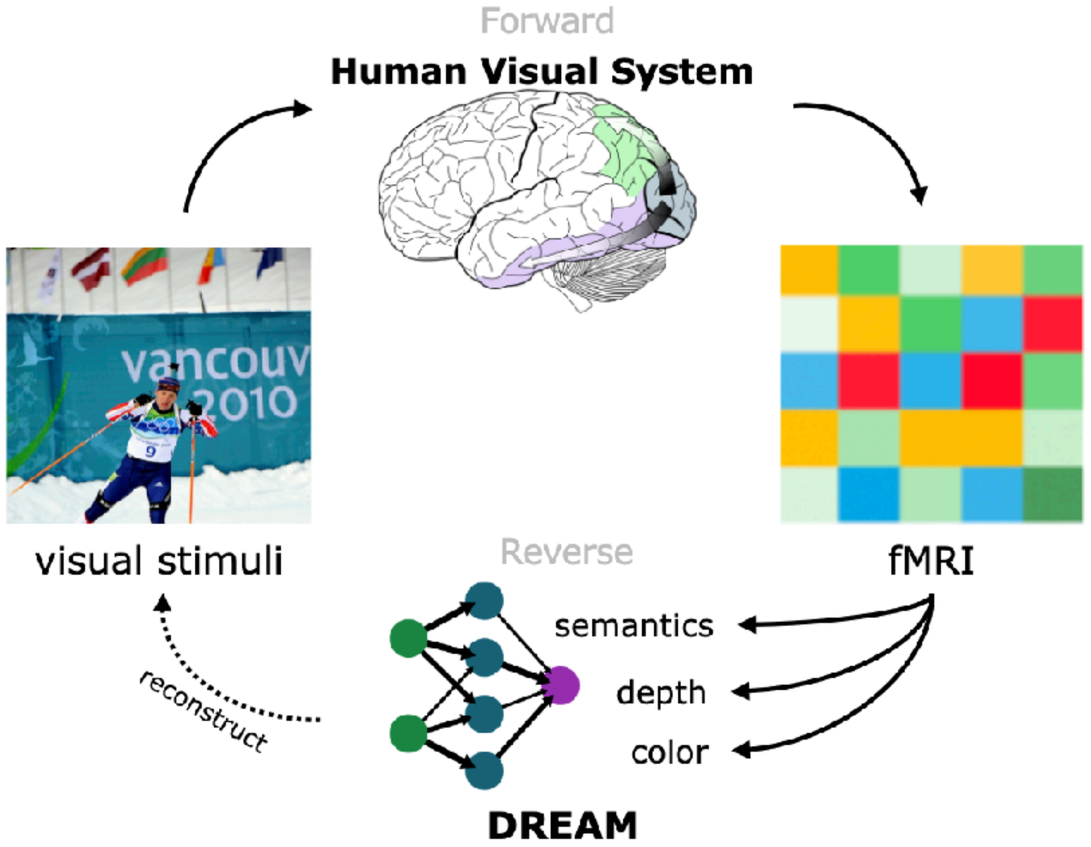
DREAM: Visual Decoding from REversing HumAn Visual SysteM

Paper Review

fMRI-to-image method

Javier Salazar Cavazos - Feb 11, 2025

The Premise



Abstract

In this work we present DREAM, an fMRI-to-image method for reconstructing viewed images from brain activities, grounded on fundamental knowledge of the human visual system. We craft reverse pathways that emulate the hierarchical and parallel nature of how humans perceive the visual world. These tailored pathways are specialized to decipher semantics, color, and depth cues from fMRI data, mirroring the forward pathways from visual stimuli to fMRI recordings.

Related Works

From our group

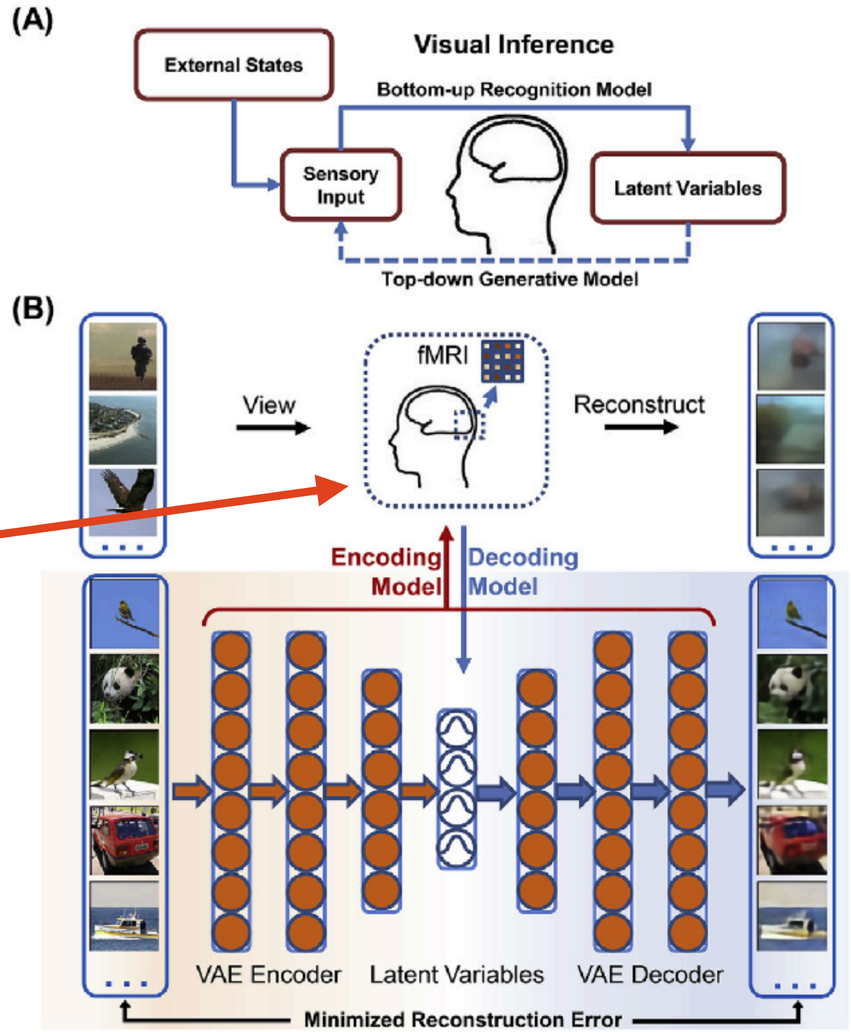
Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex

Kuan Han [b,c], Haiguang Wen [b,c], Junxing Shi [b,c], Kun-Han Lu [b,c], Yizhen Zhang [b,c], Di Fu [b,c], Zhongming Liu [a,b,c,*]

[a] *Weldon School of Biomedical Engineering, USA*
[b] *School of Electrical and Computer Engineering, USA*
[c] *Purdue Institute for Integrative Neuroscience, Purdue University, West Lafayette, IN, 47906, USA*

May 2019

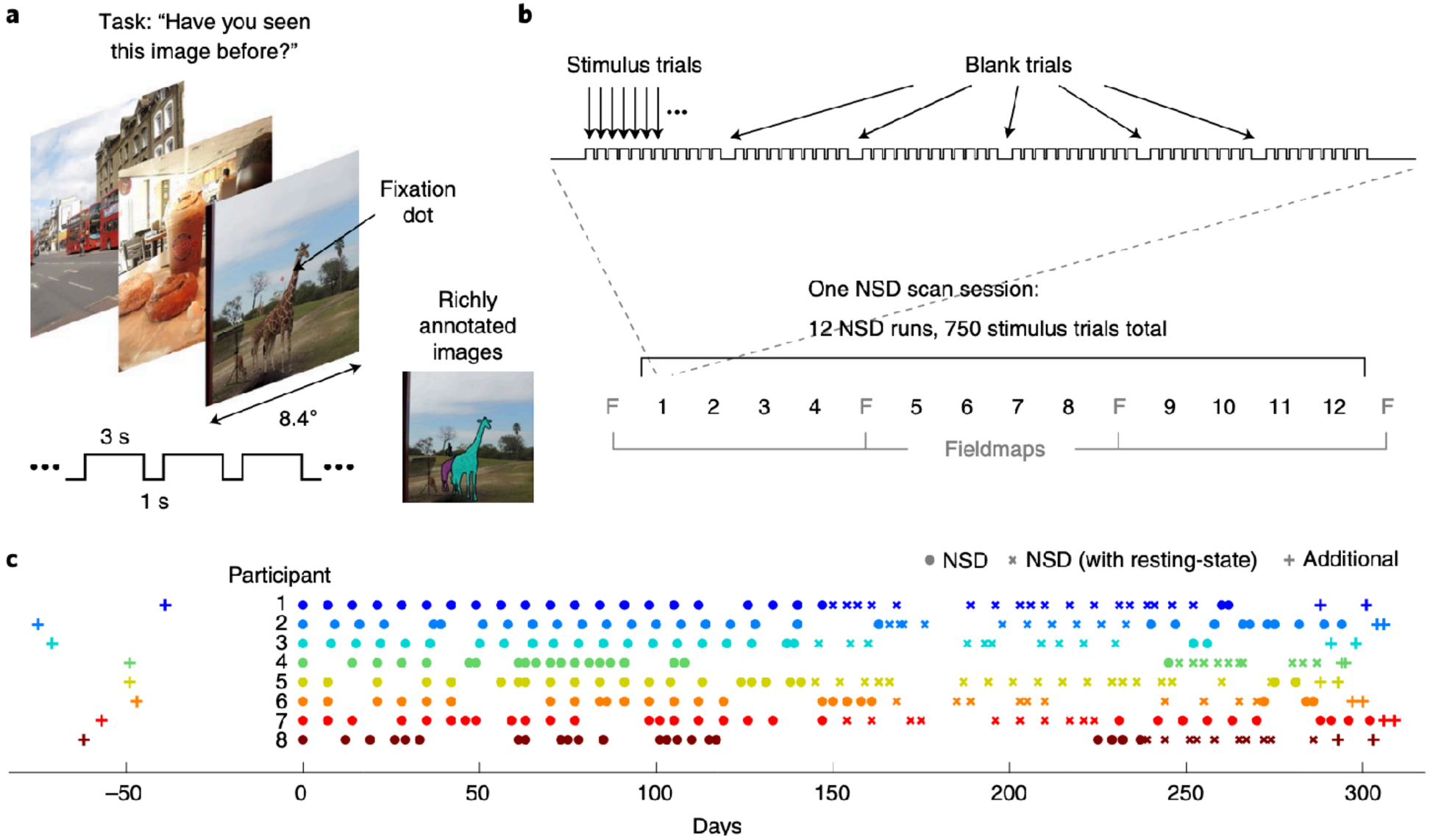
$$y_i = \boldsymbol{w}_i^T \mathbf{z} + b_i + \varepsilon_i$$



