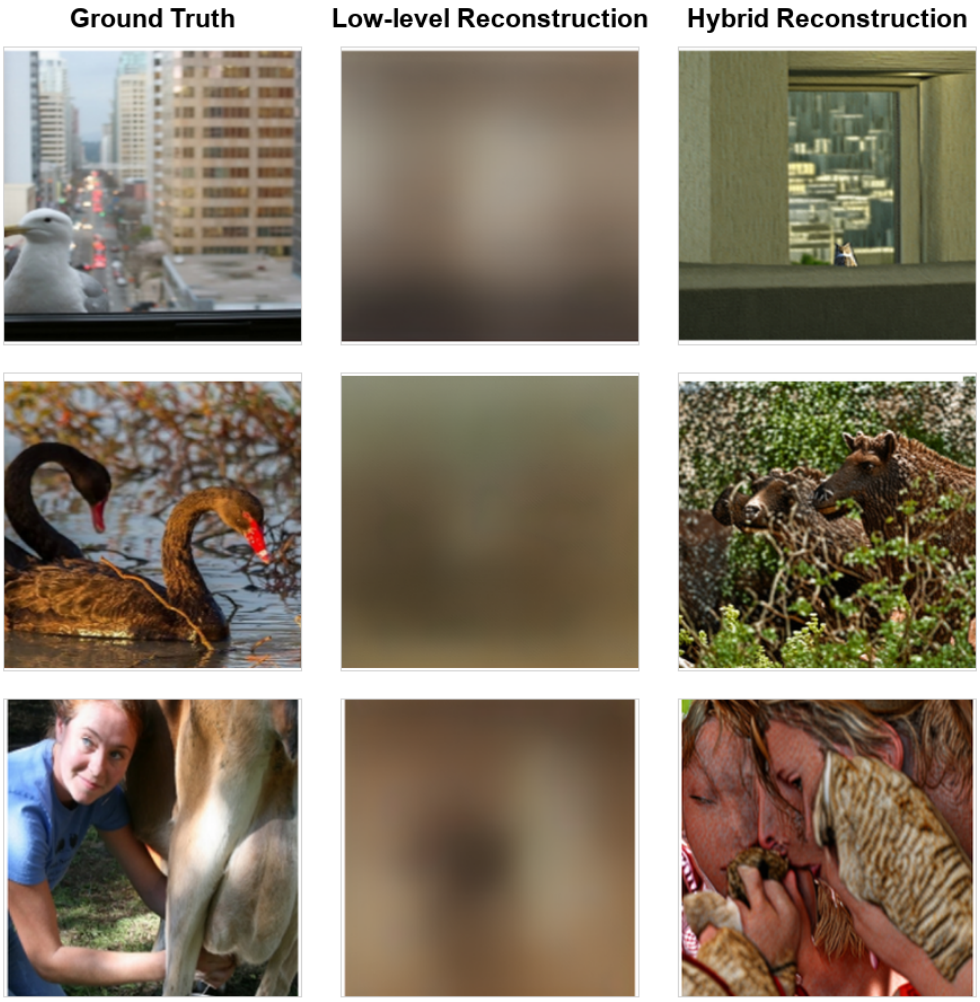
Supplementary Material

Naturalistic Stimulus Reconstruction from fMRI: A Primer in the Natural Scenes Dataset

