What Drives Improvement in MindEye2 fMRI-to-Image Reconstruction?

Hannah Faghihi Kevin Bretz Amirreza Davachi Maftukhatul Q. Virati

Abstract

Reconstructing perceived images from functional magnetic resonance imaging (fMRI) activity is hard because the signal is noisy, high-dimensional and strongly subject-specific. Recent decoding pipelines produce strong perceptual reconstructions by combining large pretrained vision-language models (VLMs) with diffusion-based image generators as priors over natural images. This report studies four ways to improve MindEye2 reconstructions, each intervening at a different point in its pipeline. Semantic Auxiliary Loss (SemAux) adds a training-time semantic regularizer during subject-specific fine-tuning. Brain-Optimized Inference (BOI) acts only at inference, generating several candidate reconstructions and keeping the one whose predicted brain response best matches the measured fMRI. A diffusion timepoint sweep varies the noise level at which image-to-image refinement begins at the unCLIP and Stable Diffusion XL (SDXL) stages. A VLM caption ablation replaces MindEye2's Generative Image-to-Text (GIT) refinement caption with VLM captions under several prompting schemas. We evaluate all four on Subject 2 of the Natural Scenes Dataset (NSD), each against its own matched MindEye2 baseline. SemAux improves image retrieval but degrades most reconstruction metrics, notably PixCorr and brain retrieval. BOI raises brain correlation at the cost of perceptual similarity, with a v2 variant recovering most of this loss for the smallest brain-correlation gain. The timepoint sweep shows an unCLIP trade-off and consistent perceptual harm from extra SDXL iterations and a fair caption comparison, as every source given the ground-truth image, shows VLM schemas beat GIT on high-level metrics. Training-time semantic guidance and inference-time brain guidance both show concrete benefits, but each needs careful balancing.

1. Introduction

Reconstructing visual perception from functional magnetic resonance imaging (fMRI) sits at the intersection of deep learning, computer vision and neuroscience. The goal is to recover the image a person viewed from their recorded fMRI activity, building on early work that identified seen natural images from brain activity (Kay et al., 2008). The task is hard because fMRI is an indirect and noisy measurement, each image produces thousands of voxel responses and both the voxel space and the response patterns vary across subjects. Collecting fMRI data for each new subject is also expensive and time-consuming, so models should work from only a small amount of subject-specific data.

Recent fMRI-to-image reconstruction methods use large pretrained vision-language models (VLMs) and diffusion-based image generators as priors over natural images (Scotti et al., 2023; Gong et al., 2025; Scotti et al., 2024). Because voxel layouts and response patterns are subject-specific, many recent methods align multiple subjects into a shared latent space and then learn a shared decoder on top, fine-tuning only a small subject-specific component on limited data (Scotti et al., 2024; Gong et al., 2025; Beliy et al., 2026). MindEye2 (Scotti et al., 2024) follows this approach in the limited-data setting and we use it as our baseline.

We investigate four directions for improving MindEye2 reconstructions. First, the brain-derived semantic embedding both conditions image generation and influences the caption used during refinement. If this brain-derived representation loses high-level semantic information during fine-tuning, the final reconstruction may remain visually plausible but semantically inaccurate. Semantic Auxiliary Loss (SemAux) therefore adds an explicit semantic supervision signal during fine-tuning, to make the learned representation more semantically informative without altering the original MindEye2 objectives. Second, the diffusion decoder is stochastic and can produce many plausible images from the same brain signal, yet the brain signal itself is never used to select among these candidates at inference time. Brain-Optimized Inference (BOI), introduced by Kneeland et al. (Kneeland et al., 2023) for MindEye1, takes advantage of this randomness by generating multiple candidate images and selecting the one whose predicted fMRI response best matches the measured

signal.

The other two directions are inference-time analyses of choices that have not been systematically studied in prior MindEye2 work. The diffusion timepoint sweep varies the number of denoising steps used at the unCLIP and the SDXL refinement stages, exposing the trade-off between preserving low-level structure from the source image and letting the diffusion prior dominate. The VLM caption ablation replaces MindEye2's Generative Image-to-Text (GIT) refinement caption with VLM-generated captions under several prompting schemas. For a fair comparison, every caption source (including GIT) is given the ground-truth image. This is a best-case test, not a deployable pipeline. The design is inspired by recent SynMind-style work that uses VLM captions to guide reconstruction (Yang et al., 2026).

Our contributions are:

- **Semantic Auxiliary Loss (SemAux)**: a Contrastive Language-Image Pretraining (CLIP) text-space regularizer added to MindEye2's subject-specific fine-tuning loss.
- **L-BOI and L-BOI v2**: two new Brain-Optimized Inference variants that score candidates in MindEye2's shared latent space and reduce the perceptual penalty of brain-guided sampling.
- **Diffusion timepoint sweep**: the first systematic sweep of the diffusion start-timepoint at both the unCLIP and Stable Diffusion XL (SDXL) refinement stages of MindEye2.
- **VLM caption ablation**: a fair comparison of six VLM schemas, the Natural Scenes Dataset (NSD) ground-truth caption and GIT, all given the ground-truth image.

Each direction is compared against its own matched MindEye2 baseline on Subject 2, so the results should be read as per-experiment rather than a combined-system comparison. Integrating the selected directions into a single pipeline is left to future work.

2. Related Work

2.1. fMRI-to-image reconstruction

Early fMRI-to-image reconstruction methods attempted to decode visual stimuli from brain activity using handcrafted visual features and classical encoding models (Kay et al., 2008; Naselaris et al., 2009). More recent approaches use deep neural networks and diffusion models to produce higher-quality natural image reconstructions. These methods typically map fMRI signals into the latent spaces of pretrained visual or multimodal models, then decode those representations into images. MindEye1 introduced a pipeline based on contrastive learning and diffusion priors (Scotti et al., 2023), while MindEye2 extended this direction with cross-subject pretraining and a stronger SDXL unCLIP reconstruction pipeline (Scotti et al., 2024).

2.2. Cross-subject and limited-data visual decoding

A major challenge in fMRI decoding is that voxel spaces and response patterns are subject-specific. Single-subject models often require many hours of fMRI data, which limits their practical usefulness. Cross-subject methods try to solve this issue by transferring information across subjects, usually by aligning subject-specific brain responses into a shared representation space. MindEye2 follows this direction by using subject-specific linear mappings into a shared latent space, followed by a shared model trained across subjects.

Recent work explores how to adapt or structure subject-specific information more carefully. MindTuner (Gong et al., 2025) studies cross-subject visual decoding using subject-specific adaptation and semantic correction, while Brain-IT (Beliy et al., 2026) introduces a transformer-based approach for mapping brain activity to image features. These works show that limited-data fMRI reconstruction can benefit from better subject adaptation and stronger semantic representations.

2.3. Semantic guidance and auxiliary supervision

Many recent fMRI-to-image reconstruction methods use CLIP-like embedding spaces because they capture high-level visual semantics such as object categories and scene content, which can then guide image generation (Radford et al., 2021; Cherti et al., 2023; Scotti et al., 2024). However, semantic embeddings alone do not guarantee exact reconstruction. Two images can be close in a semantic embedding space while still differing in object identity, spatial layout, color, or details. This matters in fMRI reconstruction, where the predicted representation is learned from noisy and limited data. Several recent approaches add semantic information, text-based correction, or auxiliary objectives to improve the semantic reliability of reconstruction pipelines (Gong et al., 2025). We follow this line by adding a lightweight auxiliary semantic objective during MindEye2 fine-tuning.

2.4. Brain-guided inference and neural encoding models

A complementary line of work evaluates reconstructions through a forward encoding model that maps images back into predicted fMRI responses. Brain-Optimized Inference (BOI) uses this signal in a closed-loop manner. It generates multiple candidates with a stochastic decoder, scores each by how well its predicted response matches the measured

fMRI and keeps the best one (Kneeland et al., 2023). We adapt this idea to MindEye2 and study alternative scoring strategies.