Related Works (VAE Results)

Dataset

NDS Dataset

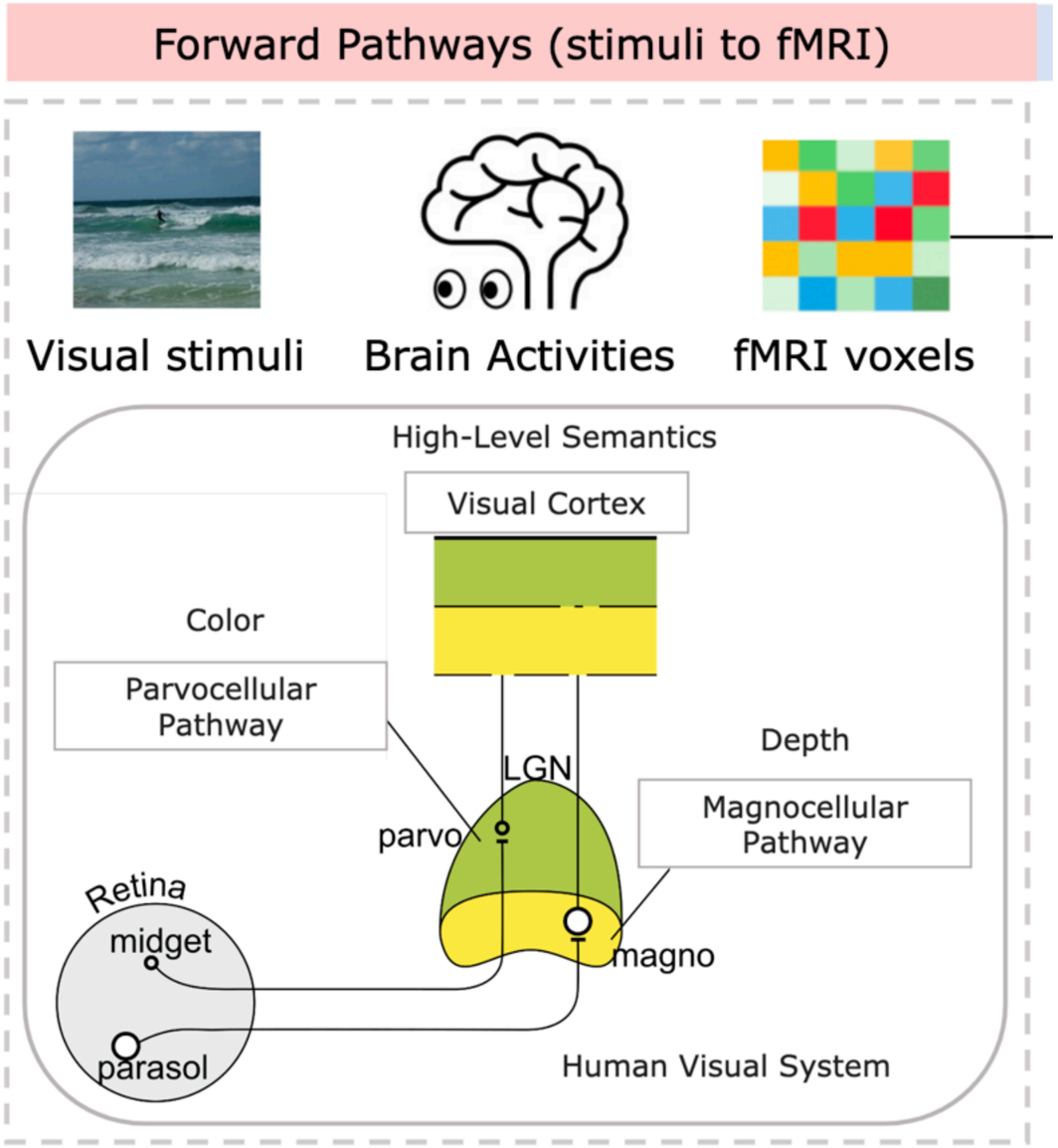

a
Task: "Have you seen this image before?"
Fixation dot
8.4°
3 s
1 s
Richly annotated images
b
Stimulus trials
Blank trials
One NSD scan session:
12 NSD runs, 750 stimulus trials total
F 1 2 3 4 F 5 6 7 8 F 9 10 11 12 F
Fieldmaps
c
NSD
NSD (with resting-state)
Additional
Participant
1
2
3
4
5
6
7
8
−50
0
50
100
150
200
250
300
Days

Dataset

- Natural Scenes Dataset (NDS) contains 8 subjects viewed images for 40 hours
- Each image was shown for 3 seconds and repeated three times over 30-40 sessions
- ~22k-30k functional MRI response trials for the ~8k images
- Images are from COCO dataset so captions are included with images
- Only 982 images common among subjects (982*3 fMRI sessions)
- BME Details: 1.8mm voxel size, 7T scanner, gradient echo, EPI, 1.6sec TR