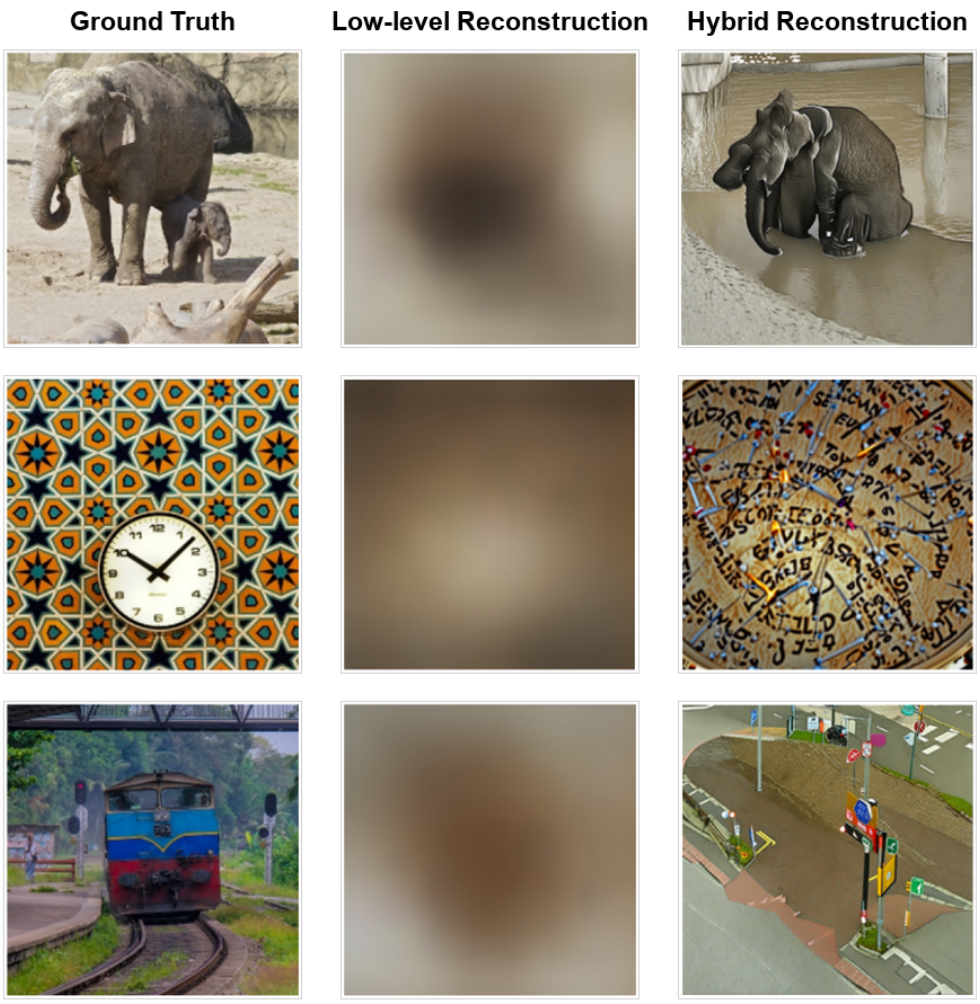
Umur Yıldız and Burcu A. Urgen

A

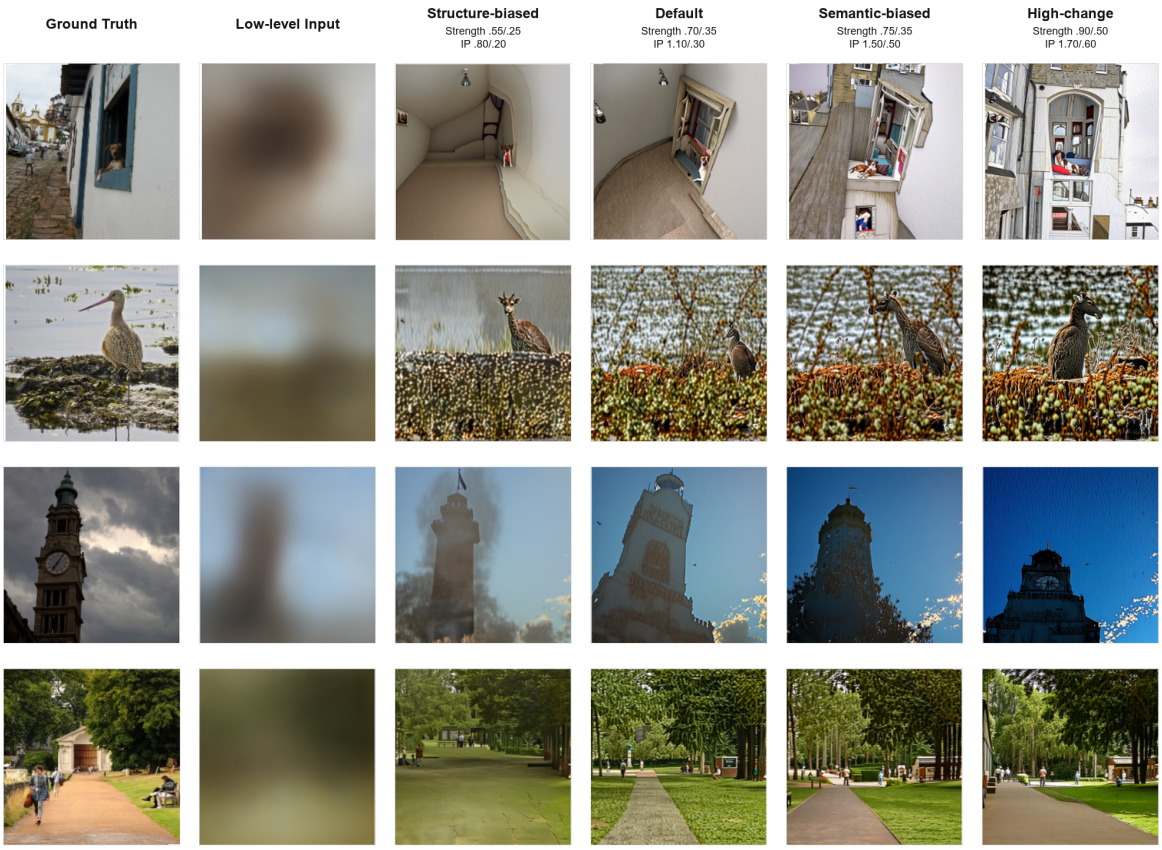


Supplementary Figure S1. Representative reconstruction failure modes. Each row shows the ground-truth image, the low-level reconstruction, and the final hybrid reconstruction. Examples illustrate qualitative failure modes observed in the generative reconstruction stage, including gist-preserving substitutions where broad scene structure is retained but the specific content changes, unrealistic object or body artifacts, and rare hard mismatches where the reconstruction differs substantially from the target. These examples highlight that visually plausible reconstructions can reflect both decoded brain-derived constraints and details supplied by the pretrained generative prior.