2.5. Image-to-image initialization and diffusion timepoints

A partial-diffusion image-to-image trajectory, where the model starts from a known image at some intermediate noise level rather than from pure noise, was introduced by SDEdit (Meng et al., 2022) for image editing and has since been adopted in fMRI reconstruction. Brain-IT (Beliy et al., 2026) initializes the unCLIP diffusion stage from a coarse brain-derived image so that the trajectory preserves low-level structure rather than re-generating it from scratch. We adopt the same kind of initialization at both the unCLIP and SDXL refinement stages of MindEye2 and treat the diffusion timepoint as a hyperparameter to sweep.

2.6. Vision-language captions for reconstruction

Several recent fMRI-to-image systems use text captions to guide the image generator. SynMind (Yang et al., 2026) uses a vision-language model (Qwen2-VL, in its MimeVis module) to caption the viewed stimuli and train a mapping from fMRI to a semantic embedding, which then conditions a Stable Diffusion generator. Our VLM caption ablation borrows the idea of VLM-sourced semantic guidance but applies it as a controlled caption-source swap inside MindEye2, comparing VLM captions against the GIT baseline with all other components fixed.

3. Method

We study four modifications to the MindEye2 pipeline (Scotti et al., 2024), summarized in Figure 2. We first recap the MindEye2 components touched by these modifications. The full pipeline is described in (Scotti et al., 2024) and illustrated in Figure 1.

3.1. MindEye2 components used in this work

MindEye2 maps fMRI voxels through a subject-specific linear mapping into a shared latent space. A shared multi-layer perceptron (MLP) backbone produces a brain-derived representation that is decoded by a diffusion prior into a sequence of OpenCLIP ViT-bigG/14 image embeddings (Cherti et al., 2023; Ramesh et al., 2022), which condition an SDXL unCLIP decoder (Podell et al., 2024). The resulting unrefined reconstruction is then refined through SDXL image-to-image conditioned on a caption generated from the same brain-derived representation by a frozen GIT model (Wang et al., 2022) and weighted-averaged with a low-level blurry reconstruction predicted through Stable Diffusion's variational autoencoder (VAE) latent space (Rombach et al., 2022; Scotti et al., 2024). Training combines a diffusion-prior loss, a contrastive retrieval loss and a low-level reconstruction loss:

$$\mathcal{L}_{\text{ME2}} = \mathcal{L}_{\text{prior}} + \alpha_1 \mathcal{L}_{\text{BiMixCo|SoftCLIP}} + \alpha_2 \mathcal{L}_{\text{low-level}}. \quad (1)$$

3.2. Semantic Auxiliary Loss (SemAux)

SemAux adds a semantic regularizer to the fine-tuning objective. A small projector head g_{sem} pools the MindEye2 backbone output across its token positions and maps the resulting vector into the CLIP text-embedding space. The target is a frozen CLIP ViT-L/14 text embedding of a weak semantic label of the viewed image. Let $\hat{\mathbf{s}}$ denote the predicted semantic embedding and $\mathbf{s}$ the target text embedding. The SemAux loss is the cosine distance between them:

$$\mathcal{L}_{\text{SemAux}} = 1 - \frac{\hat{\mathbf{s}} \cdot \mathbf{s}}{\|\hat{\mathbf{s}}\| \, \|\mathbf{s}\|}. \quad (2)$$

The total fine-tuning loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ME2}} + \lambda_{\text{sem}} \mathcal{L}_{\text{SemAux}}, \quad (3)$$

where λ_{sem} is a small scalar that prevents the auxiliary term from dominating the original MindEye2 objectives. Implementation details (projector architecture, pooling) are in Appendix B. The value of λ_{sem} and other training settings are given in Section 4.2.

3.3. Brain-Optimized Inference (BOI)

MindEye2 reliably recovers a scene's semantic content, but fine details such as color, texture and spatial layout are guessed by the diffusion prior rather than inferred from the brain signal. If multiple plausible reconstructions can be generated from the same fMRI response, the signal itself can be used to select the most faithful one. Kneeland et al. (Kneeland et al., 2023) demonstrated this on MindEye1, improving brain correlation by up to 26% across visual cortex regions.

BOI is applied after the trained MindEye2 model has produced an unrefined reconstruction and does not modify the trained weights. Given an initial reconstruction x_0 and the measured fMRI response $\mathbf{v}$, BOI iteratively generates N candidates via SDXL image-to-image at a sequence of noise levels $[\tau_1, \ldots, \tau_T]$, scores each candidate c by how well a forward encoder GNet (St-Yves et al., 2023) predicts $\mathbf{v}$ from it and keeps the best-scoring candidate as the starting point for the next iteration (Algorithm 1).

We evaluate three scoring strategies, which differ only in the function $\text{score}(\cdot, \cdot)$ used in line 6 of Algorithm 1.

BOI Voxel. Candidates are scored using Pearson correlation between the predicted and measured fMRI responses in

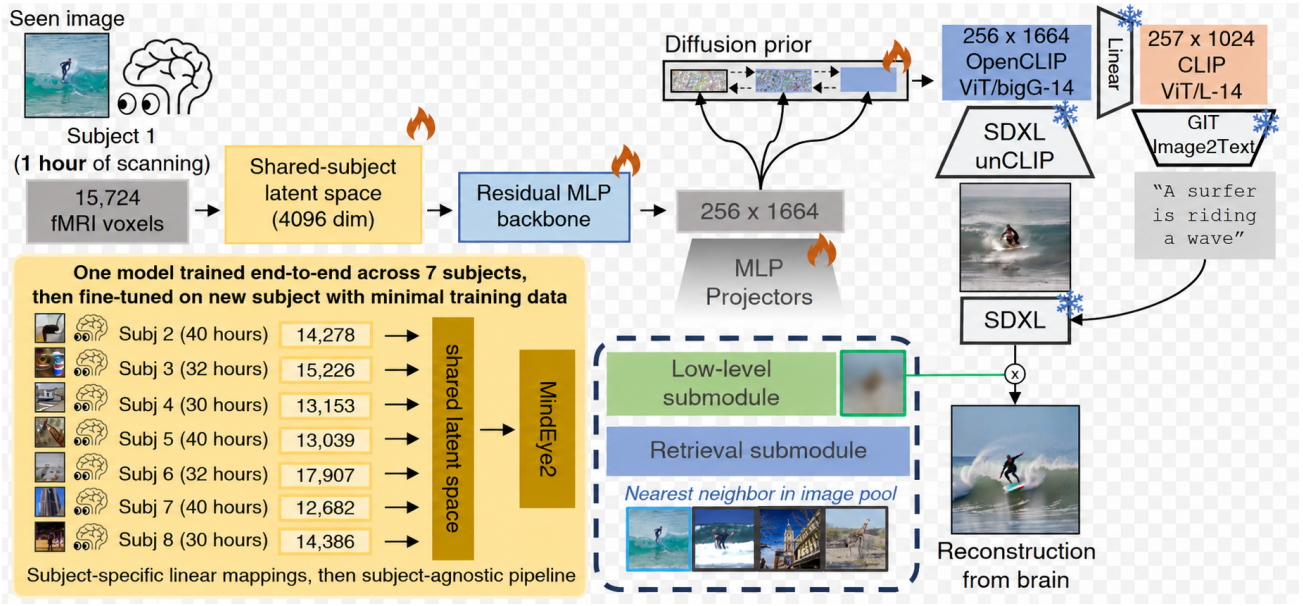


Figure 1. The MindEye2 reconstruction pipeline: subject-specific fMRI voxels are projected into a shared latent space, a backbone MLP predicts OpenCLIP image embeddings via a diffusion prior, an SDXL unCLIP decoder produces an unrefined image and a caption predicted from the brain-derived representation guides an SDXL image-to-image refinement that is combined with a low-level blurry reconstruction. Our four extensions are shown separately in Figure 2.