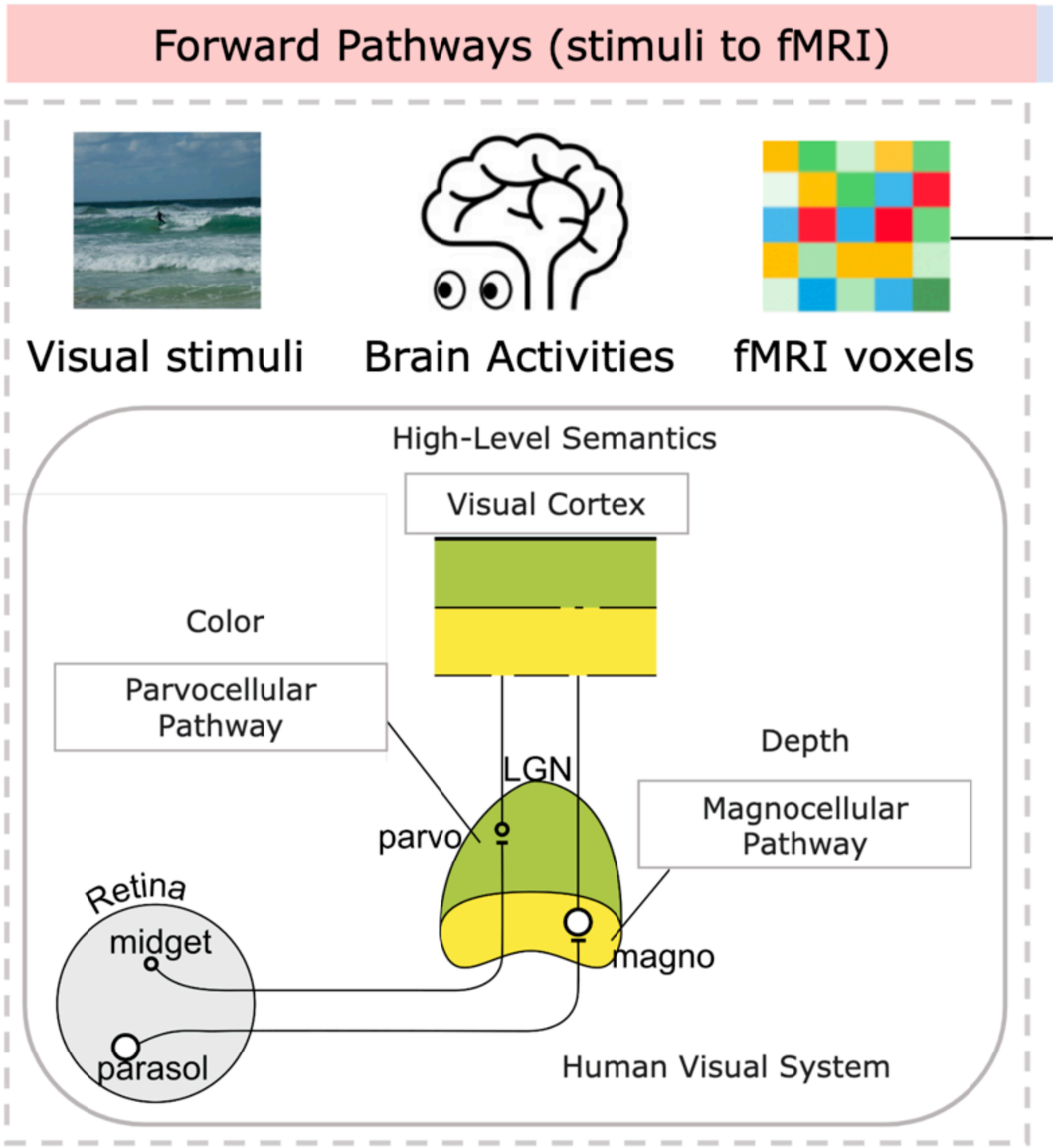
Method

Method Overview

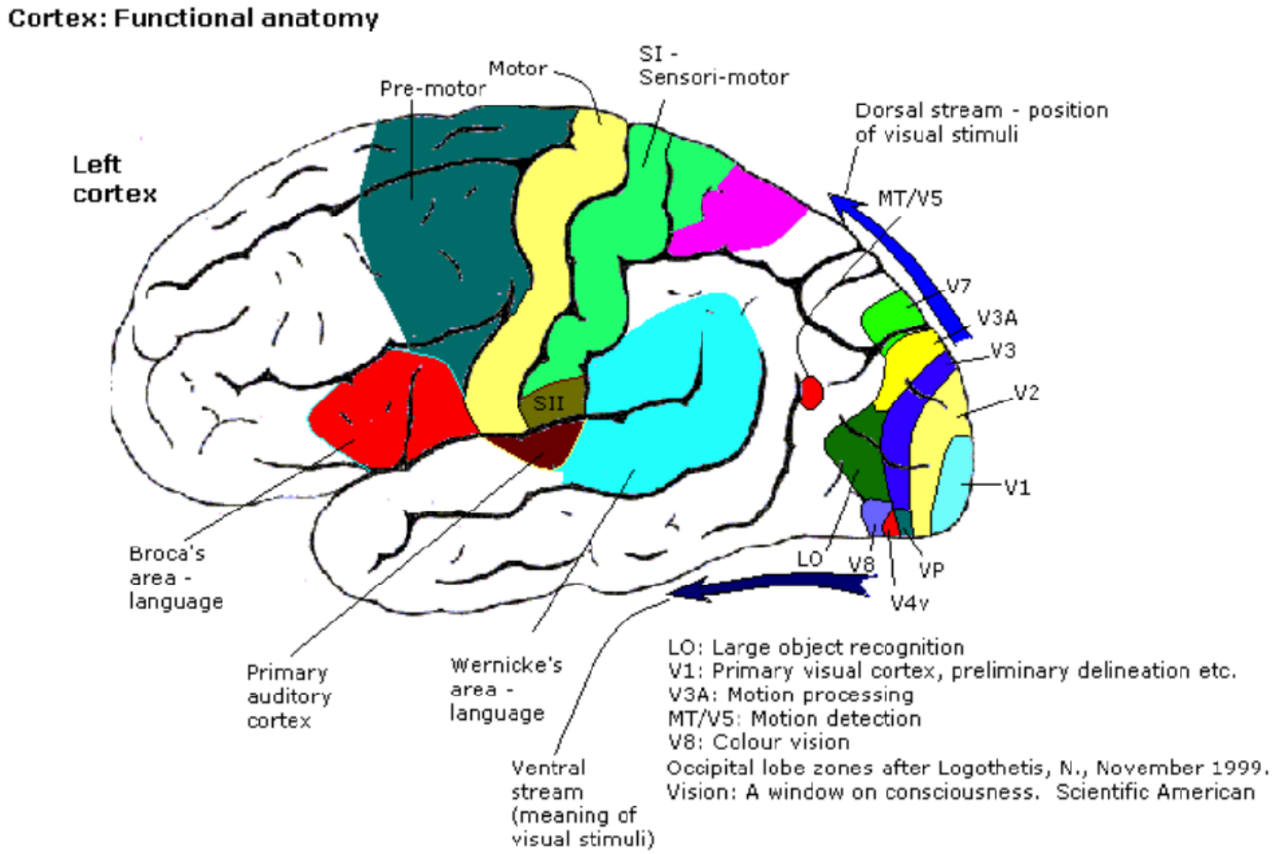


- Connections between retina and brain can be broken into two pathways
- Midget cells / Parvo responsible for color info
- Parasol cells / magno responsible for motion and depth info

Method Overview



- Visual cortex ROIs: V1, V2, V3, hV4, VO, PHC, MT,MST, LO, IPS as seen below



Method Overview

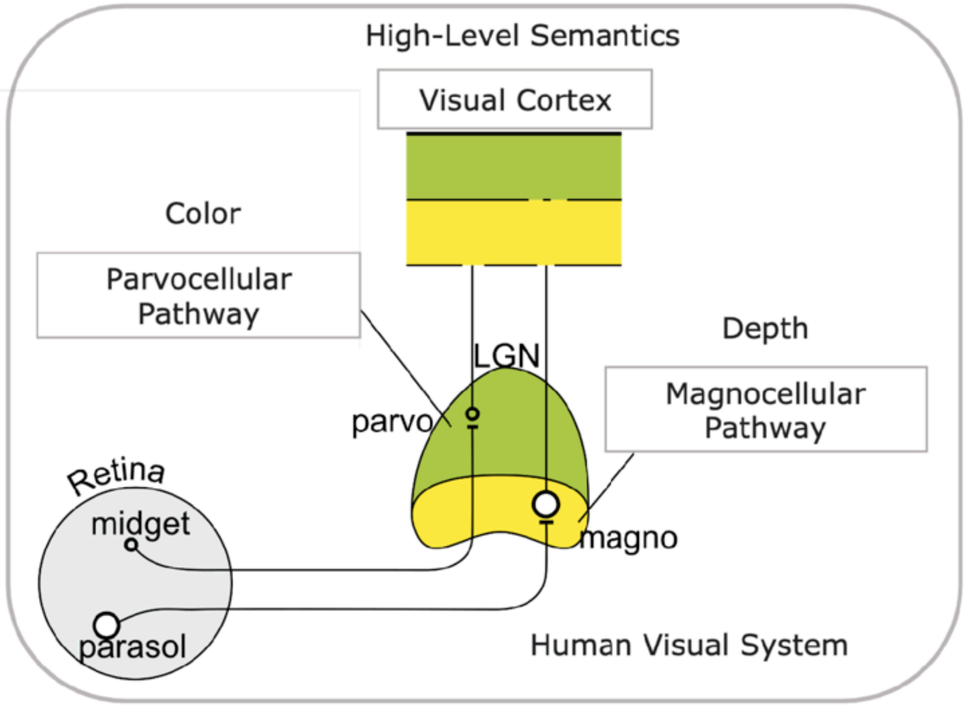
Forward Pathways (stimuli to fMRI)

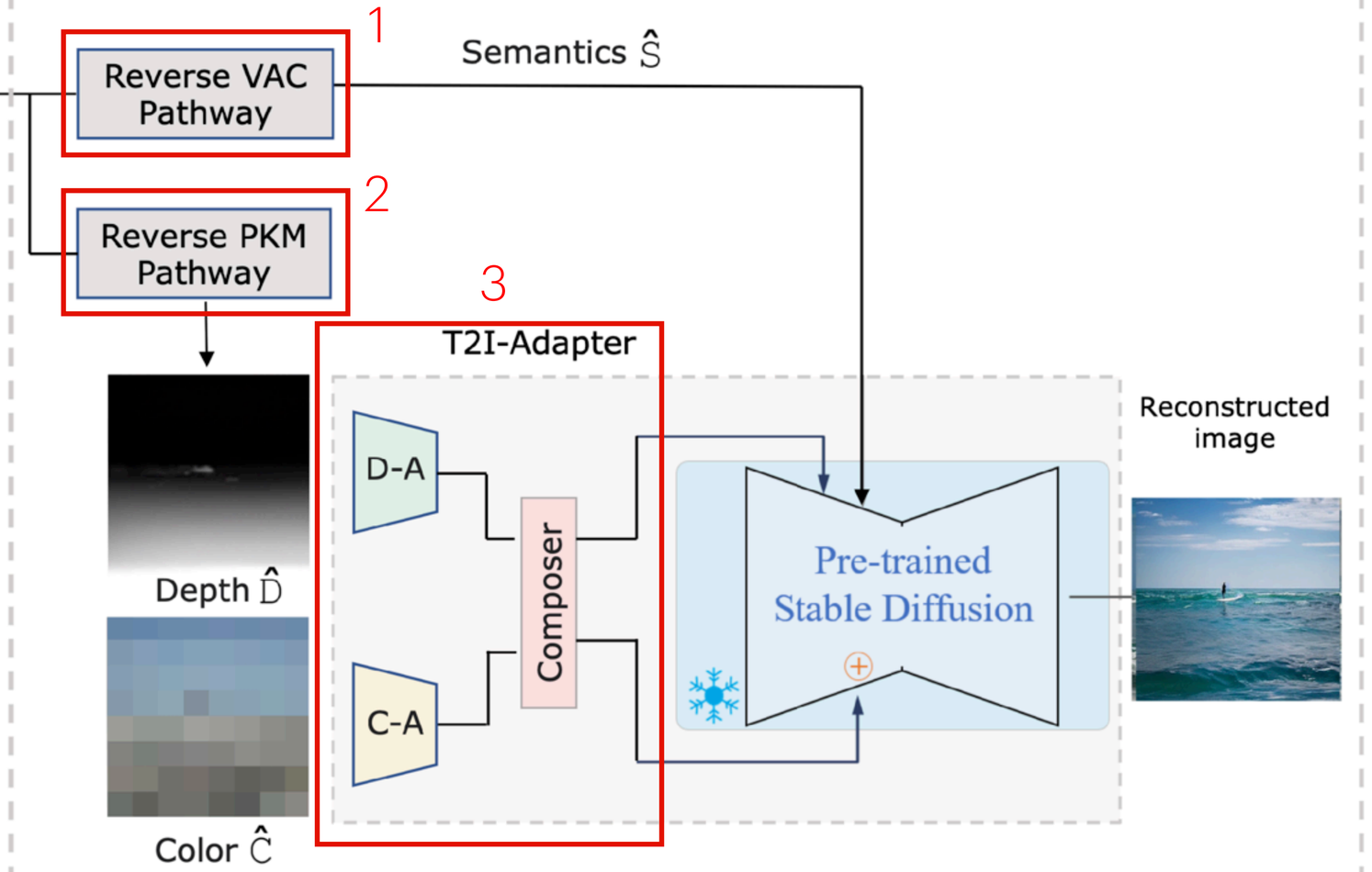
Reverse Pathways (fMRI to Semantics, Color and Depth to Image)

