DREAM: Visual Decoding from REversing HumAn Visual SysteM
— Supplementary Material —

Weihao Xia[1] ✉ Raoul de Charette[2] Cengiz Oztireli[3] Jing-Hao Xue[1]
[1]University College London [2]Inria [3]University of Cambridge
{weihao.xia.21,jinghao.xue}@ucl.ac.uk, raoul.de-charette@inria.fr, aco41@cam.ac.uk

This document includes further analyses on the background knowledge, experiments, and new results of our method. We first provide more details on the NSD neuroimaging dataset in Sec. 1 and extend background knowledge of the Human Visual System in Sec. 2, which together shed light on our design choices. We then detail T2I-Adapter in Sec. 3. Sec. 4 provides thorough implementation of DREAM, including architectures, representations and metrics. Finally, in Sec. 5 we further demonstrate the ability of our method with new results of cues deciphering, reconstruction, and reconstruction across subjects.

1. NSD Dataset

The Natural Scenes Dataset (NSD) [1] is currently the largest publicly available fMRI dataset. It features in-depth recordings of brain activities from 8 participants (subjects) who passively viewed images for up to 40 hours in an MRI machine. Each image was shown for three seconds and repeated three times over 30-40 scanning sessions, amounting to 22,000-30,000 fMRI response trials per participant. These viewed natural scene images are sourced from the Common Objects in Context (COCO) dataset [11], enabling the utilization of the original COCO captions for training.

The fMRI-to-image reconstruction studies that used NSD [8, 16, 22] typically follow the same procedure: training individual-subject models for the four participants who finished all scanning sessions (participants 1, 2, 5, and 7), and employing a test set that corresponds to the common 1,000 images shown to each participant. For each participant, the training set has 8,859 images and 24,980 fMRI tests (as each image being tested up to 3 times). Another 982 images and 2,770 fMRI trials are common across the four individuals. We use the preprocessed fMRI voxels in a 1.8-mm native volume space that corresponds to the "nsd-general" brain region. This region is described by the NSD authors as the subset of voxels in the posterior cortex that are most responsive to the presented visual stimuli. For fMRI data spanning multiple trials, we calculate the average response as in prior research [12]. Tab. 1 details the characteristics of the NSD dataset and the region of interests (ROIs) included in the fMRI data.

Table 1. **Details of the NSD dataset.**

Training	Test	ROIs	Subject ID	Voxels
8859	982	V1, V2, V3, hV4, VO, PHC, MT, MST, LO, IPS	sub01	11694
			sub02	9987
			sub05	9312
			sub07	8980

2. Detailed Human Visual System

Our approach aims to decode semantics, color, and depth from fMRI data, thus inherently bounded by the ability of fMRI data to capture the ad hoc brain activities. It is crucial to ascertain whether fMRI captures the alterations in the respective human brain regions responsible for processing the visual information. Here, we provide a comprehensive examination of the specific brain regions in the human visual system recorded by the fMRI data.

The flow of visual information [2] in neuroscience is presented as follows. Fig. 1 presents a comprehensive depiction of the functional anatomy of the visual perception. Sensory input originating from the Retina travels through the LGN in the thalamus and then reaches the Visual Cortex. **Retina** is a layer within the eye comprised of photoreceptor and glial cells. These cells capture incoming photons and convert them into electrical and chemical signals, which are then relayed to the brain, resulting in visual perception. Different types of information are processed through the parvocellular and magnocellular pathways, details of which are elaborated in the main paper. **LGN** then channels the conveyed visual information into the **Visual Cortex**, where it diverges into two streams in **Visual Association Cortex** (VAC) for undertaking intricate processing of high-level semantic contents from the visual image.