Algorithm 1 Brain-Optimized Inference (BOI)

Require: Initial reconstruction x_0, measured fMRI $\mathbf{v}$, candidates N, iterations T, noise schedule $[\tau_1, \ldots, \tau_T]$
Ensure: Best reconstruction x^*

```
x ← x_0
for t = 1 T do
    for i = 1 N do
        c_i ← SDXL-img2img(x, τ_t)
        v̂_i ← GNet(c_i)
        s_i ← score(v̂_i, v)
    end for
    Select best candidate: x ← c_{arg max_i s_i}
end forx
```

raw voxel space:

$$s(c) = \frac{(\hat{\mathbf{v}}^c - \bar{\hat{v}}^c) \cdot (\mathbf{v} - \bar{v})}{\|\hat{\mathbf{v}}^c - \bar{\hat{v}}^c\| \, \|\mathbf{v} - \bar{v}\|}, \tag{4}$$

and the best candidate is selected as $c^* = \arg\max_c s(c)$. This is the original formulation of Kneeland et al. (Kneeland et al., 2023).

L-BOI: latent-space scoring. Both the predicted and measured voxels are projected through MindEye2's trained ridge-regression weights $W \in \mathbb{R}^{1024 \times V}$ into a 1024-dimensional shared latent space and each candidate is scored using cosine similarity:

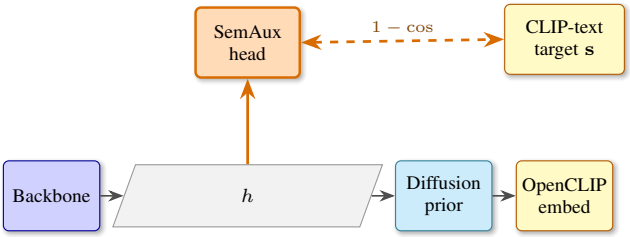
$$s(c) = \frac{\hat{\mathbf{z}}^c \cdot \mathbf{z}}{\|\hat{\mathbf{z}}^c\| \, \|\mathbf{z}\|}, \tag{5}$$

where $\hat{\mathbf{z}}^c = W\hat{\mathbf{v}}^c$ and $\mathbf{z} = W\mathbf{v}$.

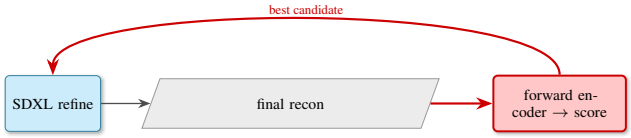
L-BOI v2. Uses the same scoring as L-BOI, but with three design changes which are initializes from MindEye2's refined reconstructions rather than unrefined ones, uses a gentler noise schedule and runs for fewer iterations. The exact noise schedules and iteration counts for all three variants are given in Section 4.3.

3.4. Diffusion timepoint sweep

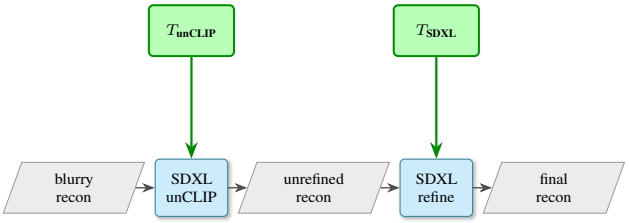
MindEye2's unCLIP stage starts from pure Gaussian noise, which discards the brain-derived low-level information the rest of the pipeline already provides through its blurry reconstruction. Following SDEdit (Meng et al., 2022) and Brain-IT (Beliy et al., 2026), we instead initialize unCLIP from this blurry reconstruction, encoded through SDXL's VAE. The diffusion then starts from an intermediate noise level T_{unCLIP} rather than from pure noise. The same parameter exists at the SDXL refinement stage as T_{SDXL}, where image-to-image already runs from the unCLIP output. We treat both as hyperparameters and sweep them independently. Lower values preserve more of the input image, while higher values give the diffusion prior more room to overwrite it.



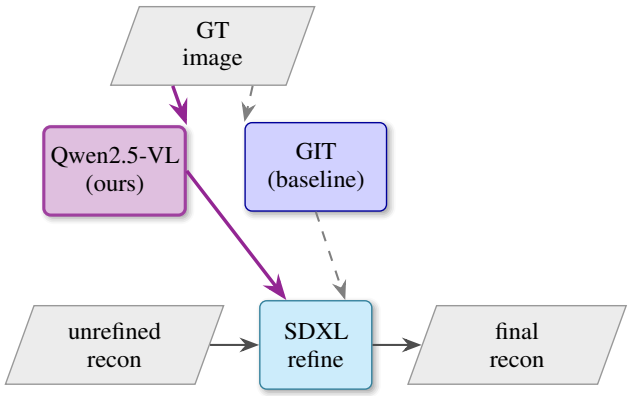
(a) SemAux adds a small head to h during fine-tuning and pulls it toward a frozen CLIP-text embedding of a weak semantic label of the viewed image.



(b) BOI keeps inference in a closed loop. A forward encoder scores each candidate reconstruction against the measured fMRI and the best one is fed back into SDXL refinement for the next round.



(c) We sweep the diffusion timepoint at the two image-to-image stages. T_{unCLIP} controls how many denoising steps start from the blurry reconstruction at the unCLIP stage and T_{SDXL} does the same at the SDXL refinement stage.



(d) We swap MindEye2's GIT caption for one from a vision-language model. For a fair comparison, both captioners are given the ground-truth (GT) image. GIT runs on its embedding (gray, baseline), Qwen2.5-VL (Bai et al., 2025) directly (purple, ours).

Figure 2. Method overview, one panel per extension. (a) SemAux modifies the fine-tuning objective at the backbone output. (b) BOI adds a closed-loop candidate selection step at inference. (c) The diffusion timepoint sweep varies a hyperparameter at the two image-to-image diffusion stages. (d) The VLM caption ablation swaps the caption source used during SDXL refinement.

3.5. VLM caption ablation

MindEye2 generates the refinement caption by passing the brain-derived OpenCLIP embedding through a frozen GIT model (Wang et al., 2022) and uses that caption as the text prompt during SDXL refinement. GIT was trained to caption natural images, not brain-derived embeddings, so the caption it produces depends on whatever survives the OpenCLIP-to-text translation and cannot be easily inspected or steered. We replace GIT with a vision-language model and vary the prompt schema to study which caption style helps the refinement stage most. For a fair comparison, both captioners are given the ground-truth image. The rest of the pipeline is unchanged.

4. Experimental Setup

4.1. Dataset

All experiments use the NSD, a large-scale 7T fMRI dataset in which subjects viewed natural images drawn from the Microsoft Common Objects in Context (MS COCO) dataset (Allen et al., 2022; Lin et al., 2014). NSD provides paired image and brain-activity samples, making it a standard benchmark for fMRI-to-image reconstruction.

Following the MindEye2 setup, the data is organized into per-subject scanning sessions, each containing 750 fMRI trials and corresponding to approximately one hour of data (Scotti et al., 2024). Training data is subject-specific, while evaluation is performed on a shared test set of images viewed by all subjects. Because the same test images are used across subjects, reconstruction quality can be compared against the same ground-truth images.

In the test set, each image is shown to a given subject up to three times. Following the standard NSD evaluation protocol used by MindEye2 and related work (Scotti et al., 2024; Gong et al., 2025; Allen et al., 2022), the repeated trial responses are averaged in raw voxel space to reduce trial-level noise before decoding. All four experiments use Subject 2 and a MindEye2 model fine-tuned on a single training session, about one hour of scanning data. We focus on the single-session setting because it would let someone reconstruct images from a single one-hour fMRI visit in-

stead of the 40 sessions that was standard practice before MindEye2. The trade-off is that single-session fine-tuning has noticeable run-to-run variance, so the baseline numbers differ slightly across the four experiments. Each experiment uses its own independently fine-tuned checkpoint, so within-experiment comparisons carry the significance. Across-experiment comparisons cannot be made directly due to the varying baselines.

4.2. SemAux setup

The semantic auxiliary weight was set to $\lambda_{\text{sem}} = 0.05$. Remaining settings are summarized in Table 1.

Table 1. Experimental setup for the SemAux comparison. Settings are shared by the baseline and SemAux model.

Setting	Value
Subject	Subject 2
Training data	1 session
Epochs	150
Global batch size	24
Learning rate	3×10^{-4}
λ_{sem}	0.05

4.3. BOI setup

BOI variants are inference-time modifications applied to a MindEye2 reconstruction. The per-variant candidate count, iteration count, noise schedule and starting reconstruction (unrefined unCLIP output vs. refined output) are listed in Table 2.