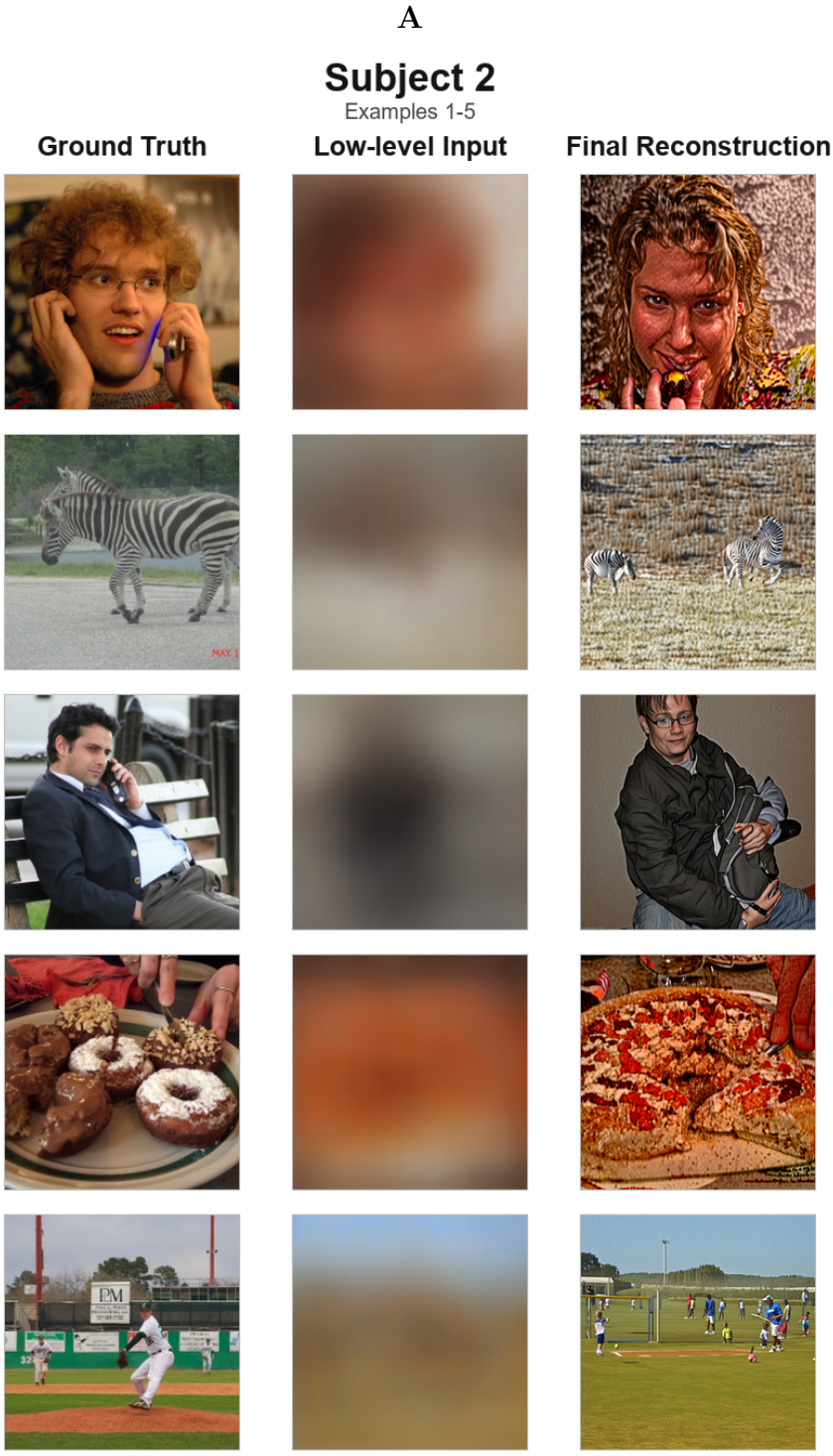
B



Supplementary Figure S1 continued. Additional examples of the same qualitative failure modes.

