6 Supplementary Material

In this section, more comparisons of captions and reconstructed images are provided, compared with state-of-the-art brain decoding pipelines.

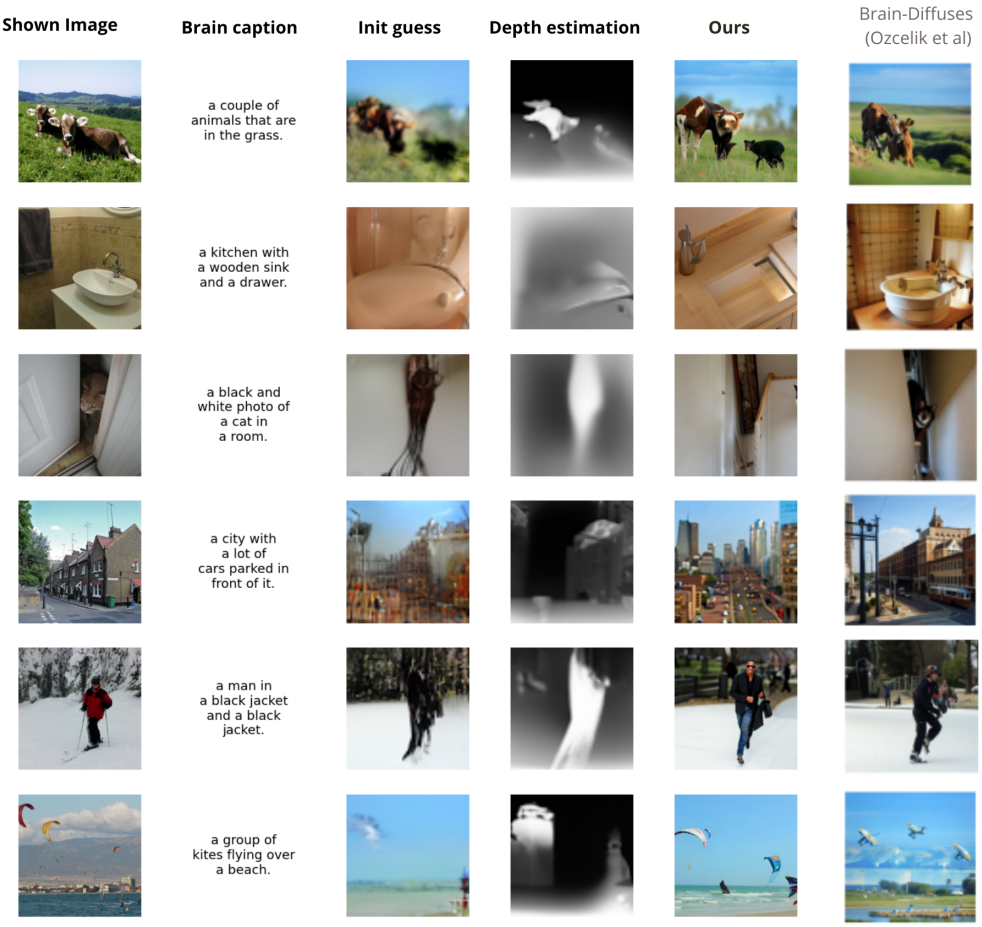

Figure A1: Comparison of our results (Columns 2-4) with the presented stimuli and other reconstruction works. The second column displays the caption derived from brain activity, the third column presents the initial guess image, the fourth column exhibits the depth-estimated images, and the fifth column showcases our final reconstruction. The last column demonstrates reconstructions from the recent BrainDiffuser work. All results are from subj01.

6.1 Ablation Study

To validate the contributions of our proposed extensions, we conducted ablation studies analyzing the impact of the depth estimation component. As shown in the attached table, we compared three model variations: 1) a baseline Stable Diffusion Img2Img pipeline using only the initial guess image, 2) a Depth2Image pipeline using only the estimated depth map, and 3) our full approach combining Stable Diffusion and ControlNet with both initial images and depth maps. Across low-level metrics like PixelCOrr and SSIM, the addition of depth information provided a consistent boost in performance. This aligns with the hypothesis that depth cues aid in capturing spatial relationships between objects and foreground-background segmentation. The full model with both initial images