Table 2. Hyperparameters for each BOI variant. N = number of candidates per iteration, T = number of iterations.

Model	N	T	Noise schedule	Initialization
Refined baseline	—	—	—	—
BOI Voxel	8	5	$[20, 15, 12, 9, 6]$	Unrefined recons
L-BOI	8	5	$[20, 15, 12, 9, 6]$	Unrefined recons
L-BOI v2	8	3	$[12, 9, 6]$	Refined recons

4.4. Diffusion timepoint setup

MindEye2's default inference schedule uses 38 unCLIP denoising steps and 25 SDXL refinement steps. We sweep $T_{\text{unCLIP}} \in \{10, 15, 20, 25, 30, 35\}$ with T_{SDXL} at its published default of 13, then fix $T_{\text{unCLIP}} = 15$ and sweep $T_{\text{SDXL}} \in \{8, 13, 18, 23\}$. The baseline initializes unCLIP from pure noise.

4.5. VLM caption ablation setup

On a 50-image NSD test subset, every captioner is given the ground-truth image, making this a best-case test of the caption source. Schema 0 reuses the NSD ground-truth caption. Schemas 1–6 use Qwen2.5-VL-72B-Instruct-AWQ (Bai et al., 2025) under different prompting styles (see Appendix C). The GIT row uses MindEye2's frozen GIT model on the ground-truth image embedding. The SDXL refinement is rerun on the same unrefined reconstructions for each caption at $T_{\text{SDXL}} = 15$, only changing the text prompt.

4.6. Evaluation metrics

We evaluate reconstructions using low-level metrics (pixel correlation, SSIM, AlexNet at layers 2 and 5), perceptual metrics drawn from pretrained image encoders (Incep, CLIP, Eff, SwAV) and retrieval metrics (image-to-brain and brain-to-image). For BOI we also report brain correlation, the Pearson correlation between the reconstruction's GNet-predicted fMRI response and the measured response in the `nsd_general` region of interest (ROI). Because these metrics capture different aspects of quality, we do not treat any single one as definitive. Full metric definitions are in Appendix A.

5. Results

5.1. SemAux results

Table 3 compares SemAux with the MindEye2 baseline. The results are mixed. SemAux improves image retrieval from 0.584 to 0.616 and ties the baseline on Incep (0.796 vs. 0.795), but underperforms it on all other metrics.

Figure 5 in the appendix shows six example reconstructions comparing the ground-truth image, the baseline reconstruction, the SemAux reconstruction and the predicted captions.

Qualitative inspection suggests that SemAux does not affect all images in the same way. For example, on the bus (row 2) and clock tower (row 6) the SemAux caption is shorter and less descriptive than the baseline yet better matches the actual scene, leading to a more faithful refinement. In other cases the change is small. A possible failure mode is that when the predicted caption becomes too generic, the refinement may produce a visually plausible image that loses specific details of the original scene.

5.2. BOI results

Table 4 reports the BOI results for Subject 2. The main pattern is that all BOI variants improve brain correlation relative to the refined baseline, but this comes at the cost of perceptual image metrics. BOI Voxel and L-BOI show the largest degradation in low-level and high-level metrics, while L-BOI v2 partially reduces this penalty by starting from refined reconstructions and using a gentler noise schedule.

The results suggest that optimizing for neural fidelity and optimizing for perceptual similarity are not always aligned. A reconstruction can better match the measured brain re-

Table 3. SemAux results on Subject 2 using one training session. Arrows indicate whether higher or lower values are better.

Model	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	ImgRet ↑	BrainRet ↑
Baseline	**0.168**	**0.290**	**0.801**	**0.895**	0.795	**0.770**	**0.842**	**0.483**	0.584	**0.382**
SemAux	0.152	0.286	0.798	0.890	**0.796**	0.769	0.845	0.485	**0.616**	0.361

Table 4. BOI results for Subject 2. Brain Corr. refers to the nsd_general region. Bold indicates the best value per column.

Model	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Brain Corr. ↑
Refined baseline	**0.194**	**0.426**	0.845	**0.909**	**0.819**	**0.812**	**0.805**	**0.463**	0.352
BOI Voxel	0.163	0.409	0.851	0.894	0.794	0.792	0.824	0.500	**0.380**
L-BOI	0.163	0.409	0.848	0.896	0.800	0.796	0.823	0.499	0.371
L-BOI v2	0.169	0.417	**0.863**	0.908	0.813	0.799	0.814	0.484	0.367

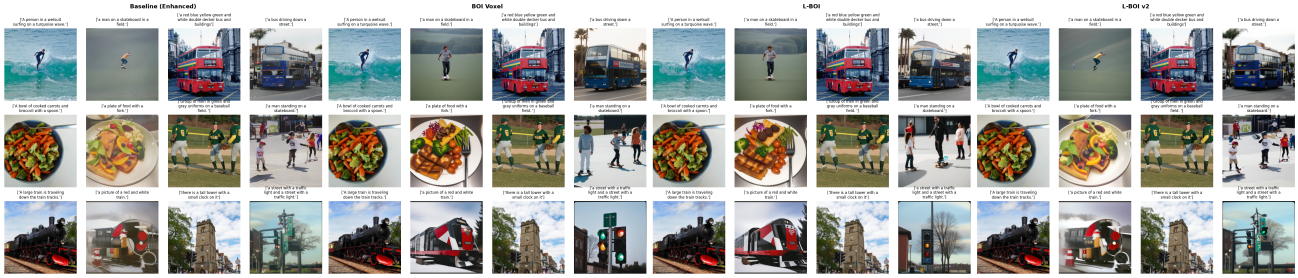


Figure 3. Visual comparison of reconstructions across all four conditions for Subject 2. From left to right: Refined baseline, BOI Voxel, L-BOI and L-BOI v2.

sponse according to an encoding model while becoming less similar to the ground-truth image according to standard perceptual metrics.

Figure 3 compares reconstructions across the four conditions. Semantic content stays consistent in every case, BOI does not introduce or correct semantic errors but operates within the range of variation that SDXL's stochastic sampling already produces from the same conditioning signal. The differences are concentrated in color fidelity and texture. The baseline and L-BOI v2 stay close to the original palette. BOI Voxel and L-BOI produce flatter, less saturated images with visible color drift on the food-plate and train columns, consistent with cumulative drift from five iterations of an aggressive noise schedule. Images with high-confidence semantic anchors such as the skateboarding and bus columns remain stable across all variants. This is a ceiling effect where MindEye2 already reconstructs the scene reliably and BOI has little room to improve.

5.3. Diffusion timepoint sweep results

Figure 4 plots two low-level metrics (PixCorr, SSIM) and two high-level metrics (Incep, CLIP) across both sweeps against the noise-init baseline (dashed). The full 8-metric table is in Appendix E. Both sweeps trade low-level fidelity against high-level perceptual quality, lowering T_{unCLIP} to 10 raises PixCorr from 0.188 to 0.250 but drops CLIP from 0.808 to 0.775, while a low T_{SDXL} of 8 gives the best Incep score of the sweep (0.819).

The same trade-off shows up qualitatively in the appendix grids (Figures 6 and 7). T_{unCLIP} sets both the noise level and the number of denoising steps, so T_{unCLIP}=10 noises the blurry init to step 10 (of 38) and denoises 10 steps, keeping the output close to the dim, smooth input. Higher T_{unCLIP} lets the SDXL prior produce sharper, more color-saturated images that drift further from the original layout. The T_{SDXL} sweep shows that aesthetic quality and reconstruction accuracy come apart that higher T_{SDXL} gives visually crisper, more photorealistic and color-saturated images, but at $T_{\text{SDXL}} = 23$ the high-level metrics fall below the noise-init baseline (CLIP 0.773 vs. 0.808, Incep 0.766 vs. 0.802). The extra refinement iterations push outputs toward generic natural images, away from faithful reproductions.