The Visual Cortex, also known as visual area 1 (V1), serves as the initial entry point for visual perception within

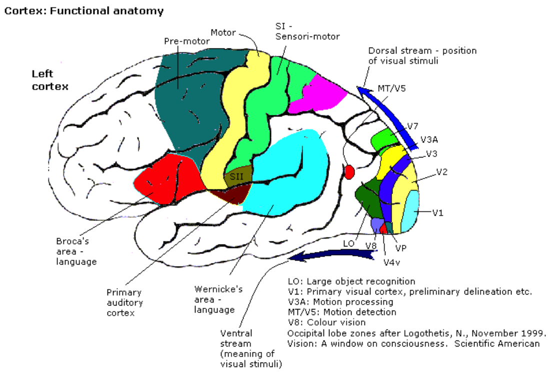
Cortex: Functional anatomy
Left cortex
Pre-motor
Motor
SI - Sensori-motor
Dorsal stream - position of visual stimuli
MT/V5
V7
V3A
V3
V2
V1
SII
LO
V8
VP
V4v
Broca's area - language
Primary auditory cortex
Wernicke's area - language
Ventral stream (meaning of visual stimuli)
LO: Large object recognition
V1: Primary visual cortex, preliminary delineation etc.
V3A: Motion processing
MT/V5: Motion detection
V8: Colour vision
Occipital lobe zones after Logothetis, N., November 1999.
Vision: A window on consciousness. Scientific American

Figure 1. **Functional Anatomy of Cortex.** The functional localization in the human brain is based on findings from functional brain imaging, which link various anatomical regions of the brain to their associated functions.
Source: Wikimedia Commons. This image is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.

the cortex. Visual information flows here first before being relayed to other regions. VAC comprises multiple regions surrounding the visual cortex, including V2, V3, V4, and V5 (also known as the middle temporal area, MT). V1 transmits information into two primary streams: the ventral stream and the dorsal stream.

- The ventral stream (black arrow) begins with V1, goes through V2 and V4, and to the inferior temporal cortex (IT cortex). The ventral stream is responsible for the "meaning" of the visual stimuli, such as object recognition and identification.
- The dorsal stream (blue arrow) begins with V1, goes through visual area V2, then to the dorsomedial area (DM/V6) and medial temporal area (MT/V5) and to the posterior parietal cortex. The dorsal stream is engaged in analyzing information associated with "position", particularly the spatial properties of objects.

After juxtaposing the explanations illustrated in Fig. 1 with the collected information demonstrated in Tab. 1, it becomes apparent that the changes occurring in brain regions linked to the processing of semantics, color, and depth are indeed present within the fMRI data. This observation emphasizes the capability to extract the intended information from the provided fMRI recordings.

3. T2I-Adapter

T2I-Adapter [14] and ControlNet [27] learn versatile modality-specific encoders to improve the control ability of text-to-image SD model [19]. These encoders extract guidance features from various conditions $\mathbf{y}$ (*e.g.* sketch, semantic label, and depth). They aim to align external control with internal knowledge in SD, thereby enhancing the precision of control over the generated output. Each encoder $\mathcal{R}$ produces n hierarchical feature maps $\mathrm{F}_{\mathcal{R}}^{i}$ from the primitive condition $\mathbf{y}$. Then each $\mathrm{F}_{\mathcal{R}}^{i}$ is added with the corresponding intermediate feature $\mathrm{F}_{\mathrm{SD}}^{i}$ in the denoising U-Net encoder:

$$\begin{aligned} \mathrm{F}_{\mathcal{R}} &= \mathcal{R}(\mathbf{y}), \\ \hat{\mathrm{F}}_{\mathrm{SD}}^{i} &= \mathrm{F}_{\mathrm{SD}}^{i} + \mathrm{F}_{\mathcal{R}}^{i}, \quad i \in \{1, 2, \cdots, n\}. \end{aligned} \tag{1}$$

T2I-Adapter consists of a pretrained SD model and several adapters. These adapters are used to extract guidance features from various conditions. The pretrained SD model is then utilized to generate images based on both the input text features and the additional guidance features. The CoAdapter mode becomes available when multiple adapters are involved, and a composer processes features from these adapters before they are further fed into the SD. Given the deciphered semantics, color, and depth information from fMRI, we can reconstruct the final images using the color and depth adapters in conjunction with SD.

4. Implementation Details

4.1. Network Architectures

The fMRI ↦ Semantics encoder $\mathcal{E}_{\mathtt{fmri}}$ maps fMRI voxels to the shared CLIP latent space [17] to decipher semantics. The network architecture includes a linear layer followed by multiple residual blocks, a linear projector, and a final MLP projector, akin to previous research [4, 20]. The learned embedding is with a feature dimension of 77×768, where 77 denotes the maximum token length and 768 represents the encoding dimension of each token. It is then fed into the pretrained Stable Diffusion [19] to inject semantic information into the final reconstructed images.