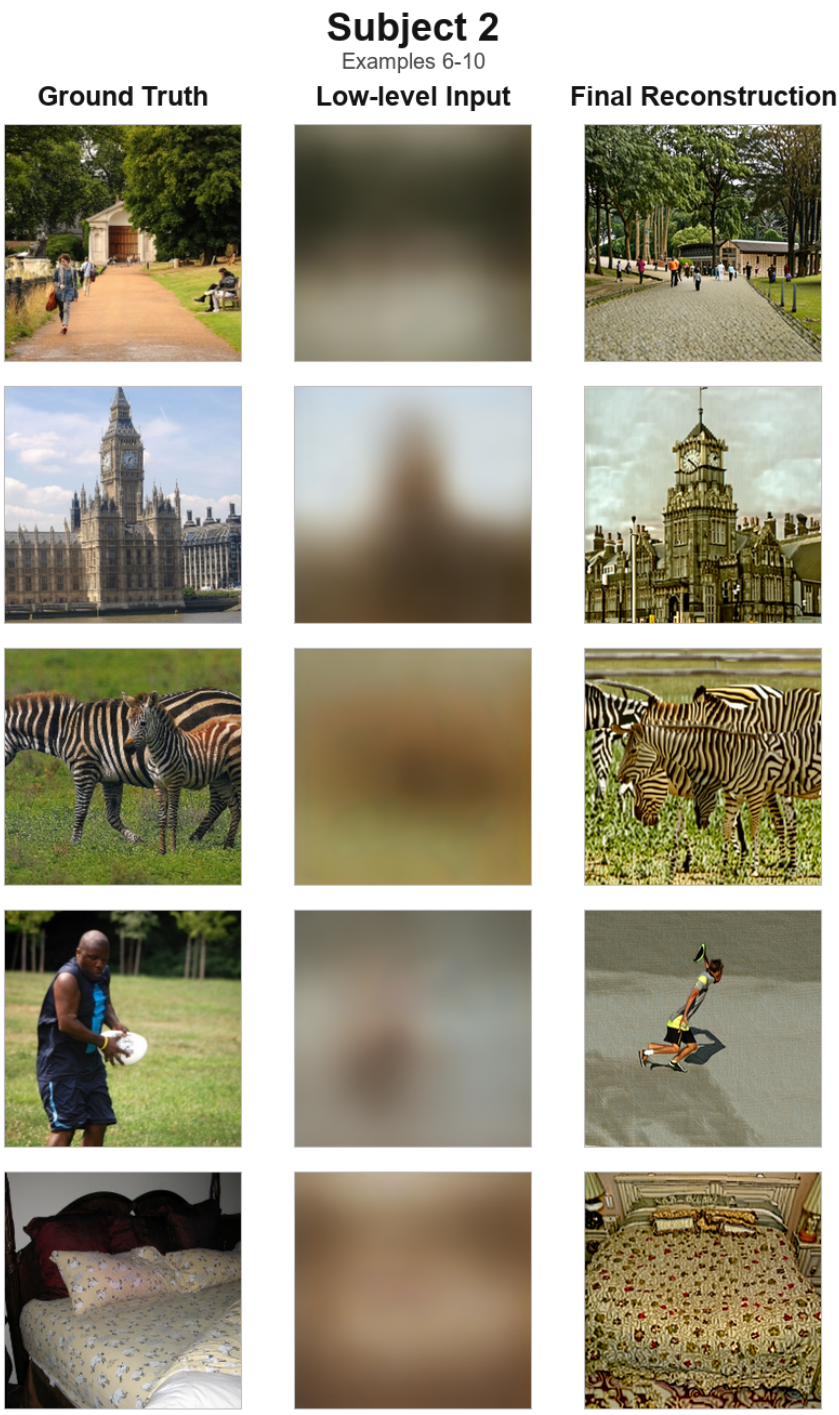


Supplementary Figure S2. Sensitivity of hybrid reconstructions to generation-stage parameters. Each row shows one held-out NSD test image, the corresponding low-level reconstruction used as the image-to-image input, and four final hybrid reconstructions generated with different denoising strength and IP-Adapter conditioning weights. The examples illustrate how changing the balance between low-level image guidance and semantic conditioning can alter both spatial layout and semantic content, even when the decoded inputs are fixed. These examples are intended as a qualitative parameter