5.4. VLM caption ablation results

Even when GIT is given the ground-truth image, all six VLM schemas and the brief NSD ground-truth caption beat it on the high-level metrics (Table 5). CLIP rises from GIT's 0.880 to 0.988 (positional), Incep from 0.900 to 0.962 (tags) and Eff drops from 0.711 to 0.602 (style). Low-level metrics are essentially tied. GIT-style captions are therefore a weaker refinement prompt than VLM-style descriptions, even before brain decoding is involved. Within the VLM schemas, dense leads on Alex(2) (0.902) and sentence + tags on SwAV (0.365), so the per-metric winners spread across schemas with no single style dominating.

Visually (Figure 8), the weak captions (GIT, short and some-

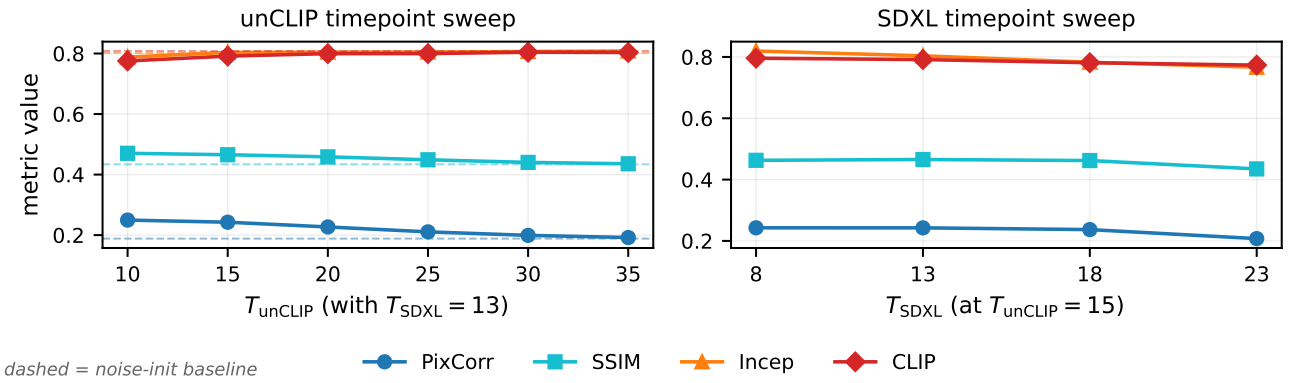


Figure 4. Diffusion timepoint sweep on Subject 2 (full test set). Left: T_{unCLIP} varied with T_{SDXL}=13 (default). Right: T_{SDXL} varied at T_{unCLIP}=15.

times NSD ground truth) miss obvious scene content like the windsurfer in column 4 and the pile of carrots in column 1. The richer schemas all capture scene content but with smaller, varying semantic inconsistencies.

6. Discussion

6.1. SemAux discussion and limitations

The SemAux results suggest that semantic supervision is not automatically beneficial for MindEye2. The improvement in image retrieval indicates that the auxiliary loss can affect the learned representation, but this did not translate into better reconstruction quality across the metrics.

A likely reason is that the semantic target is coarse. SemAux predicts a CLIP text embedding from weak semantic labels, so the supervision mainly encourages broad category or scene information. This can help high-level identification, but it may also reduce attention to details such as layout, texture, color and object relations. This is consistent with the decrease in low-level metrics such as PixCorr and SSIM.

The qualitative examples also show a possible failure mode of caption-guided refinement. If the predicted caption becomes too generic, SDXL may generate a visually plausible but oversimplified image. In this case, semantic guidance can move the reconstruction toward a confident but incorrect interpretation.

The current experiment has several limitations. The semantic labels are weak and coarse, λ_{sem} was not selected through a full hyperparameter search and the experiment was limited to one subject and one training-session setting. More precise labels, such as averaged multi-caption embeddings or multi-level semantic labels, may be a better semantic signal. Broader evaluation across subjects and loss weights is needed before SemAux can be considered a reliable improvement.

6.2. BOI discussion and limitations

All three BOI variants improve brain correlation (+0.015 to +0.028 on `nsd_general`) at the cost of perceptual quality. BOI Voxel and L-BOI degrade nearly all perceptual metrics relative to the baseline, while L-BOI v2 partially recovers this loss through better initialization and a gentler noise schedule. L-BOI v2 nearly matches the baseline on Alex(5) (0.908 vs. 0.909) and Incep (0.813 vs. 0.819) while still improving brain correlation, which suggests that its design choices reduce the perceptual penalty of brain-guided optimization. Comparing scoring strategies, L-BOI beats BOI Voxel on high-level perceptual metrics (Incep 0.800 vs. 0.794 and CLIP 0.796 vs. 0.792) but falls slightly short on brain correlation (0.371 vs. 0.380).

The design choices of L-BOI v2 were motivated by three problems visible in the Voxel and L-BOI runs. Both earlier variants started from unrefined reconstructions, which are weaker than the refined ones. So BOI was being compared against a stronger baseline than the image it started from, which is an unfair disadvantage from the first step. Starting from refined reconstructions fixes this. The aggressive noise schedule was another problem of five rounds of heavy noise caused the images to drift further from the original, hurting low-level metrics even when individual steps were improving brain alignment. Cutting to three gentler steps limits that drift. Finally, each test image was seen only three times by the subject, so the brain signal is noisy and with fewer iterations there are fewer chances for a bad score to push the reconstruction in the wrong direction.

Despite these incremental gains, several structural limitations prevented a more decisive improvement. First, MindEye2 already reconstructs the correct semantic category reliably, so BOI is left selecting among candidates that all represent the same scene, which gives the brain score little discriminative power at the category level. Second, GNet was trained on natural photographs, making its voxel pre-

Table 5. VLM caption ablation on Subject 2 (50-image NSD test subset). Every caption source is given the ground-truth image, so only the caption varies across rows. GIT uses MindEye2's caption model on the ground-truth image embedding. Schemas 1–6 use Qwen2.5-VL on the ground-truth image. All rows refine through SDXL at $T_{\text{SDXL}} = 15$. Bold marks the best value per column.

Schema	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
GIT (on GT image)	0.225	0.419	0.864	0.944	0.900	0.880	0.711	0.411
0. NSD ground truth	**0.231**	**0.420**	0.876	0.968	0.914	0.969	0.642	0.388
1. Short	0.227	0.419	0.872	0.971	0.944	0.983	0.634	0.384
2. Dense	0.229	0.416	**0.902**	0.967	0.940	0.985	0.617	0.374
3. Tags	0.227	0.414	0.891	0.972	**0.962**	0.974	0.624	0.376
4. Sentence + tags	0.228	0.413	0.899	0.966	0.949	0.982	0.605	**0.365**
5. Style	0.229	0.411	0.892	**0.973**	0.942	0.976	**0.602**	0.372
6. Positional	0.229	**0.420**	0.891	0.966	0.953	**0.988**	0.622	0.368

dictions less accurate for SDXL-generated images and degrading the quality of the scoring signal. These limitations are not unique to our setting. The results in the original BOI paper (Kneeland et al., 2023) indicate a similar PixCorr decrease of about 16.5% (0.310 to 0.259 in their Table 1) when BOI is applied to MindEye1, which suggests the perceptual cost is inherent to the method. The key difference is that MindEye2's stronger baseline leaves less room for improvement, making this cost more visible.

6.3. Diffusion timepoint sweep discussion

The unCLIP sweep shows a clear trade-off between low-level fidelity and high-level semantic content. The pixel-level gains at $T_{\text{unCLIP}} = 10$ (PixCorr 0.188 to 0.250, SSIM 0.434 to 0.470) come with high-level drops (CLIP 0.808 to 0.775, Incep 0.802 to 0.788). As T_{unCLIP} grows the picture inverts, by $T_{\text{unCLIP}} = 30$ the low-level metrics give back most of the gain and the high-level metrics return to or past the baseline. No single T_{unCLIP} wins across all metrics, so the right value depends on whether pixel-level fidelity or perceptual similarity matters more.