The fMRI ↦ Depth & Color encoder $\mathcal{E}$ and decoder $\mathcal{D}$ decipher depth and color information from the fMRI data. Given that spatial palettes are generated by first downsampling (with bicubic interpolation) an image and then upsampling (with nearest interpolation) it back to its original resolution, the primary objective of the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ shifts towards predicting RGBD images from fMRI data. The architecture of $\mathcal{E}$ and $\mathcal{D}$ is built on top of [7], with inspirations drawn from VDVAE [5].

4.2. Representations of Semantics, Color and Depth

This section serves as an introduction to the possible choices of representations for semantics, color, and depth. We currently use CLIP embedding, depth map [18], and spatial color palette [14] to facilitate subsequent processing of T2I-Adapter [14] in conjunction with a pretrained Stable Diffusion [19] for image reconstruction from deciphered cues. However, there are other possibilities that can be utilized within our framework.

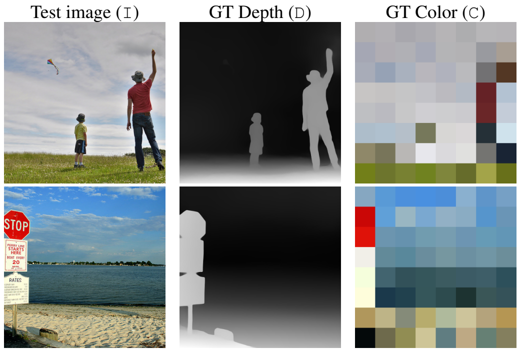

Figure 2. **Depth and Color Representations.** We present pseudo ground truth samples of Depth (MiDaS prediction [18]) and Color (×64 downsampling of the test image) for a NSD input image.

Semantics. The Stable Diffusion utilizes a frozen CLIP ViT-L/14 text encoder to condition the model on text prompts. It is with a feature space dimension of 77×768, where 77 denotes the maximum token length and 768 represents the encoding dimension of each token. The CLIP ViT-L/14 image encoder is with a feature space dimension of 257×768. We maps flattened voxels to an intermediate space of size 77×768, corresponding to the last hidden layer of CLIP ViT/L-14. The learned embeddings inject semantic information into the reconstructed images.

Depth. We select depth as the structural guidance for two main reasons: alignment with the human visual system, and better performance demonstrated in our preliminary experiments. Following prior research [14, 27], we use the MiDaS predictions [18] as the surrogate ground truth depth maps, which are visualized in Fig. 2.

Color. There are many representations that can provide the color information, such as histogram and probabilistic palette [9, 23] However, ControlNet [27] and T2I-Adapter [14] only accept spatial inputs, which leaves no alternative but to utilize the *spatial color palettes* as the color representation. In practice, spatial color palettes resemble coarse resolution images, as seen in Fig. 2, and are generated by first ×64 downsampling (with bicubic interpolation) an image and then upsampling (with nearest interpolation) it back to its original resolution.

During the image reconstruction phase, the spatial palettes contribute the color and appearance information to the final images. These spatial palettes are derived from the image estimated by the RGBD decoder in R-PKM. We refer to the images produced at this stage as the "initial guessed image" to differentiate them from the final reconstruction. The initial guessed image offers color cues but it also contains inaccuracies. By employing a ×64 downsampling, we can effectively extract necessary color details from this image while minimizing the side effects of inaccuracies.

Other Guidance. In the realm of visual decoding with pre-trained diffusion models [14, 26, 27], any guidance available in these models can be harnessed to fill in gaps of missing information, thereby enhancing performance. This spatial guidance includes representations such as sketch [21], Canny edge detection, HED (Holistically-Nested Edge Detection) [25], and semantic segmentation maps [3]. These alternatives could potentially serve as the intermediate representations for the reverse pathways in our method. HED and Canny are edge detectors, which provide object boundaries within images. However, during our preliminary experiments, both methods were shown to face challenges in providing reliable edges for all images. Sketches encounter similar difficulties in providing reliable guidance. The semantic segmentation map provides both structural and semantic cues. However, it overlaps in function with CLIP semantics and depth maps, and leads to diminished performance gain on top of the other two representations.