sweep rather than a replacement for the fixed reference configuration used for the reported metrics.



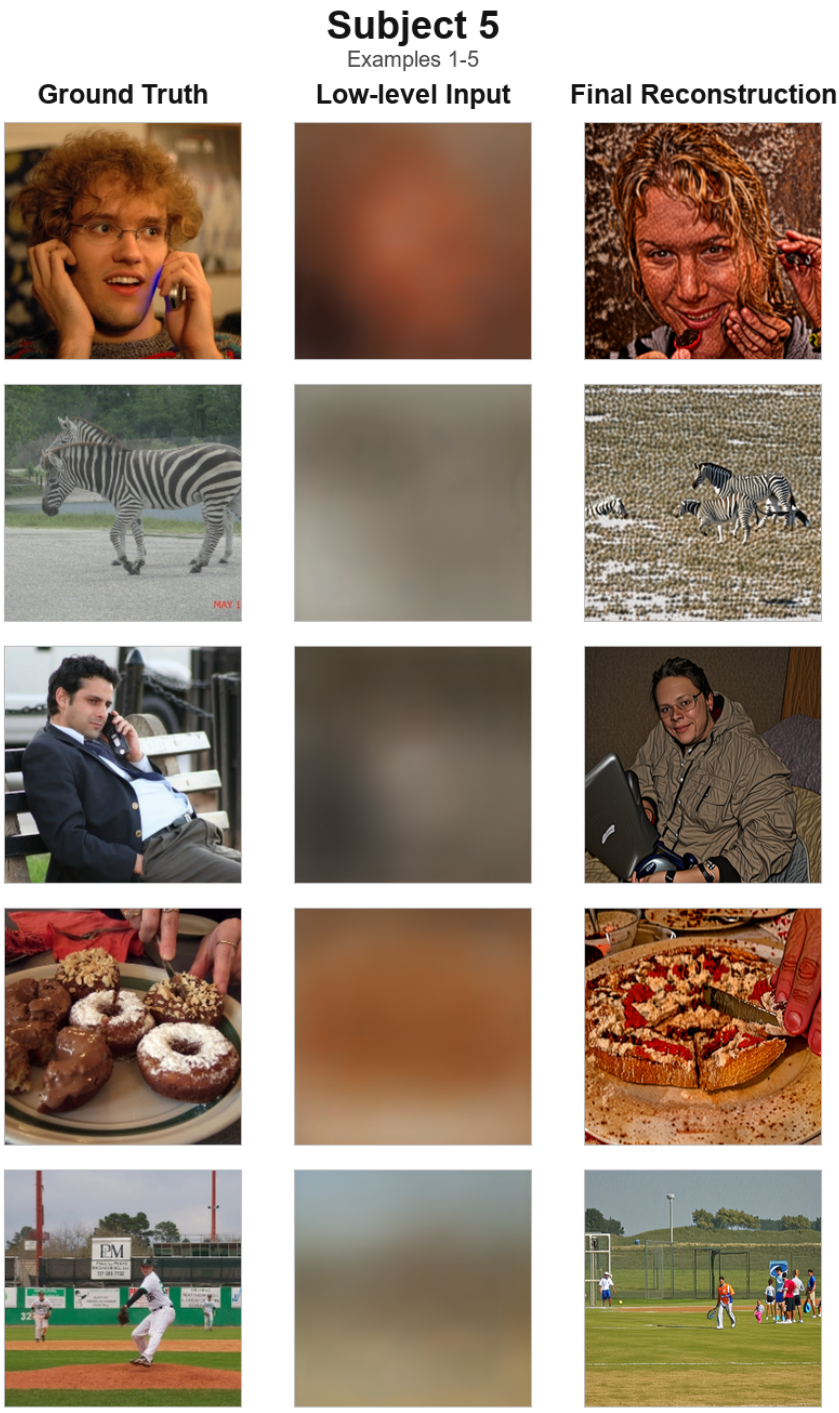
Supplementary Figure S3. Cross-subject qualitative check. Panels show all ten exported held-out test examples for each additional NSD subject using the same tutorial pipeline. For readability, each subject is split across two panels. Each row contains the ground-truth image, the low-level image-to-image input, and the final hybrid reconstruction. The examples are intended as a compact portability check beyond the reference subject, rather than a fully optimized multi-subject benchmark.