The SDXL stage is more asymmetric. Increasing T_{SDXL} at $T_{\text{unCLIP}} = 15$ hurts almost every metric (Alex(2) drops from 0.804 at $T_{\text{SDXL}} = 18$ to 0.762 at $T_{\text{SDXL}} = 23$). Decreasing it to $T_{\text{SDXL}} = 8$ is the best setting on Incep and stays close to the unCLIP-only setting on the rest. The unCLIP output is already close to the final image, so each extra refinement step is more likely to drift than to refine.

6.4. VLM caption ablation discussion

The setup only changes the refinement caption. The unrefined reconstruction is still MindEye2's brain-derived output, while every captioner is given the ground-truth image. This tells us which caption style works best as an SDXL refinement prompt, not what would happen end-to-end. Every alternative caption, even the brief NSD ground truth, lifts the high-level metrics far above GIT while leaving low-level metrics tied. Since GIT here gets a clean input, the gap is not just brain-decoding noise. The likely reason is GIT's brief, generic image-caption style, while VLM schemas give SDXL richer descriptions with more concrete content to refine toward.

Positional captions lead on CLIP, likely because explicit left/right and foreground/background cues fix SDXL's spatial layout without locking in the content.

7. Conclusion and Next Steps

We presented four directions for improving MindEye2 reconstructions on Subject 2 of NSD, which are SemAux as a training-time modification, BOI as an inference-time modification, a diffusion timepoint sweep at the unCLIP and SDXL refinement stages and a VLM caption ablation that replaces MindEye2's GIT-generated refinement caption. SemAux improves image retrieval but slightly reduces most reconstruction metrics. BOI variants improve brain correlation at the cost of perceptual similarity, with L-BOI v2 giving the best compromise. The timepoint sweep reveals that T_{unCLIP} trades pixel-level fidelity against high-level semantic quality and that extra SDXL refinement iterations promote aesthetic appearance over reconstruction accuracy as more iterations produce visually sharper images but push the high-level metrics below the noise-init baseline. The VLM caption ablation finds that the VLM schemas and the brief NSD ground-truth caption outperform GIT on high-level metrics, even when every captioner is given the ground-truth image.

Each direction is evaluated separately against its own matched MindEye2 baseline and on a single subject. Two next steps matter equally and both were out of reach within the time and compute budget for this report. First, confirming the trends on the other NSD subjects (1, 5, 7) covered by MindEye2, since single-subject results cannot establish generality. Second, combining the strongest variants into a single pipeline and evaluating the joint system, since the four directions intervene at different points in MindEye2 and should be complementary in principle. Other open directions include more informative semantic targets for SemAux (image-specific or averaged multi-caption embeddings), per-image choice of T_{unCLIP} and T_{SDXL} driven by a GNet brain-correlation signal and investigating why GIT is so weak as a refinement prompt.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., and Kay, K. N. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1): 116–126, 2022.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Beliy, R., Zalcher, A., Kogman, J., Wasserman, N., and Irani, M. Brain-it: Image reconstruction from fmri via brain-interaction transformer. In *International Conference on Learning Representations*, 2026.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Gong, Z., Zhang, Q., Bao, G., Zhu, L., Xu, R., Liu, K., Hu, L., and Miao, D. 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.

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

Kneeland, R., Ojeda, J., St-Yves, G., and Naselaris, T. Brain-optimized inference improves reconstructions of fmri brain activity. *arXiv preprint arXiv:2312.07705*, 2023.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., and Gallant, J. L. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Scotti, P. S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Cohen, E., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K. A., and Abraham, T. M. 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.

Scotti, P. S., Tripathy, M., Villanueva, C. K. T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K. A., and Abraham, T. M. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

St-Yves, G., Allen, E. J., Wu, Y., Kay, K., and Naselaris, T. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature Communications*, 14(1):3329, 2023.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, 2019.

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. Git: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Yang, L., Yang, M., Li, K., Zhang, H., Pang, K., and Song, Y.-Z. Synmind: Reducing semantic hallucination in fMRI-based image reconstruction. *arXiv preprint arXiv:2601.17857*, 2026.

A. Evaluation metrics

Reconstructions are evaluated using a combination of low-level, perceptual and retrieval metrics, following the protocol established in MindEye2 (Scotti et al., 2024).

Low-level metrics. Pixel correlation (PixCorr) is the Pearson correlation between flattened ground-truth and reconstructed pixel intensities. SSIM is the structural similarity index (Wang et al., 2004) averaged over local windows. Both capture intensity- and structure-level fidelity at the pixel scale. AlexNet(2) and AlexNet(5) measure 2-way percent-correct using AlexNet (Krizhevsky et al., 2012) activations at early and mid layers of the ground-truth and reconstructed images, capturing edge- and texture-level visual features.

Perceptual (high-level) metrics. InceptionV3 (Incep) (Szegedy et al., 2016) and CLIP (Radford et al., 2021) report 2-way percent-correct in their respective feature spaces. EfficientNet (Eff) (Tan & Le, 2019) and SwAV (Caron et al., 2020) report feature distances (lower is better) between ground-truth and reconstructed images, as additional perceptual signals.

Retrieval metrics. ImgRet (image retrieval) is the top-1 accuracy of selecting the correct image from a pool of candidates given the brain-derived embedding. BrainRet is the reverse. Both probe whether the learned representation is discriminative enough to identify the correct sample at the pool level.

Brain correlation (BOI only). For each reconstruction, we run the GNet forward encoder to obtain a predicted fMRI response and report the Pearson correlation between this predicted response and the measured response in the `nsd_general` ROI. This metric is used to assess how closely a reconstruction matches the brain activity that originally evoked the viewed image.

B. SemAux architecture details

SemAux is attached to the MindEye2 backbone before the diffusion prior. At this stage the backbone output is a sequence of 256 token vectors, each of dimension 1664, matching the token structure of the OpenCLIP ViT-bigG/14 image embedding. The SemAux head first averages these tokens into a single 1664-dimensional vector that summarizes the high-level information in the brain-derived representation. This pooled vector is then passed through a small projector head $g_{\text{sem}} : \mathbb{R}^{1664} \to \mathbb{R}^{768}$ whose output is a predicted semantic embedding (not a class prediction) in the CLIP ViT-L/14 text-embedding space. During training, the target $\mathbf{s}$ is built from a weak semantic text label of the viewed NSD image, encoded with a frozen CLIP ViT-L/14 text encoder and ℓ_2-normalized.

C. VLM caption schemas

Each VLM schema instructs Qwen2.5-VL (Bai et al., 2025) to describe the ground-truth image in a different style. The schemas, indexed 0–6 in our results table, are:

- **0 (NSD ground truth):** the NSD caption associated with the viewed image, included as a brief human-written reference (not an idealized upper bound).
- **1 (short):** a one-sentence description focused on the main subject and action.
- **2 (dense):** a multi-sentence description covering subject, action, setting, lighting, composition and style.
- **3 (tags):** a comma-separated list of tags covering subject, action, setting, lighting, style.
- **4 (sentence+tags):** a short sentence followed by a tag list.
- **5 (style):** a one-sentence description emphasizing visual style and rendering.
- **6 (positional):** a short sentence followed by explicit positional or compositional cues (left/right, foreground/background).

For each schema we keep the rest of the MindEye2 pipeline fixed and only vary the text prompt passed to the refinement stage.