4.3. Evaluation Methodology

Metrics for Visual Decoding. For visual decoding metrics, we employ the same suite of eight evaluation criteria as previously used in research [8, 16, 20, 22]. PixCorr, SSIM, AlexNet(2), and AlexNet(5) are categorized as low-level, while Inception, CLIP, EffNet-B, and SwAV are considered high-level. Following [16], we downsampled the generated images from a 512×512 resolution to a 425×425 resolution (corresponding to the resolution of ground truth images in the NSD dataset) for PixCorr and SSIM metrics. For the other metrics, the generated images were adjusted based on the input specifications of each respective network. It should be noted that not all evaluation outcomes are available for earlier models, depending on the metrics they chose to experiment with. Our quantitative comparisons with MindEye [20], Takagi *et al.* [22], and Gu *et al.* [8] are made according to the exact same test set, *i.e.*, the 982 images that are shared for all 4 subjects. Lin *et al.* [10] disclosed their findings exclusively for Subject 1, with a custom training-test dataset split.

Metrics for Depth and Color. We additionally measure consistency of our extracted depth and color. We borrow some common metrics from depth estimation [13] and color correction [24] to assess depth and color consistencies in the final reconstructed images. For depth metrics, we report Abs Rel (absolute error), Sq Rel (squared error), RMSE (root mean squared error), and RMSE log (root mean squared logarithmic error) — detailed in [13].

For color metrics, we use CD (Color Discrepancy) [24] and STRESS (Standardized Residual Sum of Squares) [6]. CD calculates the absolute differences between the ground truth I and the reconstructed image $\hat{I}$ by utilizing the nor-