B



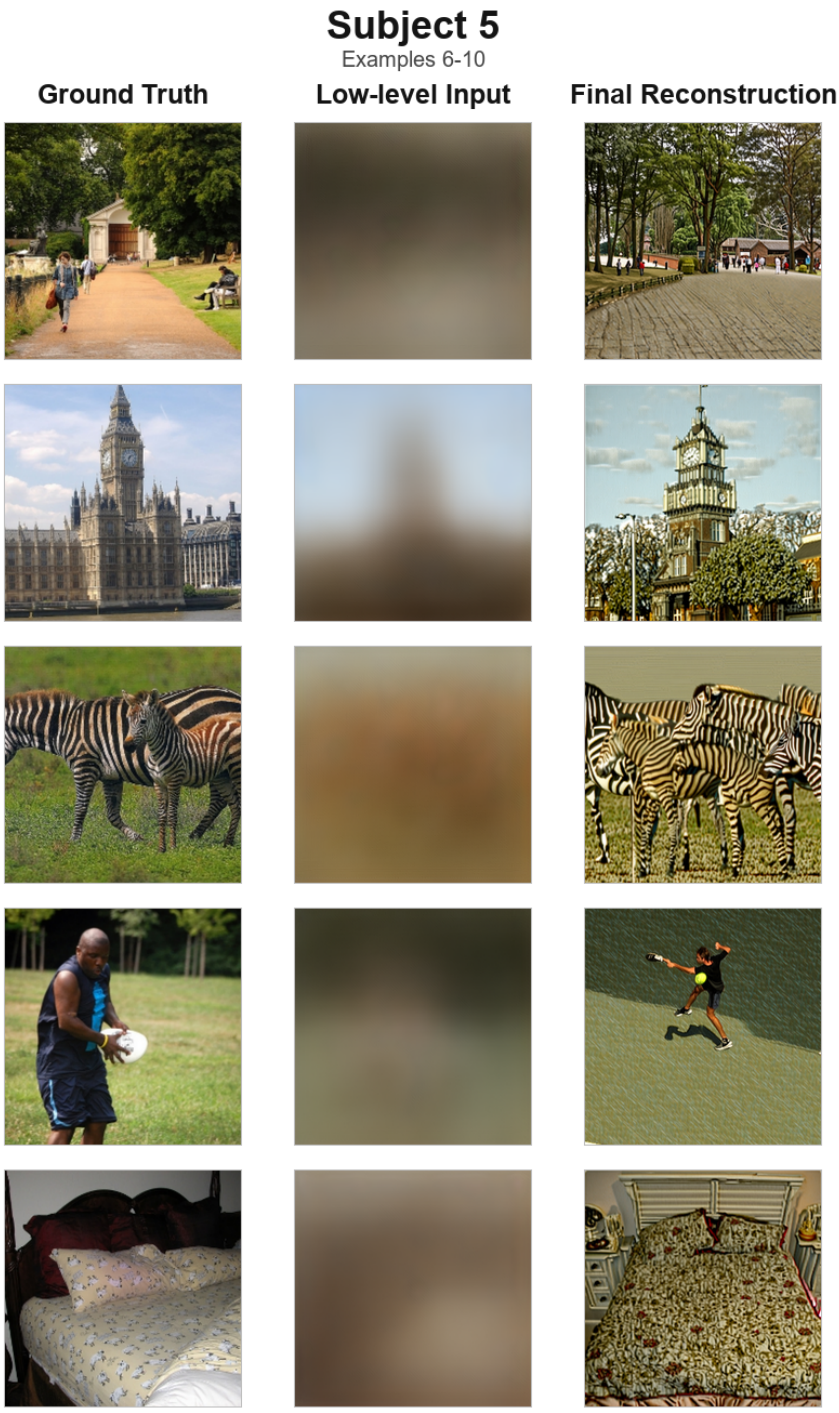
Supplementary Figure S3 continued. Remaining held-out test examples for Subject 2.