Visual stimuli

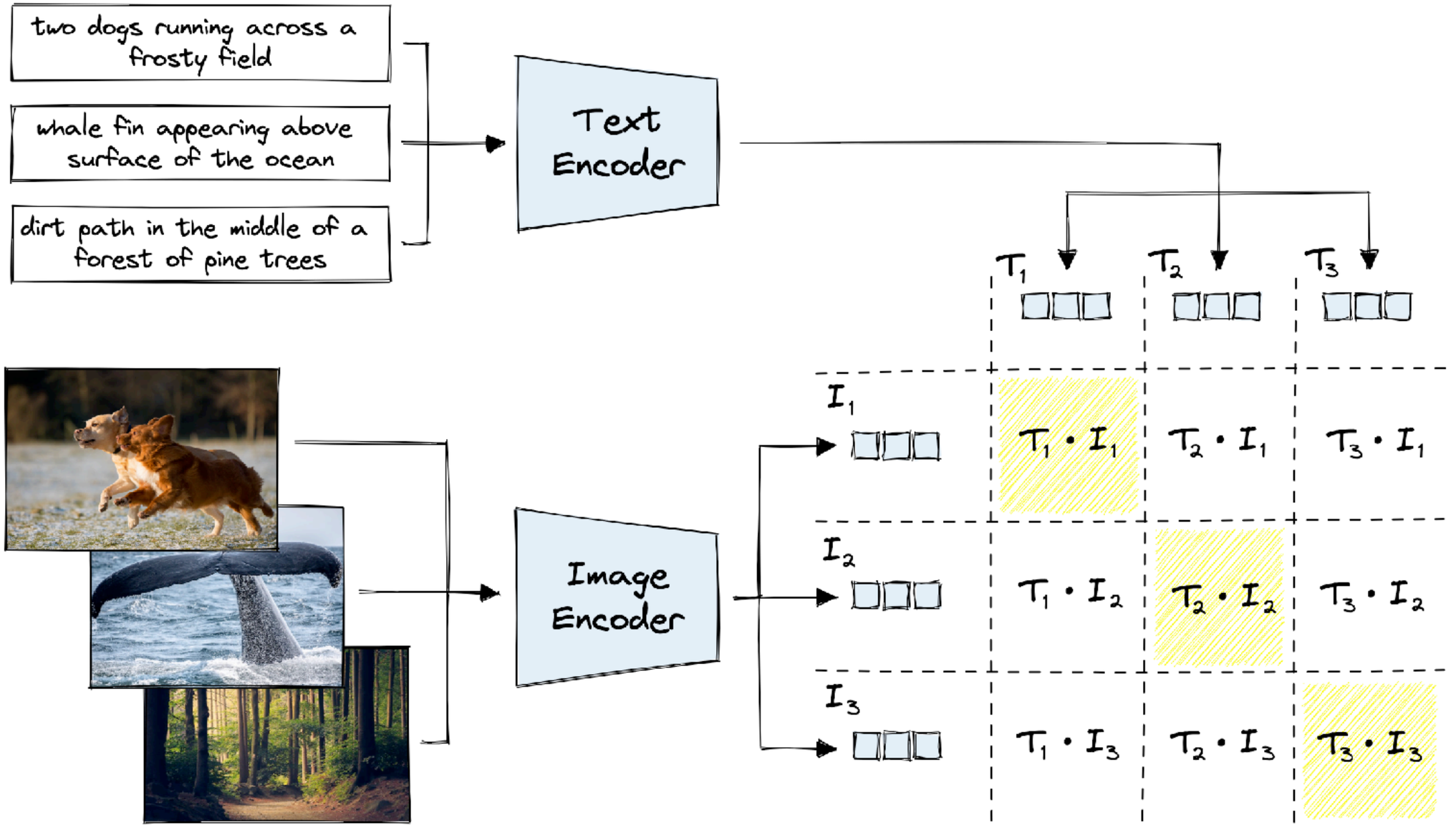
Brain Activities

fMRI voxels

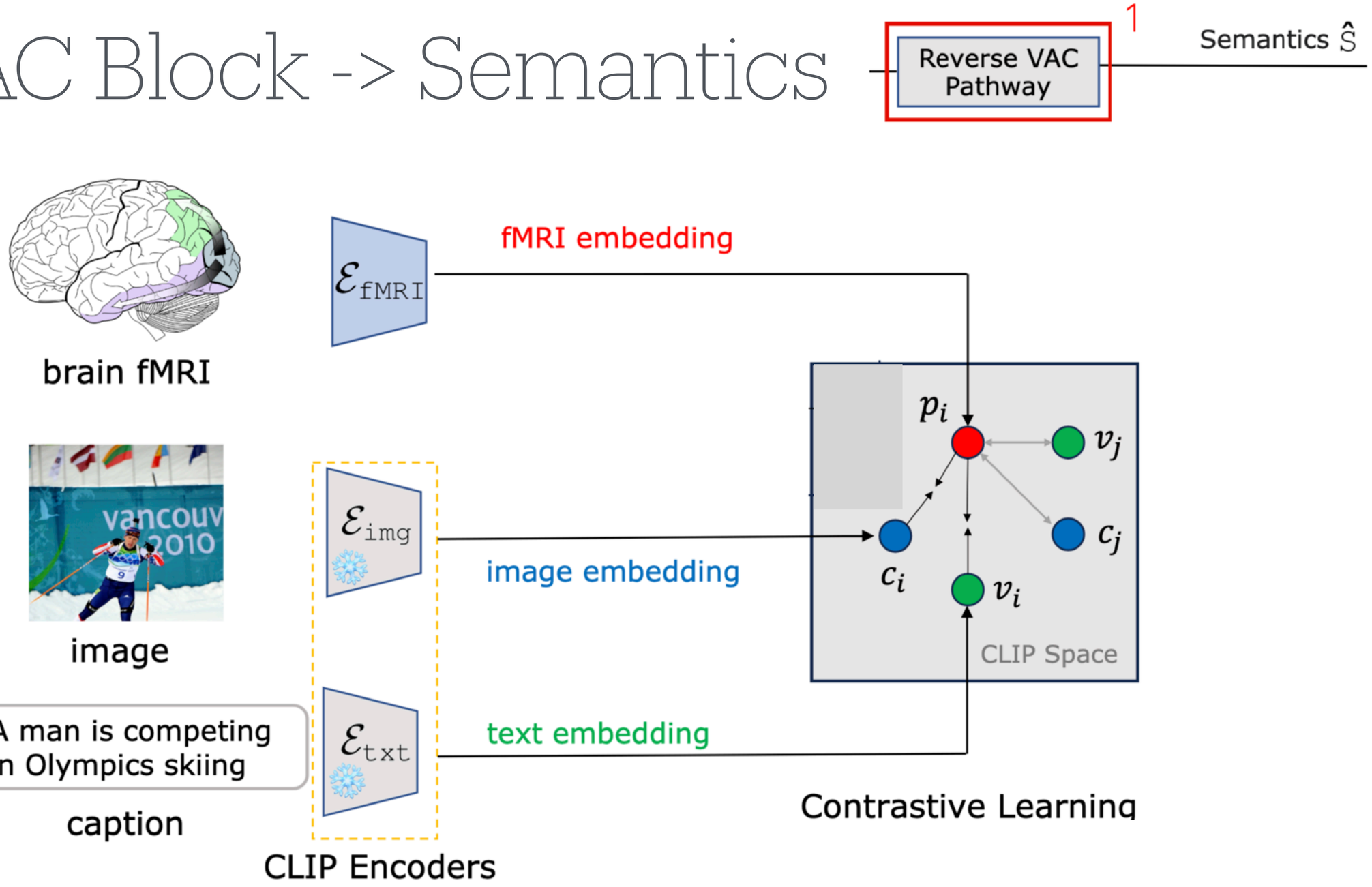




1. RVAC - Background on CLIP



1. RVAC Block -> Semantics