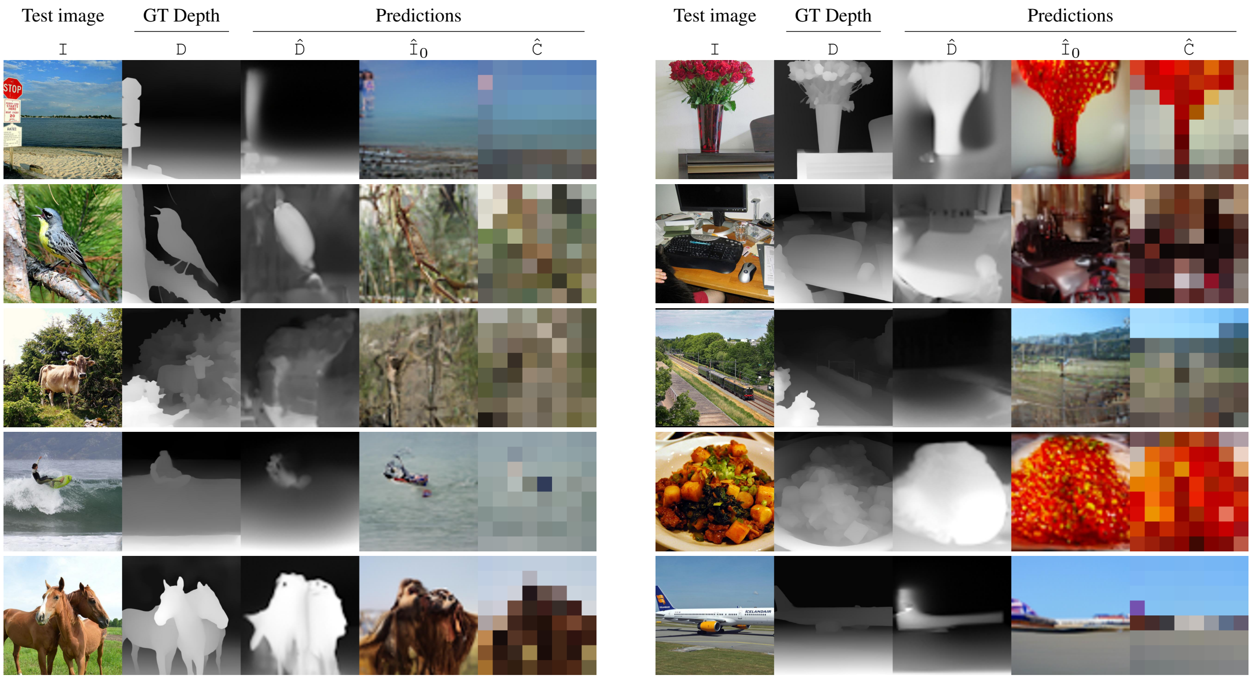

Figure 3. **DREAM Decoding of Depth and Color.** We display the test image corresponding to fMRI, alongside the depth ground truth ($\mathtt{D}$) and the depth/color predictions ($\hat{\mathtt{D}}, \hat{\mathtt{C}}$). The R-PKM component predicts depth maps and the initial guessed RGB images ($\hat{\mathtt{I}}_0$). The predicted spatial palettes are derived from these initial guessed images. The results highlight the proficiency of our R-PKM module in capturing and converting intricate aspects from fMRI recordings into essential cues for visual reconstructions.

malized histograms of images segmented into bins:

$$\mathrm{CD}(I, \hat{I}) = \sum \left| \mathcal{H}(I) - \mathcal{H}(\hat{I}) \right|, \tag{2}$$

where $\mathcal{H}(\cdot)$ represents the histogram function over the given range (*e.g*. [0, 255]) and number of bins. In simpler terms, this equation computes the absolute difference between the histograms of the two images for all bins and then sums them up. The number of bins for histogram is set as 64. STRESS calculates a scaled difference between the ground-truth $\mathtt{C}$ and the estimated color palette $\hat{\mathtt{C}}$:

$$\mathrm{STRESS} = 100\sqrt{\frac{\sum_{i=1}^{n}\left(F\hat{\mathtt{C}}_i - \mathtt{C}_i\right)^2}{\sum_{i=1}^{n}\mathtt{C}_i^2}}, \tag{3}$$

where n is the number of samples and F is calculated as

$$F = \frac{\sum_{i=1}^{n}\hat{\mathtt{C}}_i\mathtt{C}_i}{\sum_{i=1}^{n}\hat{\mathtt{C}}_i^2}. \tag{4}$$

5. Additional DREAM Results

This section presents additional results of our method, to showcase the effectiveness of DREAM. Sec. 5.1 presents the fMRI ↦ depth & color results, which demonstrates how the deciphered and represented color and depth information helps to boost the performance of visual decoding. Sec. 5.2 provides more examples of fMRI test reconstructions from subject 1. The results shows that the extracted essential cues from fMRI recordings lead to enhanced consistency in appearance, structure, and semantics when compared to the viewed visual stimuli. Sec. 5.3 provides results of all four subjects.

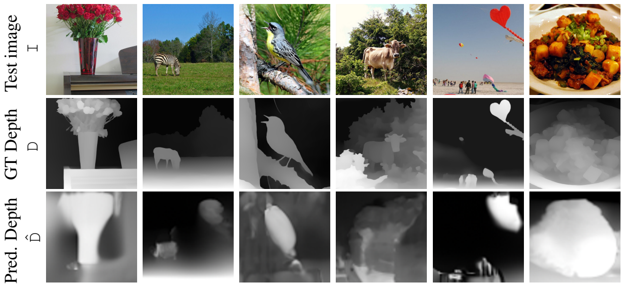

Figure 4. **Sample Depth**. We show sample depth maps ($\hat{\mathtt{D}}$) deciphered from fMRI using R-PKM, alongside the ground-truth depth ($\mathtt{D}$) estimated from MiDaS [18] on the original test image (I).

5.1. Depth & Color Deciphering

Fig. 3 showcases additional depth and color results deciphered from the R-PKM component. Overall, it is able to capture and translate these intricate aspects from fMRI recordings to spatial guidance crucial for more accurate im-