D. Example reconstructions: SemAux

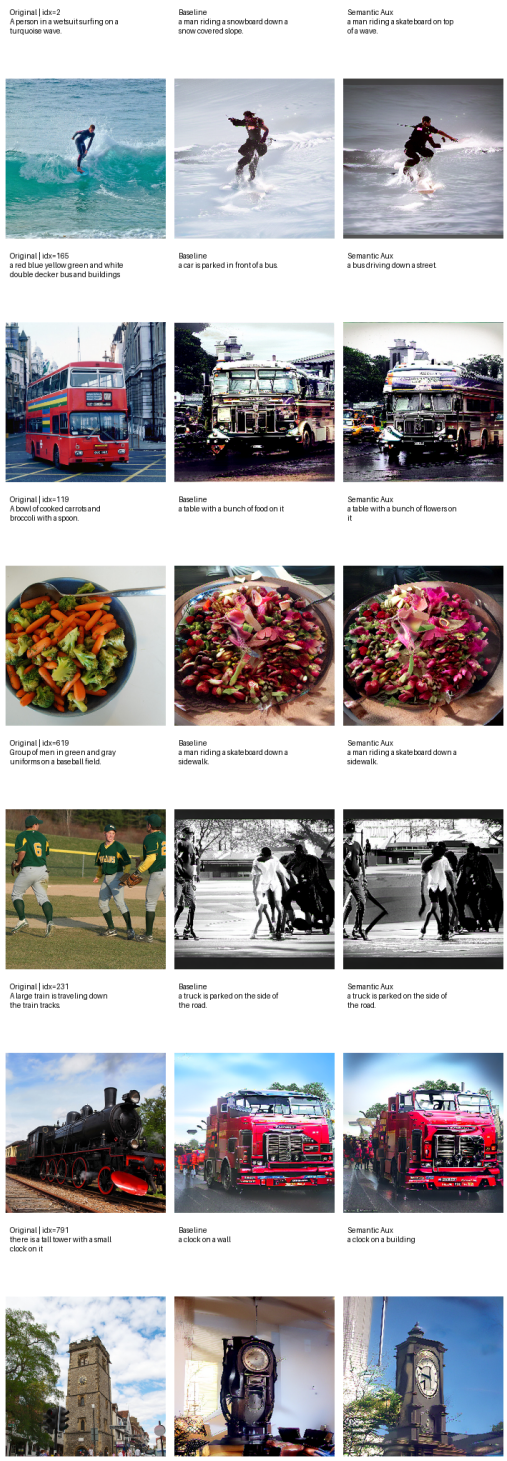


Figure 5. Qualitative comparison of ground-truth images, MindEye2 baseline reconstructions, SemAux reconstructions and predicted captions.

E. Diffusion timepoint sweep: full numerical results

Table 6. Diffusion timepoint sweep on Subject 2 (full test set). The baseline initializes unCLIP from pure noise. The first block sweeps T_{unCLIP} with $T_{\text{SDXL}} = 13$ (the MindEye2 published default). The second block sweeps T_{SDXL} at $T_{\text{unCLIP}} = 15$. Bold marks the best value per column.

Variant	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
Baseline (noise init)	0.188	0.434	0.842	**0.904**	0.802	**0.808**	**0.801**	0.469
$T_{\text{unCLIP}} = 10$ (at T_{SDXL}=13)	**0.250**	**0.470**	0.797	0.867	0.788	0.775	0.833	0.502
$T_{\text{unCLIP}} = 15$ (at T_{SDXL}=13)	0.243	0.466	0.826	0.886	0.804	0.791	0.818	0.483
$T_{\text{unCLIP}} = 20$ (at T_{SDXL}=13)	0.227	0.459	0.841	0.896	0.805	0.800	0.811	0.474
$T_{\text{unCLIP}} = 25$ (at T_{SDXL}=13)	0.211	0.449	**0.846**	0.901	0.805	0.800	0.805	0.471
$T_{\text{unCLIP}} = 30$ (at T_{SDXL}=13)	0.199	0.440	0.840	**0.904**	0.807	0.804	0.803	**0.468**
$T_{\text{unCLIP}} = 35$ (at T_{SDXL}=13)	0.192	0.436	0.844	**0.904**	0.808	0.803	0.803	**0.468**
$T_{\text{SDXL}} = 8$ (at T_{unCLIP}=15)	0.243	0.463	0.834	0.899	**0.819**	0.796	0.809	0.474
$T_{\text{SDXL}} = 18$ (at T_{unCLIP}=15)	0.237	0.462	0.804	0.860	0.783	0.781	0.832	0.502
$T_{\text{SDXL}} = 13$ (at T_{unCLIP}=15)	0.243	0.466	0.826	0.886	0.804	0.791	0.818	0.483
$T_{\text{SDXL}} = 23$ (at T_{unCLIP}=15)	0.208	0.435	0.762	0.822	0.766	0.773	0.841	0.522

F. Example reconstructions: T_{unCLIP} sweep

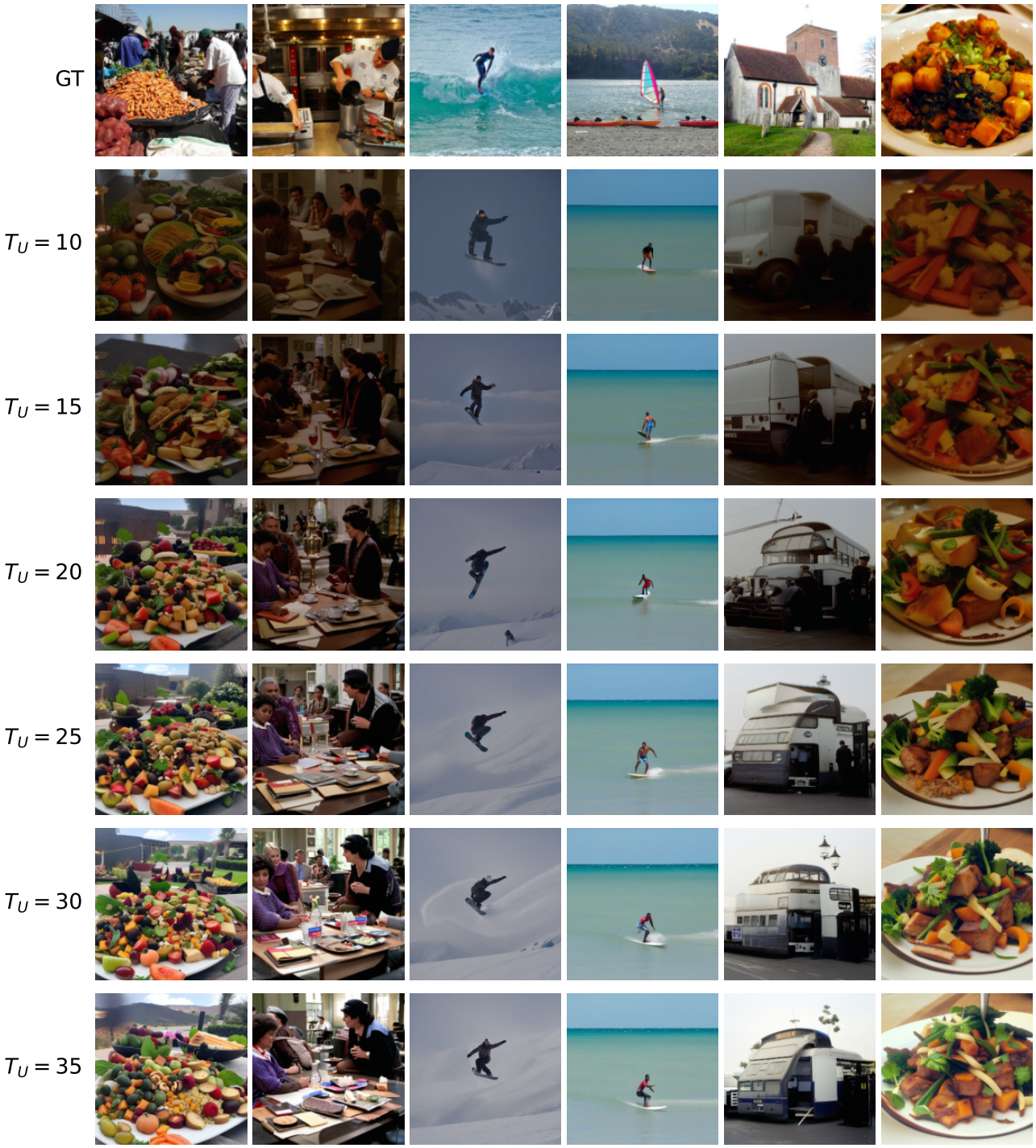


Figure 6. Same six Subject 2 test images reconstructed at each T_{unCLIP} in the sweep, all with $T_{\text{SDXL}} = 13$ (the MindEye2 published default). Qualitative interpretation is in Section 5.3.