Method Overview

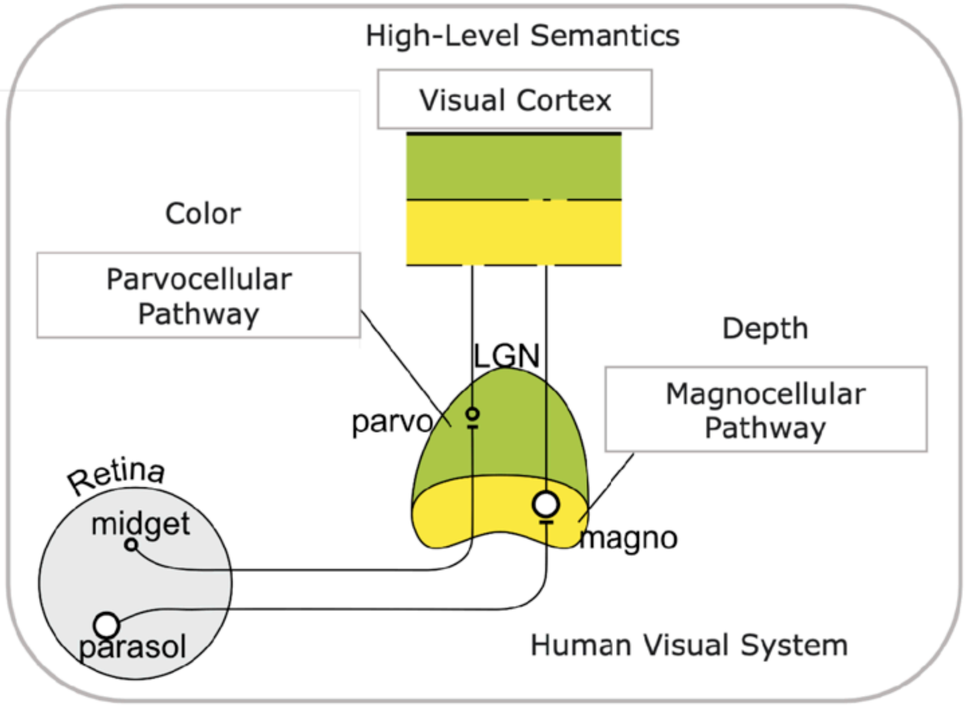
Forward Pathways (stimuli to fMRI)

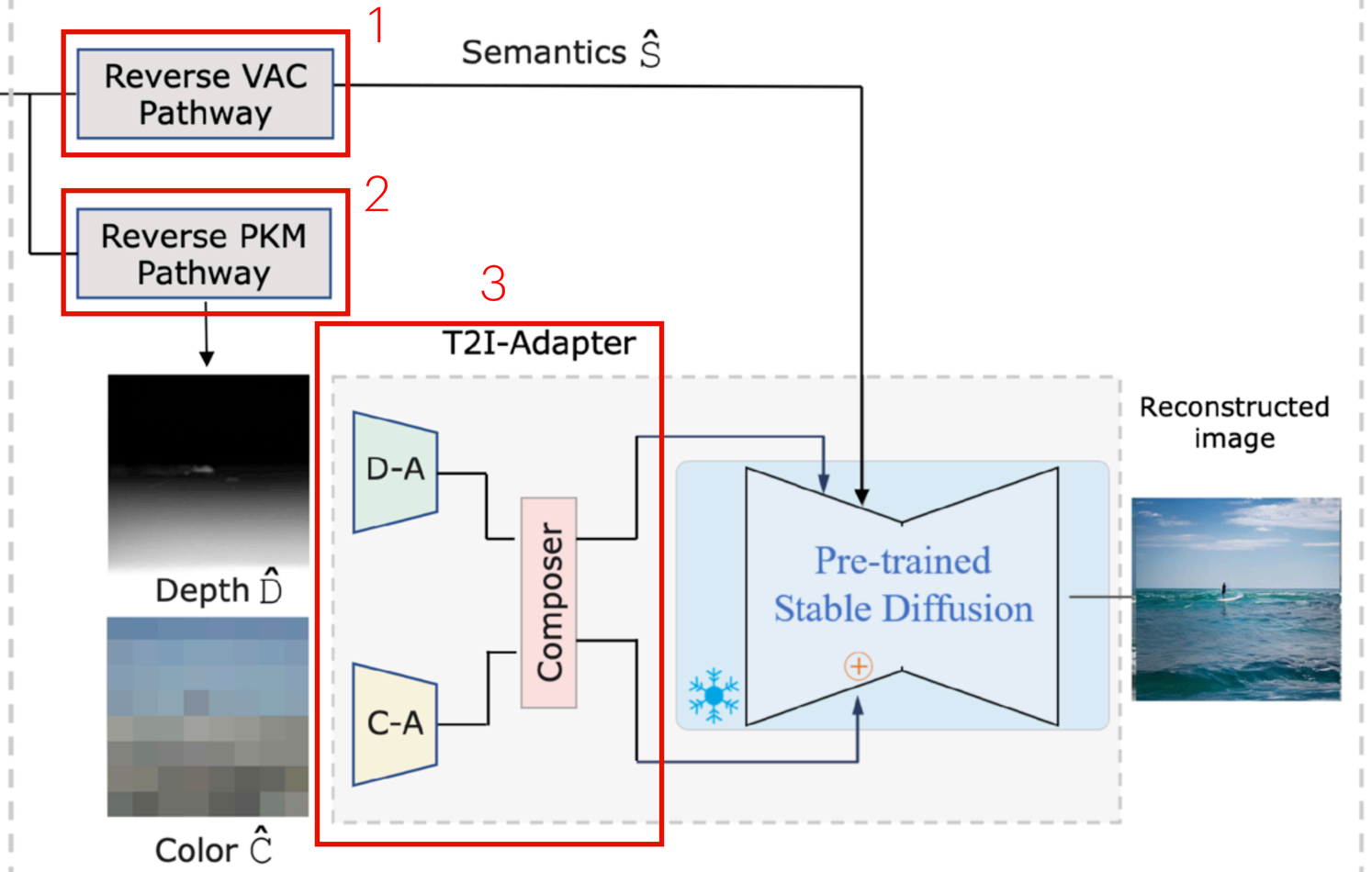
Reverse Pathways (fMRI to Semantics, Color and Depth to Image)