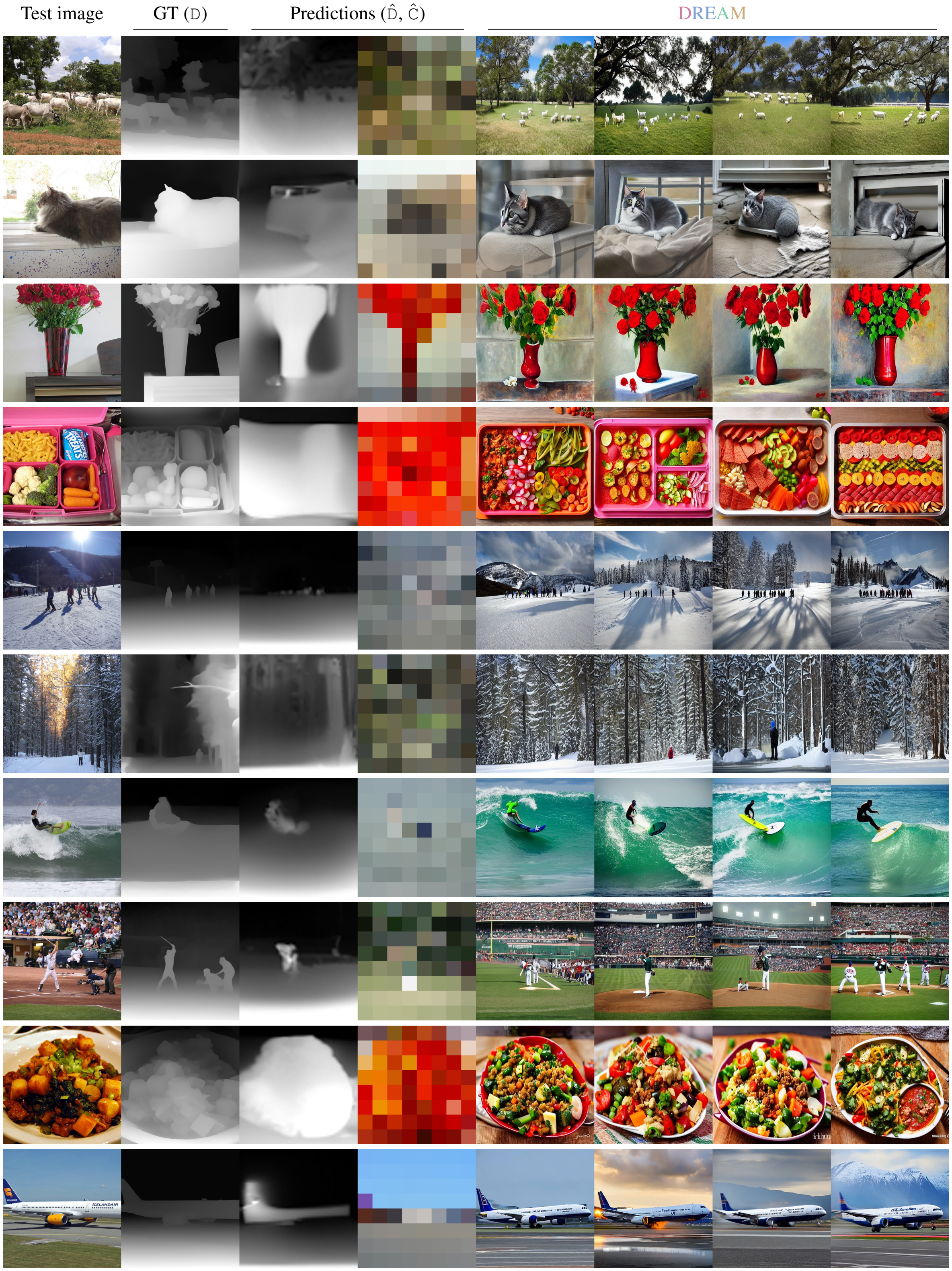

Figure 5. **DREAM Reconstructions.** We show reconstruction for subject 1 (sub01) from the NSD dataset. Our approach extracts essential cues from fMRI recordings, leading to enhanced consistency in appearance, structure, and semantics when compared to the viewed visual stimuli. The results are randomly selected. The illustrated depth, color, and final images demonstrate that the deciphered and represented color and depth cues help to boost the performance of visual decoding.