C



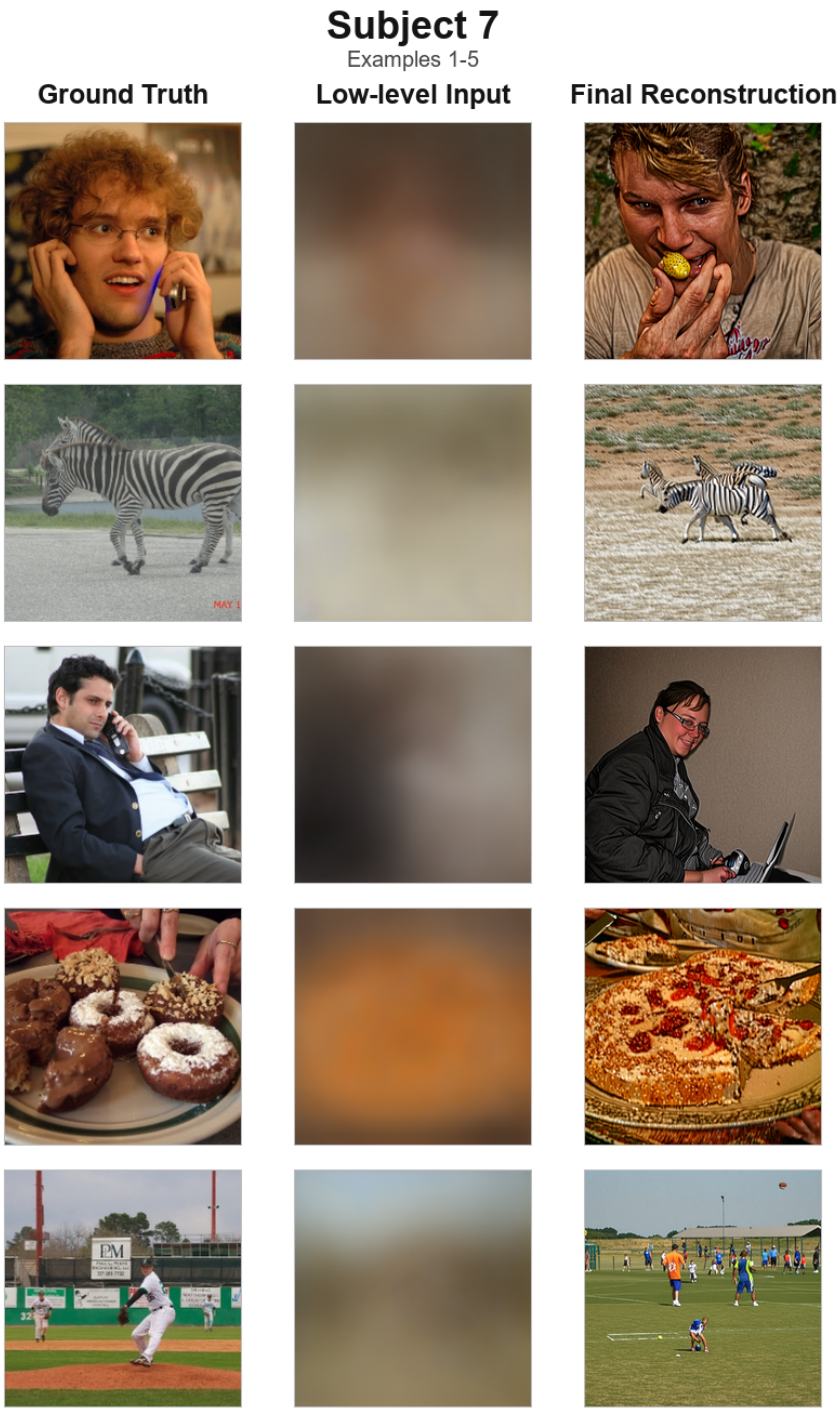
Supplementary Figure S3 continued. Held-out test examples for Subject 5.