Visual stimuli

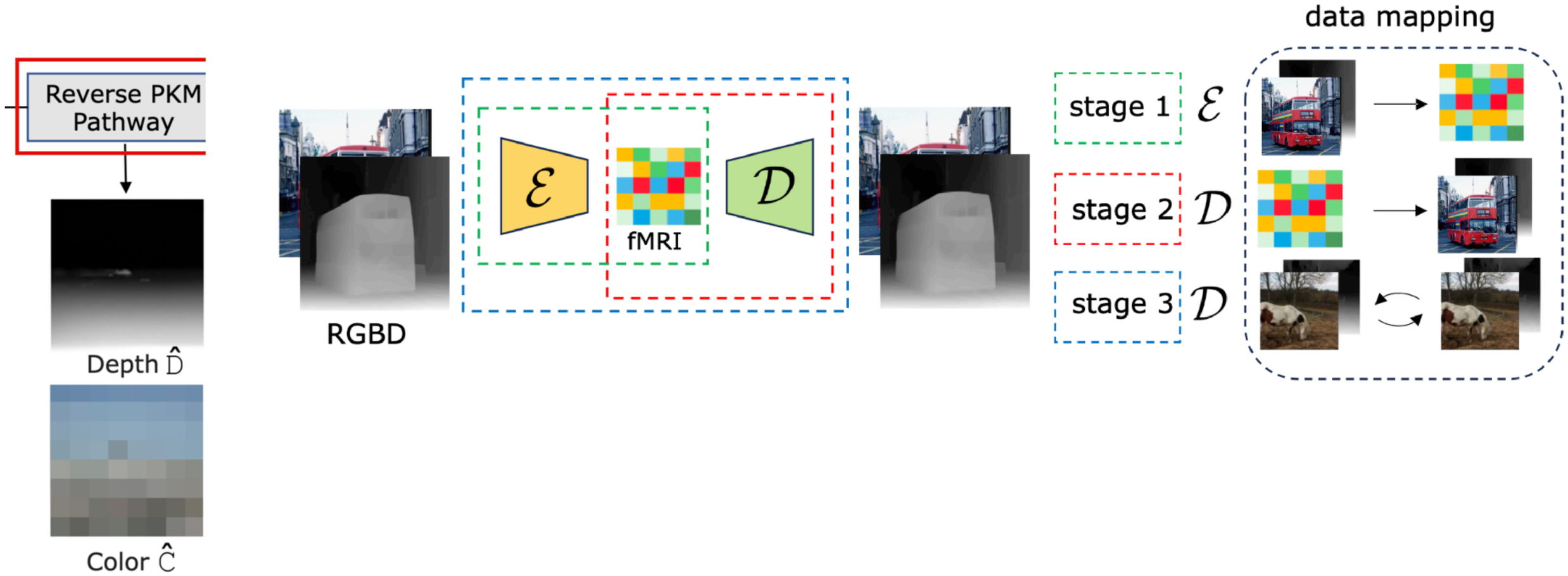
Brain Activities

fMRI voxels

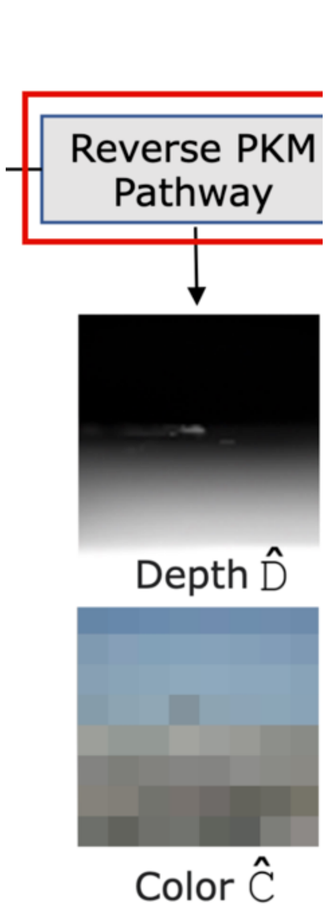


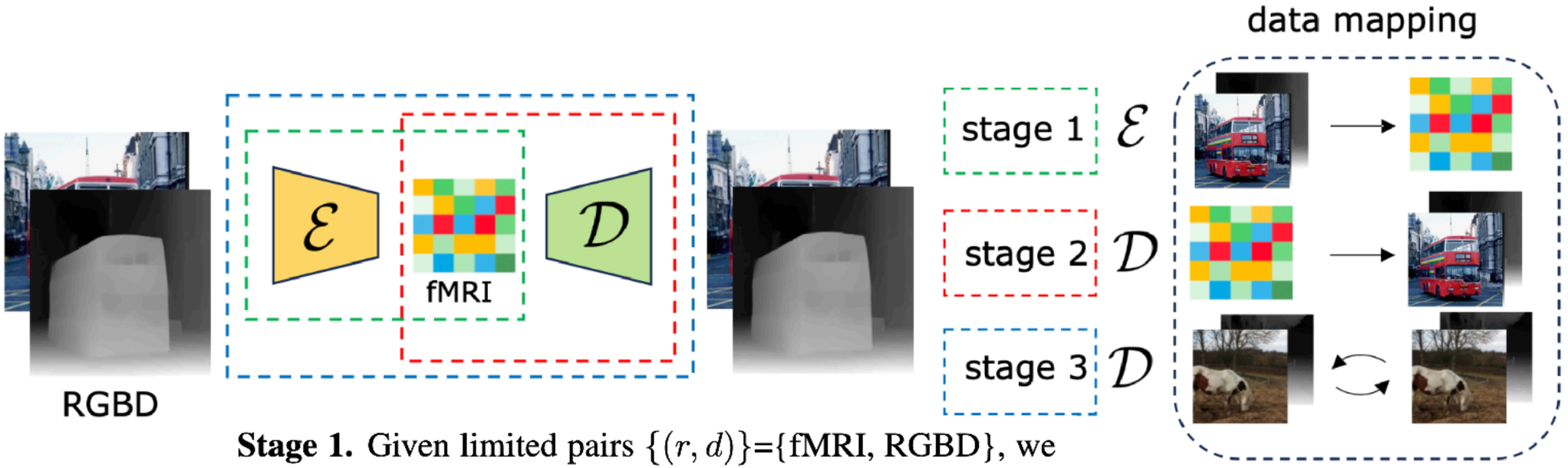


2. RPKM Block - Depth & Color



RPKM Block - Depth & Color



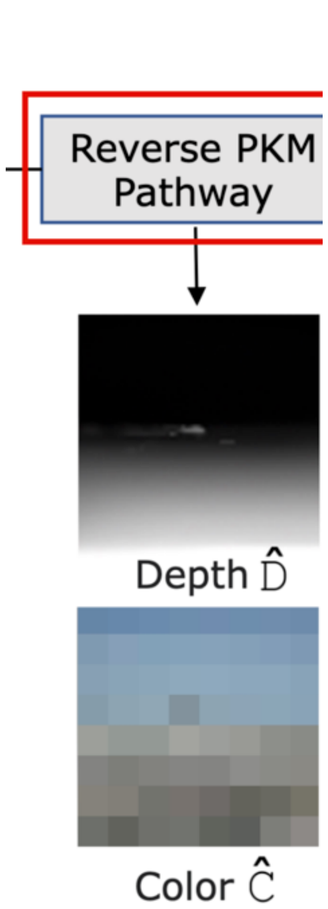


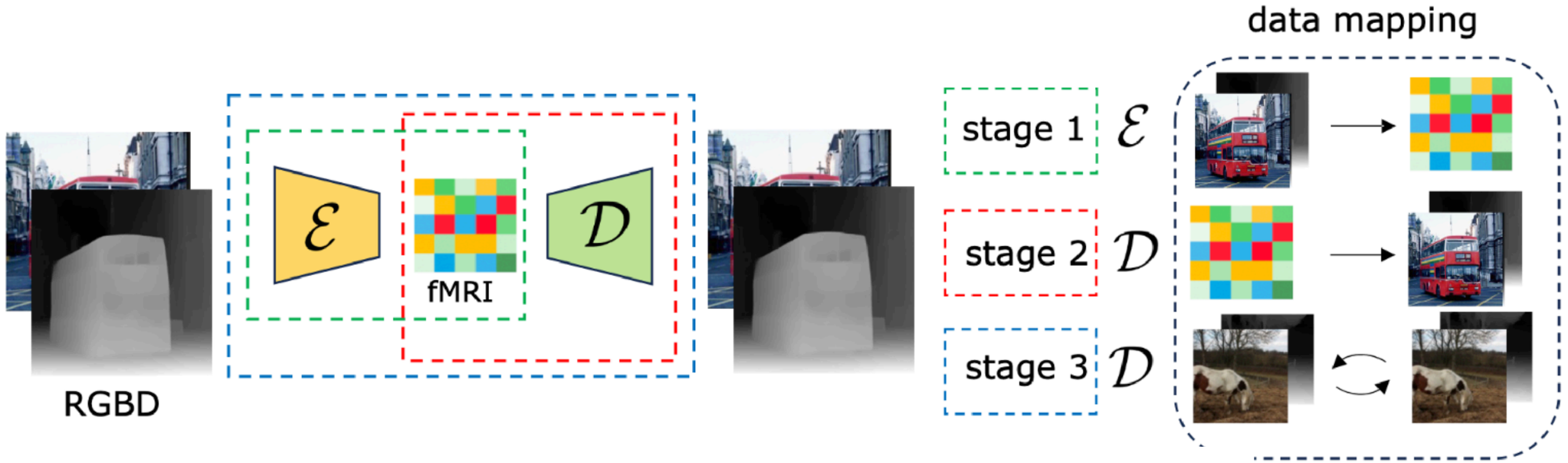
Stage 1. Given limited pairs $\{(r, d)\}$={fMRI, RGBD}, we first train an encoder to map RGBD to their corresponding fMRI data. To compensate for the absence of depth in fMRI datasets, we use MiDaS-estimated depth maps [34] as surrogate ground-truth depth. The encoder is trained with a convex combination of mean square error and cosine proximity between the input r and its predicted counterpart $\hat{r}$:

$$\mathcal{L}_r(r, \hat{r}) = \beta \cdot \mathrm{MSE}(r, \hat{r}) - (1 - \beta) \cos(\angle(r, \hat{r})), \quad (5)$$

where β is determined empirically as a hyperparameter.

RPKM Block - Depth & Color



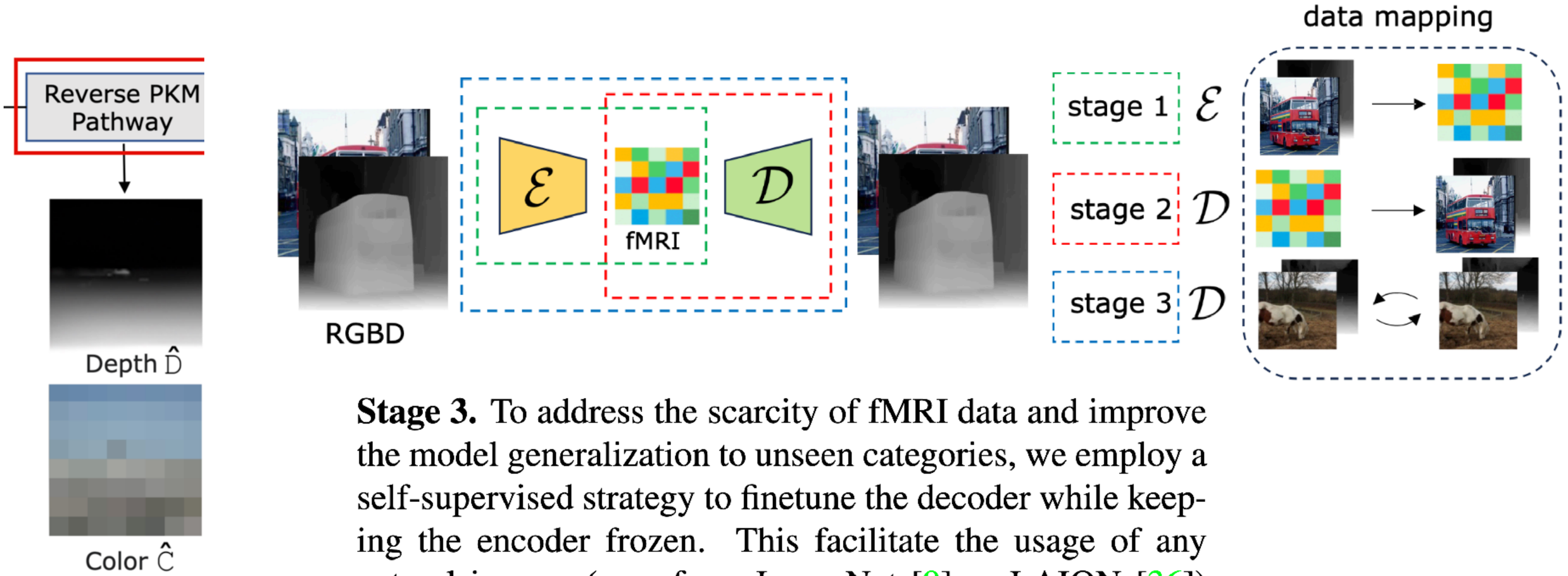


Stage 2. Similar to stage 1, we now train the decoder with pairs $\{(r, d)\}$ in a supervised manner:

$$\mathcal{L}_s(d, \hat{d}) = \|d - \hat{d}\|_1 + \mathcal{J}(\hat{d}), \tag{6}$$