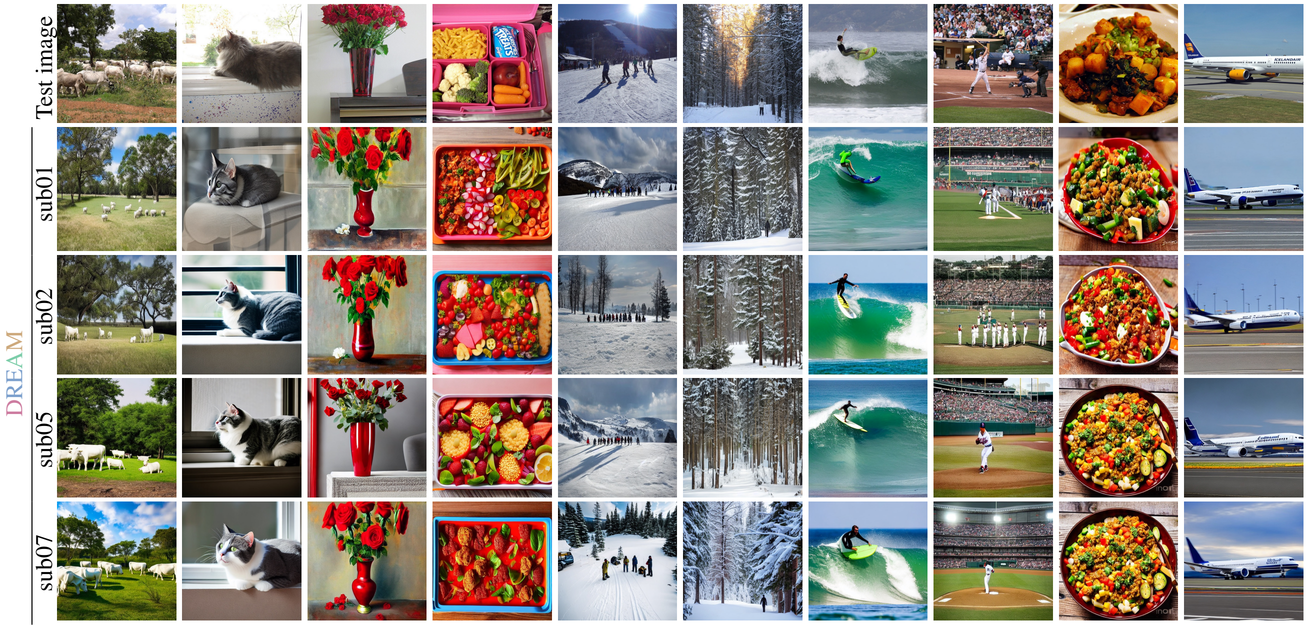

Figure 6. **Subject-Specific Results.** We visualize subject-specific outputs of DREAM on the NSD dataset. For each subject, the model is retrained because the brain activity varies across subjects. Overall, it consistently reconstructs the test image for all subjects while we note that some reconstruction inaccuracies are shared across subjects (cf. Sec. 5.3). Quantitative metrics are in Tab. 2.

Table 2. **Subject-Specific Evaluation.** Quantitative evaluation of the DREAM reconstructions for the participants (sub01, sub02, sub05, and sub07) of the NSD dataset. Performance is stable accross all participants, and consistent with the results reported in the main paper. Some example visual results can be found in Fig. 6.

Subject	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
sub01	.288	.338	95.0%	97.5%	94.8%	95.2%	.638	.413
sub02	.273	.331	94.2%	97.1%	93.4%	93.5%	.652	.422
sub05	.269	.325	93.5%	96.6%	93.8%	94.1%	.633	.397
sub07	.265	.319	92.7%	95.4%	92.6%	93.7%	.656	.438

age reconstructions.

For depth, the second and third columns show example depth reconstruction alongside their corresponding estimated ground truth obtained from the original RGB images. Results show that the depth estimated, while far from perfect, is sufficient to provide coarse guidance on the scene structure and object position/orientation for our reconstruction guidance purpose.

The last two columns show the color results. The predicted spatial palettes are generated by downscaling the "initial guessed images" denoted $\hat{\mathtt{I}}_0$ (not to be confused with $\hat{\mathtt{I}}$) which corresponds to the RGB channels of the R-PKM RGBD output. As discussed in Sec. 4.2, employing a $\times 64$ downsampling on the "initial guessed images" achieves a trade-off between efficiently extracting essential color cues and effectively mitigating the inaccuracies in these images. Despite not accurately preserving the color of local regions due to the resolution, the produced color palettes provide a relevant constraint and guidance on the overall color tone. Additional depth outputs are in Fig. 4.