G. Example reconstructions: T_{SDXL} sweep

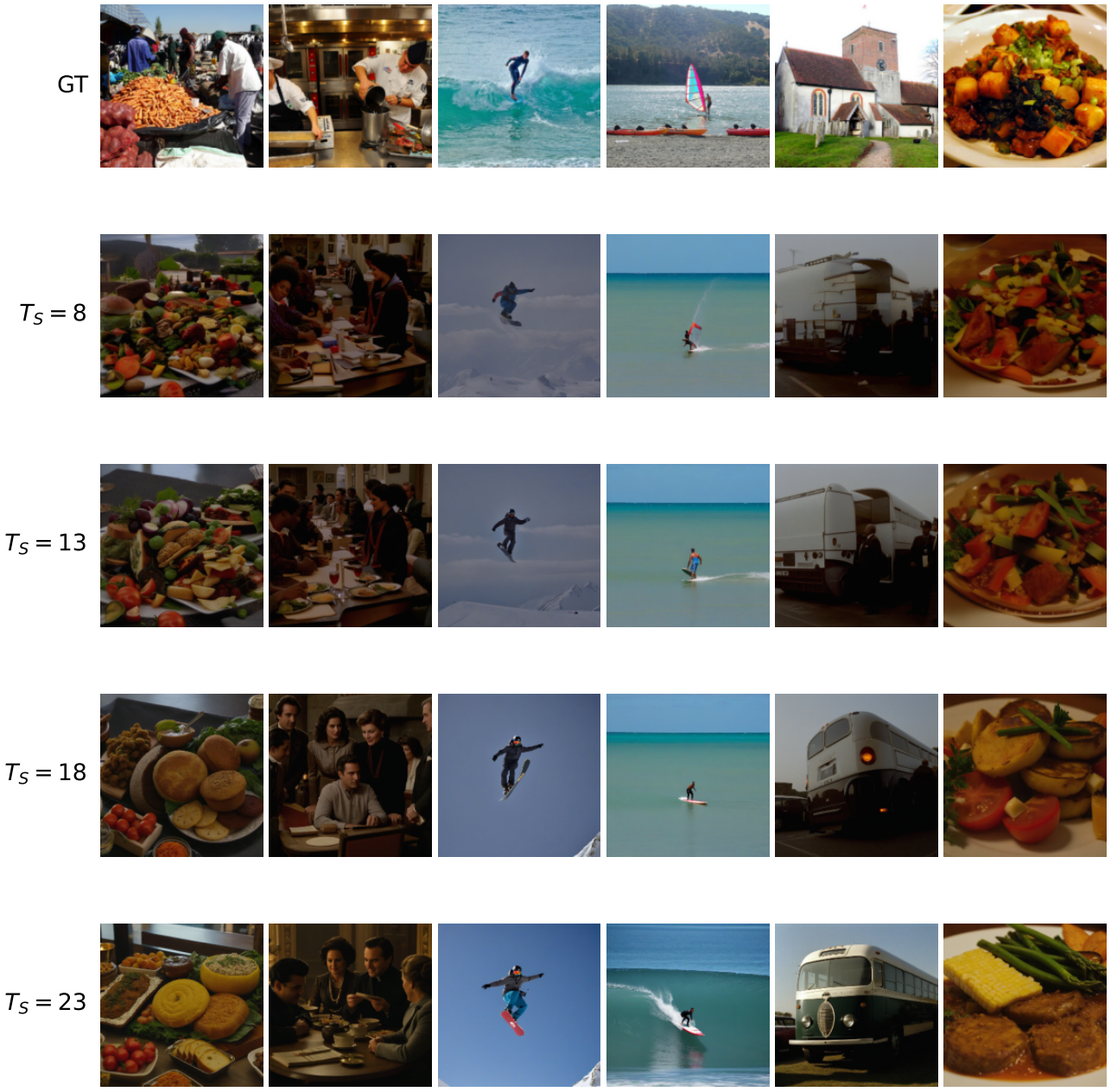


Figure 7. Same six Subject 2 test images reconstructed at each T_{SDXL} in the sweep, all with $T_{\text{unCLIP}} = 15$. Qualitative interpretation is in Section 5.3.

H. Example reconstructions: VLM caption ablation

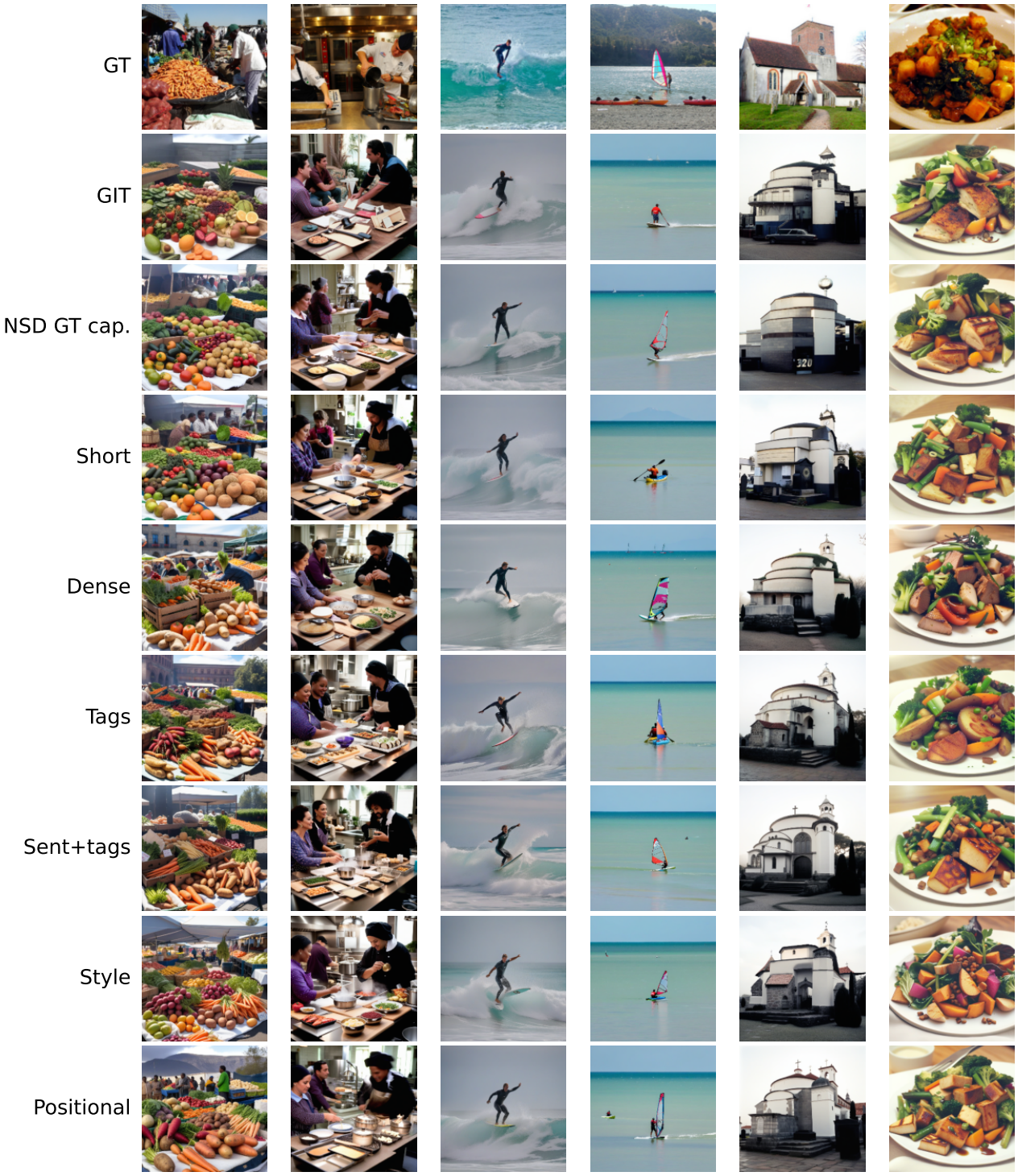


Figure 8. Same six Subject 2 test images refined with each caption source from Table 5: GIT on the ground-truth image embedding, the NSD ground-truth caption and the six VLM caption schemas (also from the ground-truth image). Qualitative interpretation is in Section 5.4.