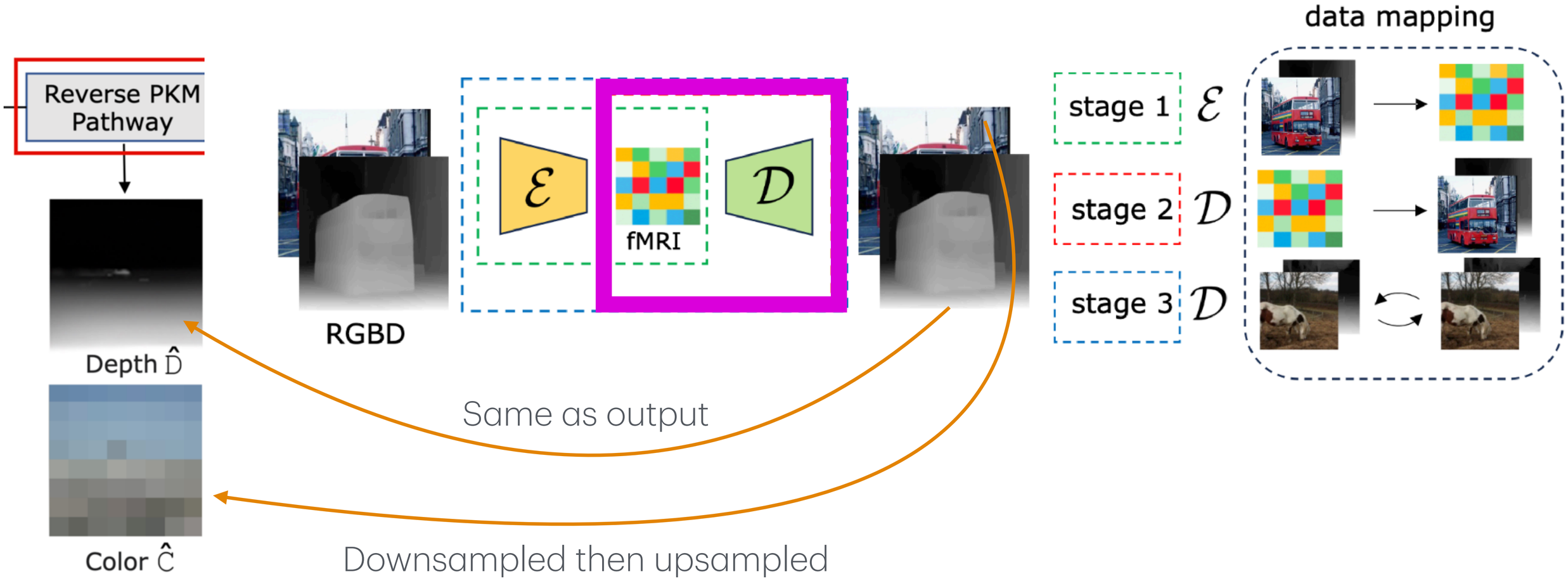
where $\hat{d} = \mathcal{D}(r)$ and the total variation regularization $\mathcal{J}(\hat{d})$ encourages spatial smoothness in the reconstructed $\hat{d}$.

RPKM Block - Depth & Color



Stage 3. To address the scarcity of fMRI data and improve the model generalization to unseen categories, we employ a self-supervised strategy to finetune the decoder while keeping the encoder frozen. This facilitate the usage of any natural images (*e.g.*, from ImageNet [9] or LAION [36]) along with their estimated depth maps

RPKM Block - Inference



Method Overview

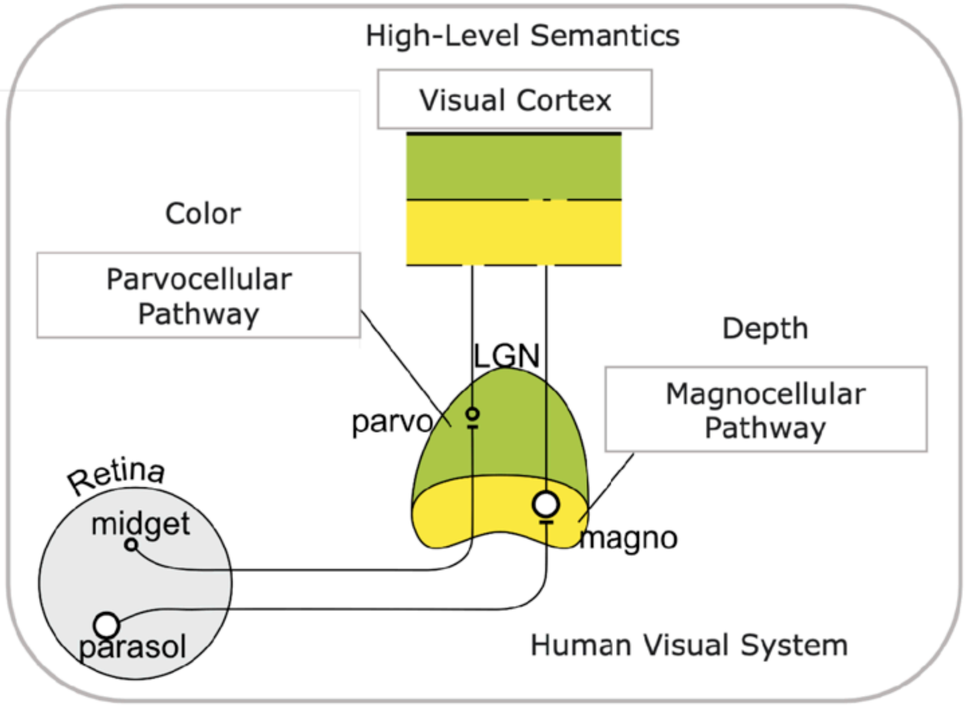
Forward Pathways (stimuli to fMRI)

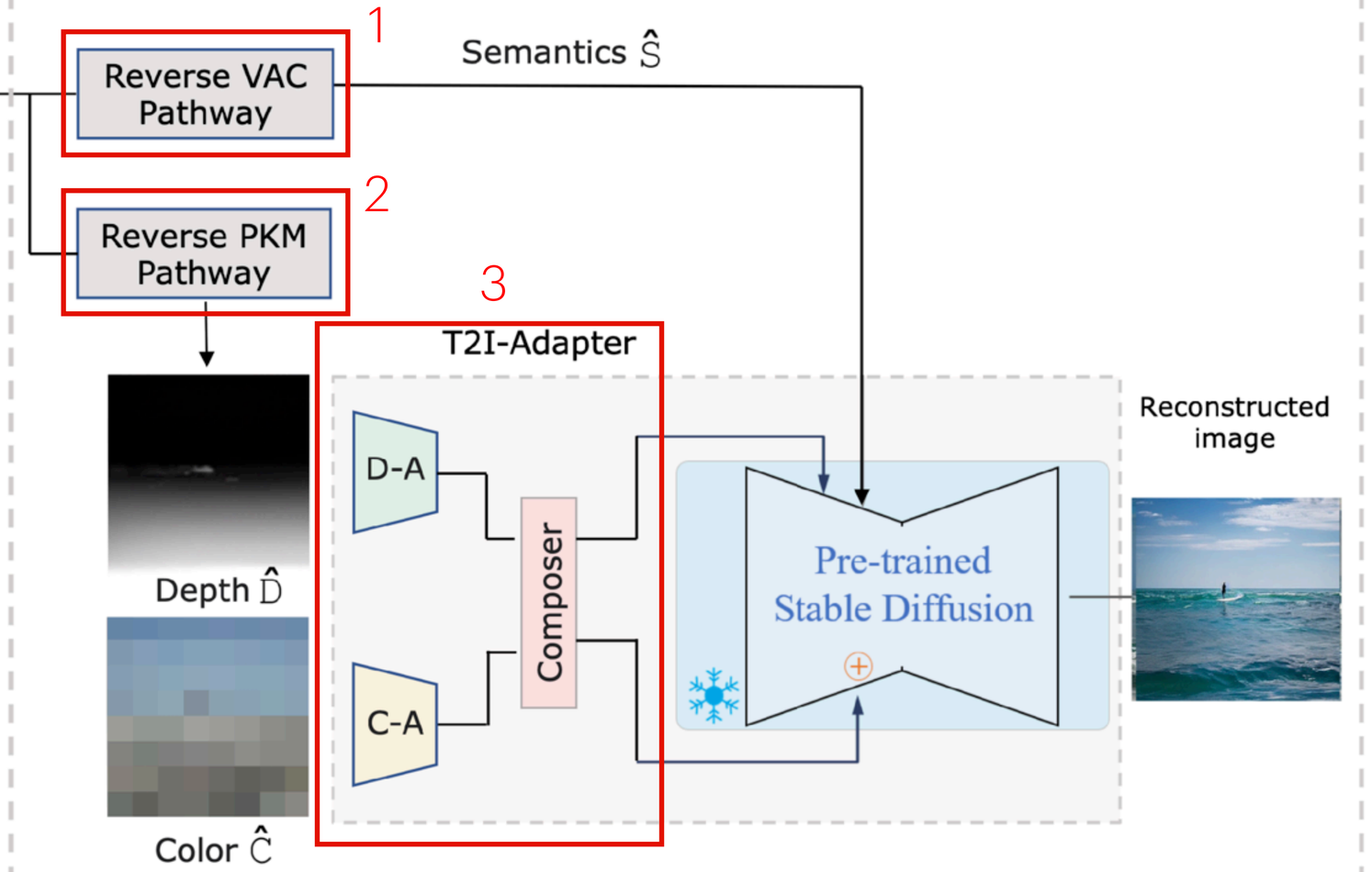
Reverse Pathways (fMRI to Semantics, Color and Depth to Image)

