**
ΔHigher Vis.	**0.8407**
ΔVisual Cortex	**0.8370**

Table 7: **Preference accuracy and reconstruction gains.** Bold denotes $r \geq 0.6$.

F HYPERPARAMETER AND GROUPING SENSITIVITY

We evaluate the sensitivity of WALLBREAKER to four design choices: subspace rotation strength, retained PCA rank, hierarchy-dependent perturbation schedule, and cortical hierarchy grouping. All configurations use the same training and evaluation protocol and are shared across participants without subject-specific retuning. Unless otherwise specified, the remaining components are fixed to the default configuration used in the main experiments.

Table 8 reports mental-imagery reconstruction performance under the evaluated configurations. Overall, the method remains stable across a range of settings, although individual hyperparameters induce different trade-offs among structural fidelity, semantic similarity, and brain alignment.

Factor	Setting	Reconstruction			Visual Cortex↑
		SSIM↑	Inception↑	CLIP↑	
Rotation strength	0	0.399	0.517	0.566	0.158
	0.20	0.398	0.525	**0.609**	0.155
	0.35 (default)	**0.401**	**0.560**	0.583	**0.169**
	0.50	0.397	0.527	0.601	0.154
PCA rank	4	0.397	0.534	**0.602**	0.167
	8 (default)	**0.401**	**0.560**	0.583	**0.169**
	16	0.398	0.530	0.563	0.151
Beta schedule	Uniform	0.394	0.509	0.591	0.153
	Default	0.401	**0.560**	0.583	**0.169**
	Strong	**0.402**	0.497	**0.602**	0.163
Hierarchy grouping	Early / Mid+High	0.399	0.548	0.576	0.158
	Early+Mid / High	0.396	0.535	0.579	0.161
	Early / Mid / High (default)	**0.401**	**0.560**	**0.583**	**0.169**

Table 8: **Sensitivity of WALLBREAKER to hyperparameter and hierarchy-grouping choices on mental imagery reconstruction.** Bold values indicate the best result within each sensitivity block. The default configuration is shared across participants and provides the most balanced performance across structural, semantic, and brain-alignment metrics.

Subspace-rotation strength. Removing subspace rotation reduces both CLIP similarity and visual-cortex correlation, supporting the contribution of intra-hierarchy perturbation. A rotation strength of 0.20 achieves the highest CLIP score, whereas the default strength of 0.35 provides the best SSIM, Inception, and visual-cortex correlation. Stronger rotation preserves semantic similarity but weakens brain alignment, indicating a trade-off rather than a uniformly optimal strength.

Retained PCA rank. Performance remains relatively stable for ranks 4, 8, and 16. Rank 4 achieves the highest CLIP similarity, while the default rank of 8 provides the strongest balance across SSIM, Inception, and visual-cortex correlation. Increasing the rank to 16 provides no further benefit, suggesting that the dominant low-rank components capture the information required by the perturbation mechanism.

Perturbation schedule. The hierarchy-dependent Beta schedules produce different balances between structural and semantic metrics. The strong schedule obtains the highest SSIM and CLIP scores but reduces Inception accuracy, whereas the uniform schedule underperforms the default on most reported metrics. The default schedule therefore offers the most balanced reconstruction and brain-alignment performance.

Hierarchy grouping. The default early/mid/high decomposition outperforms both coarser two-group alternatives on all reported metrics. Merging mid-level regions with either the early- or high-level group consistently reduces performance, supporting the treatment of mid-level cortex as a distinct hierarchy component.

Overall, these results show that WALLBREAKER is not dependent on a single narrow hyperparameter setting. We use the default configuration because it provides the most balanced performance across structural, semantic, and brain-alignment metrics.

G MULTI-SUBJECT AND LOSO GENERALIZATION

Standard fMRI-to-image reconstruction is typically evaluated with subject-specific models because cortical anatomy and functional organization vary substantially across participants. We additionally examine whether the advantage of WALLBREAKER over the component-matched MIRAGE baseline persists under multi-subject training and generalizes to unseen participants.