D



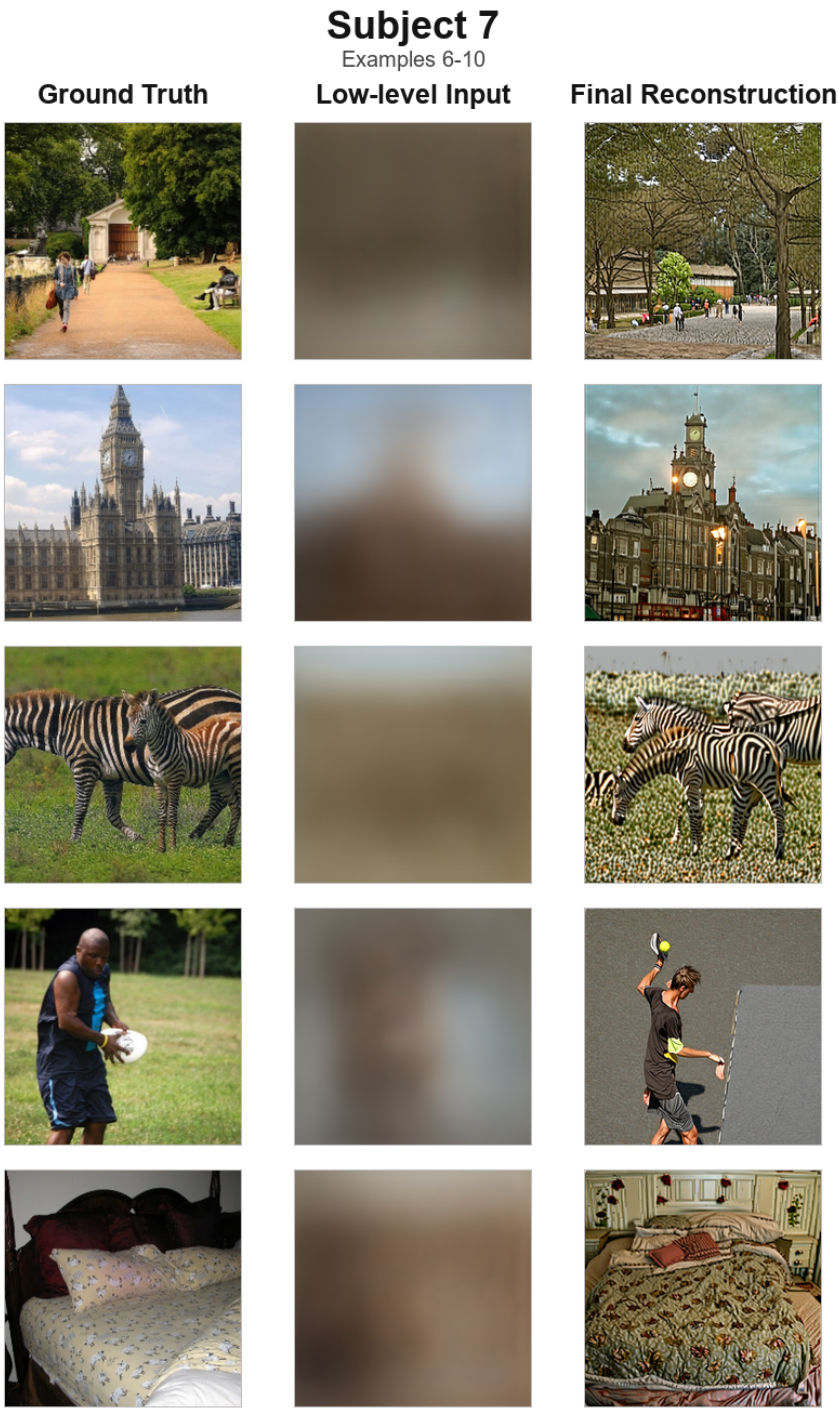
Supplementary Figure S3 continued. Remaining held-out test examples for Subject 5.

E



Supplementary Figure S3 continued. Held-out test examples for Subject 7.

F



Supplementary Figure S3 continued. Remaining held-out test examples for Subject 7.

Supplementary Table S1. Cross-subject quantitative check.

Subject	Output	PixCorr	SSIM	InceptionV3	CLIP
2	Low-level only	.393	.538	.521	.535
2	Semantic only	.144	.400	.963	.934
2	Hybrid	.311	.416	.969	.935
5	Low-level only	.316	.532	.539	.540
5	Semantic only	.138	.397	.946	.946
5	Hybrid	.290	.410	.935	.925
7	Low-level only	.346	.533	.519	.542
7	Semantic only	.131	.405	.960	.940
7	Hybrid	.286	.410	.942	.926

Note. Metrics are reported for Subjects 2, 5, and 7 over the first 100 held-out test images using the same tutorial pipeline. Values are shown for the low-level-only output, the semantic-only output, and the final hybrid reconstruction. The table is intended as a compact portability check rather than a fully optimized multi-subject benchmark. Columns report PixCorr, SSIM, InceptionV3 identification accuracy, and CLIP identification accuracy.