Visual stimuli

Brain Activities

fMRI voxels





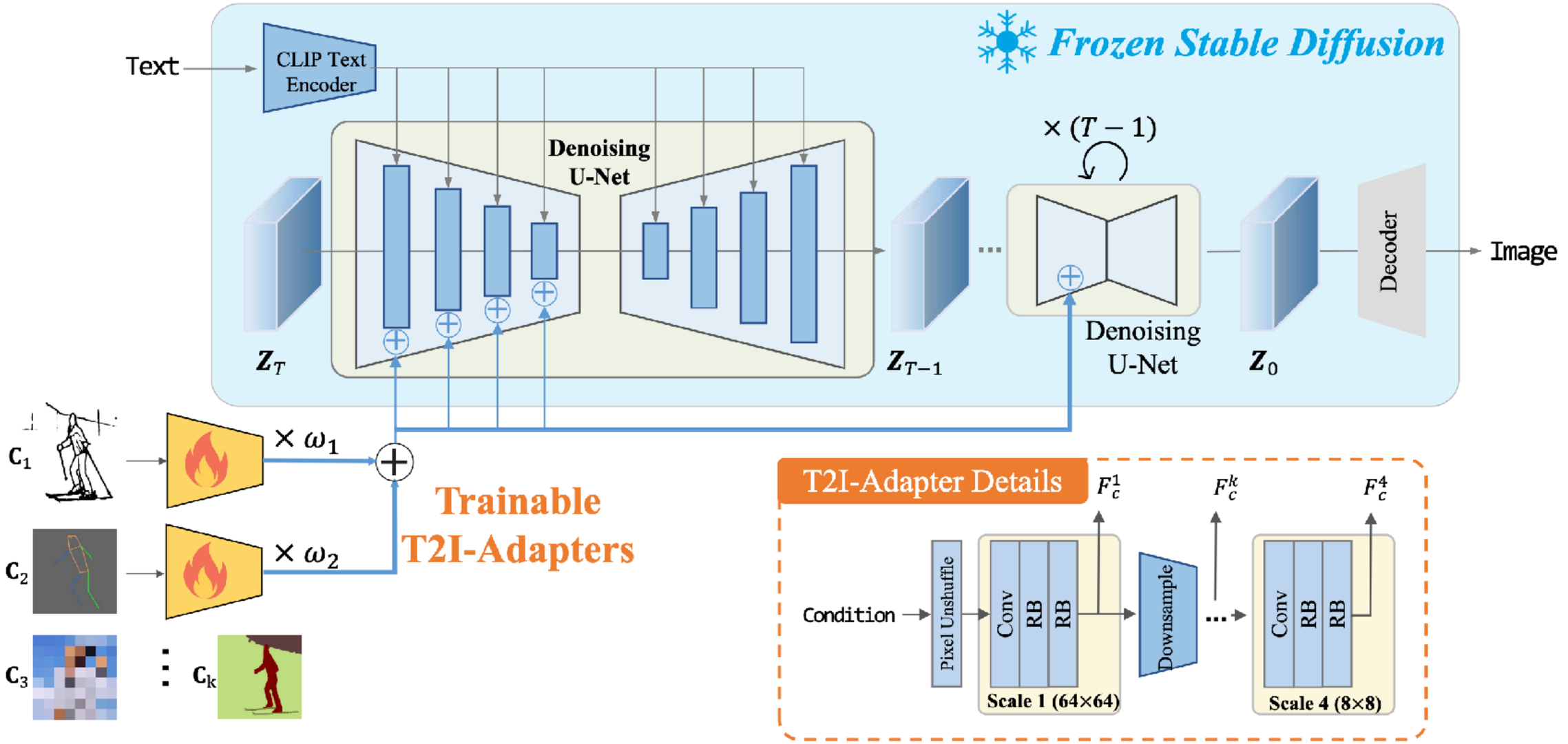
3. T2I Adapter

T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models

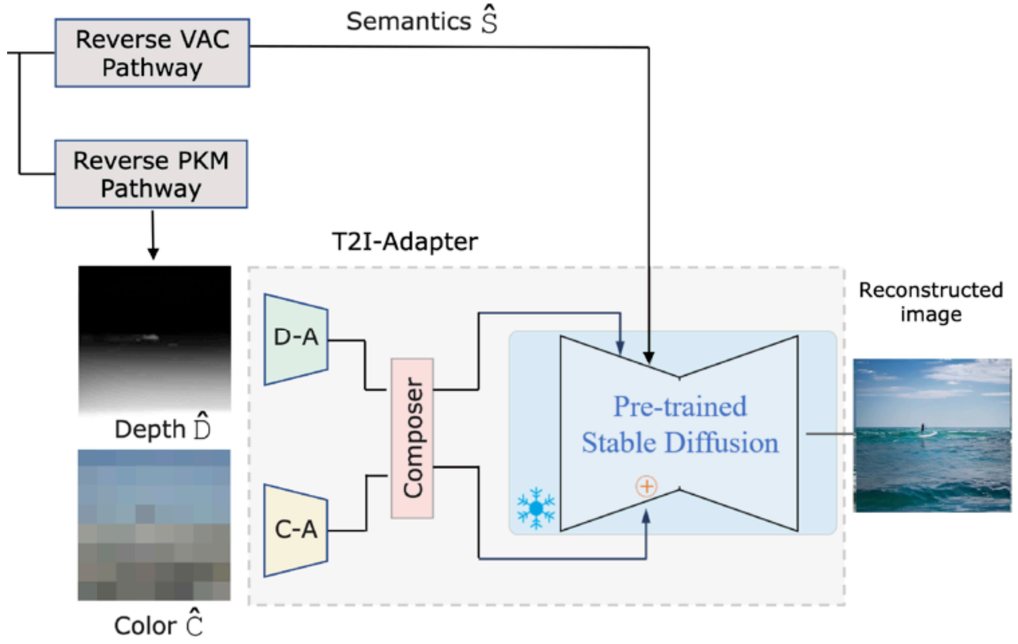
Chong Mou[*1,2] Xintao Wang[†2] Liangbin Xie[*2,3,4] Yanze Wu[2] Jian Zhang[†1]
Zhongang Qi[2] Ying Shan[2] Xiaohu Qie[2]
[1]Peking University Shenzhen Graduate School [2]ARC Lab, Tencent PCG [3]University of Macau [4]Shenzhen Institute of Advanced Technology

https://github.com/TencentARC/T2I-Adapter

T2I Adapter


Frozen Stable Diffusion
Text
CLIP Text Encoder
Denoising U-Net
$\times (T-1)$
Z_T
Z_{T-1}
Denoising U-Net
Z_0
Decoder
Image
C_1
C_2
C_3
C_k
$\times \omega_1$
$\times \omega_2$
Trainable T2I-Adapters
T2I-Adapter Details
F_c^1
F_c^k
F_c^4
Condition
Pixel Unshuffle
Conv
RB
RB
Downsample
Scale 1 (64×64)
Scale 4 (8×8)

Diffusion



We utilize Stable Diffusion (SD) [35] to reconstruct the final image from the predicted CLIP embedding $\hat{\mathrm{S}}$ and the additional guidance from predicted color palette $\hat{\mathrm{C}}$ and depth map $\hat{\mathrm{D}}$. Such guidance is produced using the color adapter $\mathcal{R}_c$ and the depth adapter $\mathcal{R}_d$ within T2I-adapter [29]. This process is formulated as follows:

$$\begin{aligned} \mathrm{F}_{\mathcal{R}} &= \omega_c \mathcal{R}_c\left(\hat{\mathrm{C}}\right) + \omega_d \mathcal{R}_d\left(\hat{\mathrm{D}}\right), \\ \hat{I} &= \mathrm{SD}\left(z, \mathrm{F}_{\mathcal{R}}, \hat{\mathrm{S}}\right), \end{aligned} \tag{7}$$

where z is a random noise, ω_c and ω_d are adjustable weights to control the relative significance of the adapters.

Results

Results

Figure 6. **Sample Visual Decoding Results from the SOTA Methods on NSD.**

Ablation Studies

My two cents

Controversy?

5.1. Experimental Setting

Dataset. We use the Natural Scenes Dataset (NSD) [2] in all experiments, which follows the standard practices in the field [15, 24, 27, 32, 37, 41]. NSD, as the largest fMRI dataset, records brain responses from eight human subjects successively isolated in an MRI machine and passively observed a wide range of visual stimuli, namely, natural images sourced from MS-COCO [25], which allows retrieving the associated captions. In practice, because brain activity patterns highly vary across subjects [18], a separate model is trained per subject. The standardized splits contain 982 fMRI test samples and 24,980 fMRI training samples. Please refer to the supplementary material for more details.

What?!?!

Controversy?

5.1. Experimental Setting

Dataset. We use the Natural Scenes Dataset (NSD) [2] in all experiments, which follows the standard practices in the field [15, 24, 27, 32, 37, 41]. NSD, as the largest fMRI dataset, records brain responses from eight human subjects successively isolated in an MRI machine and passively observed a wide range of visual stimuli, namely, natural images sourced from MS-COCO [25], which allows retrieving the associated captions. In practice, because brain activity patterns highly vary across subjects [18], a separate model is trained per subject. The standardized splits contain 982 fMRI test samples and 24,980 fMRI training samples. Please refer to the supplementary material for more details.

I imagine there is high correlation between these two, no class withholding?

Discussion/Conclusion

- What do BME people think about this work?
- What other criticisms are there with this method?
- Are general population-based methods feasible? (No subject specific models)
- Is this approach a glorified classifier?

Thanks for listening :)