Evaluation protocols. We consider three training protocols:

- **Subject-specific:** a separate model is trained and evaluated for each participant.
- **Multi-subject:** a single model is trained jointly on data from all four participants and evaluated on each participant.

Protocol	Method	Reconstruction			Visual Cortex↑
		SSIM↑	Alex(2)↑	CLIP↑	
Subject-specific	MIRAGE	0.398±0.015	0.639±0.060	0.575±0.072	0.168±0.021
	WALLBREAKER	**0.401±0.013**	**0.648±0.049**	**0.583±0.078**	**0.169±0.021**
Multi-subject	MIRAGE	0.392±0.011	0.489±0.097	0.501±0.048	0.137±0.035
	WALLBREAKER	**0.396±0.008**	**0.584±0.094**	**0.517±0.038**	**0.157±0.028**
LOSO	MIRAGE	0.355±0.030	0.457±0.085	**0.451±0.089**	0.065±0.030
	WALLBREAKER	**0.367±0.017**	**0.462±0.080**	0.449±0.046	**0.073±0.027**

Table 9: **Cross-subject generalization on mental imagery reconstruction.** We compare WALLBREAKER with the component-matched MIRAGE baseline under subject-specific, multi-subject, and leave-one-subject-out (LOSO) protocols. Results are reported as mean ± standard deviation across the four evaluation participants. Bold values indicate the best result within each training protocol and metric.

- **Leave-one-subject-out (LOSO):** each participant is held out in turn, with the model trained on the remaining three participants; results are averaged across the four folds.

WALLBREAKER and MIRAGE use identical reconstruction backbones and matched training protocols in all settings. Their differences therefore reflect the hierarchy-aware components rather than changes to the pretrained representations or image generator.

Multi-subject training. WALLBREAKER outperforms MIRAGE on all four reported metrics. In particular, AlexNet(2) increases from 0.489 to 0.584, CLIP similarity from 0.501 to 0.517, and visual-cortex correlation from 0.137 to 0.157. The hierarchy-aware components therefore retain their advantage when training data are shared across participants.

Generalization to unseen participants. Under LOSO evaluation, WALLBREAKER improves SSIM from 0.355 to 0.367, AlexNet(2) from 0.457 to 0.462, and visual-cortex correlation from 0.065 to 0.073, while approximately matching MIRAGE on CLIP (0.449 versus 0.451). These results indicate a modest relative advantage when no training data from test participant are available.

Both methods nevertheless exhibit substantial degradation from the subject-specific to the LOSO setting; for example, the visual-cortex correlation of WALLBREAKER decreases from 0.169 to 0.073. Thus, the results support relative cross-subject robustness but do not establish subject-independent or population-level reconstruction. Explicit cross-subject alignment remains an important direction for future work.

H COMPUTATIONAL COST

WALLBREAKER retains the pretrained representation models and image-generation backbone used by MIRAGE. The hierarchy-driven perturbation is parameter-free and applied only during training, while the hierarchy-preference estimator introduces only 30 additional trainable parameters. As shown in Table 10, WALLBREAKER increases decoder training time from 24.768 to 24.770 seconds per epoch. The latency of mapping an fMRI sample to the reconstruction conditions increases from 0.378 to 0.395 milliseconds, corresponding to an additional 0.017 milliseconds per sample. The reported conditioning latency excludes image synthesis because both methods use the same frozen generator with identical sampling settings. The comparison therefore isolates the additional cost of hierarchy-preference estimation and adaptive conditioning. Overall, WALLBREAKER introduces negligible parameter and computational overhead without increasing the capacity of the generative backbone.

Metric	MIRAGE	WALLBREAKER
Decoder training (s/epoch)	24.768±0.009	24.770±0.009
fMRI conditioning (ms/sample)	0.378	0.395
Total pipeline parameters	7.134B	7.134B
Additional trainable parameters	–	30

Table 10: **Computational overhead relative to MIRAGE.**