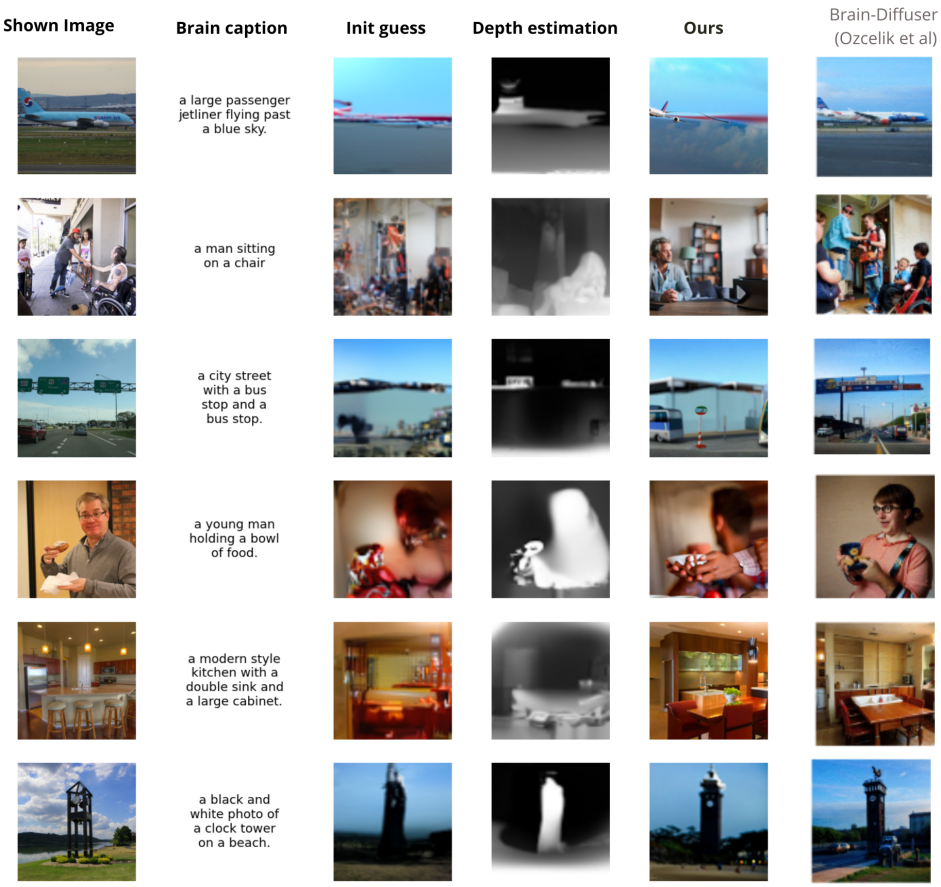

Figure A2: Comparison of our results (Columns 2-4) with the presented stimuli and other reconstruction works. The second column displays the caption derived from brain activity, the third column presents the initial guess image, the fourth column exhibits the depth-estimated images, and the fifth column showcases our final reconstruction. The last column demonstrates reconstructions from the recent BrainDiffuser work. All results are from subj01.

and depth performed the best, indicating that the two components are complementary. Qualitatively, the depth maps appeared to enhance object boundaries and 3D perspective. These results suggest that incorporating depth estimates helps the model reconstruct more accurate and realistic representations of the visual stimuli. The depth component specifically seems to benefit lower-level aspects like shapes and spatial relationships, which are critical for humans to perceive two images as highly similar Hermann et al. [2020]. By guiding the image reconstruction process with depth information extracted from brain activity, our approach can generate images that better match human perceptual judgments.

Ablation study	Low level metrics			High level		
Variant	PixCorr	SSIM	AlexNet (2)	AlexNet (5)	Inception	CLIP
Text + init	0.1204	0.1941	0.5815	0.7454	0.7974	0.8768
Stable Diffusion depth	0.3333	0.3106	0.8493	0.9654	0.8248	0.8778
ControlNet	0.**3379**	**0.3178**	**0.8707**	**0.9674**	**0.8238**	**0.8788**

Table 3: Ablation Study: Performance Metrics of Different Model Variants. Text + init is the plain Stable Diffusion Img2Img pipeline with initial guess image and captions predicted by the brain. Stable Diffusion depth is a variant pipeline that takes as input the initial guess image and captions and internally tries to estimate a depth map from the initial guess. ControlNet is external conditioning for the StableDiffusion Img2Img pipeline, so the inputs are the initial guess, the captions, and the depth maps estimated from the brain. This latter method is the one used in the paper and values (higher is better) show that this particular combination improves performance. Overall, this ablation study shows that including information about depth improves performances, particularly on low-level features.