Although depth and color guidance are sufficient to reconstruct images reasonably resembling the test one, it is yet unclear if better depth and color cues can be extracted from the fMRI data or if depth and color are doomed to be coarse estimation due to loss of data in the fMRI recording.

5.2. More Image Reconstruction Results

More examples of image reconstruction for subject 1 are shown in Fig. 5. From left to right: the first two columns display the test images and their corresponding ground truth depth maps. The third and fourth columns depict the predicted depth and color, respectively, in the form of depth maps and spatial palettes. The remaining columns represent the final reconstructed images. The results are randomly selected. The illustrated final images demonstrate that the deciphered and represented color and depth cues help to boost the performance of visual decoding. Overall, DREAM evidently extracts good-enough cues from the fMRI recordings, leading to consistent reconstruction of the appearance, structure, and semantics of the viewed visual stimuli.

5.3. Subject-Specific Results

We used the same standardized training-test data splits as other NSD reconstruction papers [15, 16, 20], training subject-specific models for each of 4 participants (sub01, sub02, sub05, and sub07). More details on the different participants can be found in Sec. 1 and Tab. 1. Fig. 6 shows DREAM outputs for all four participants, with individual subject evaluation metrics reported in Tab. 2. More sub01 results can be found in Fig. 5. Overall, DREAM proves to work well regardless of the subject. However, it is interesting to note that some reconstructions mistakes are shared across subjects. For example, fMRIs of the *vase flowers* picture (3rd column) are often reconstructed as paintings, except for sub05, and the *food plate* (2nd rightmost column) which is taken at an angle is almost always reconstructed as a more top-view photography. These consistent mistakes across subjects may suggest dataset biases.

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. 1

[2] Per Brodal. *The central nervous system: structure and function*. oxford university Press, 2004. 1

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 3

[4] 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 *CVPR*, pages 22710–22720, 2023. 2

[5] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. In *ICLR*, 2021. 2

[6] Pedro A Garcia, Rafael Huertas, Manuel Melgosa, and Guihua Cui. Measurement of the relationship between perceived and computed color differences. *JOSA A*, 24(7):1823–1829, 2007. 3

[7] Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022. 2

[8] Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. Decoding natural image stimuli from fmri data with a surface-based convolutional network. In *MIDL*, 2023. 1, 3

[9] Guillaume Le Moing, Tuan-Hung Vu, Himalaya Jain, Patrick Pérez, and Matthieu Cord. Semantic palette: Guiding scene generation with class proportions. In *CVPR*, pages 9342–9350, 2021. 3

[10] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind Reader: Reconstructing complex images from brain activities. *NeurIPS*, 35:29624–29636, 2022. 3

[11] 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 *ECCV*, pages 740–755, 2014. 1

[12] Yizhuo Lu, Changde Du, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. *arXiv preprint arXiv:2303.14139*, 2023. 1

[13] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021. 3

[14] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 3

[15] Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *IJCNN*, pages 1–8. IEEE, 2020. 7

[16] Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of Perceived Images from fMRI Patterns and Semantic Brain Exploration using Instance-Conditioned GANs. In *IJCNN*, pages 1–8. IEEE, 2022. 1, 3, 7

[17] 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 *ICML*, pages 8748–8763. PMLR, 2021. 2

[18] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 44(3):1623–1637, 2020. 2, 3, 4

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[20] Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *arXiv preprint arXiv:2305.18274*, 2023. 2, 3, 7

[21] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *ICCV*, pages 5117–5127, 2021. 3

[22] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, pages 14453–14463, 2023. 1, 3

[23] Yi Wang, Menghan Xia, Lu Qi, Jing Shao, and Yu Qiao. Palgan: Image colorization with palette generative adversarial networks. In *ECCV*, pages 271–288. Springer, 2022. 3

[24] Menghan Xia, Jian Yao, Renping Xie, Mi Zhang, and Jinsheng Xiao. Color consistency correction based on remapping optimization for image stitching. In *ICCV workshops*, pages 2977–2984, 2017. 3

[25] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, pages 1395–1403, 2015. 3

[26] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022. 3

[27] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 3