Mind Reader: Reconstructing complex images from brain activities

Anonymous Author(s)
Affiliation
Address
email

Abstract

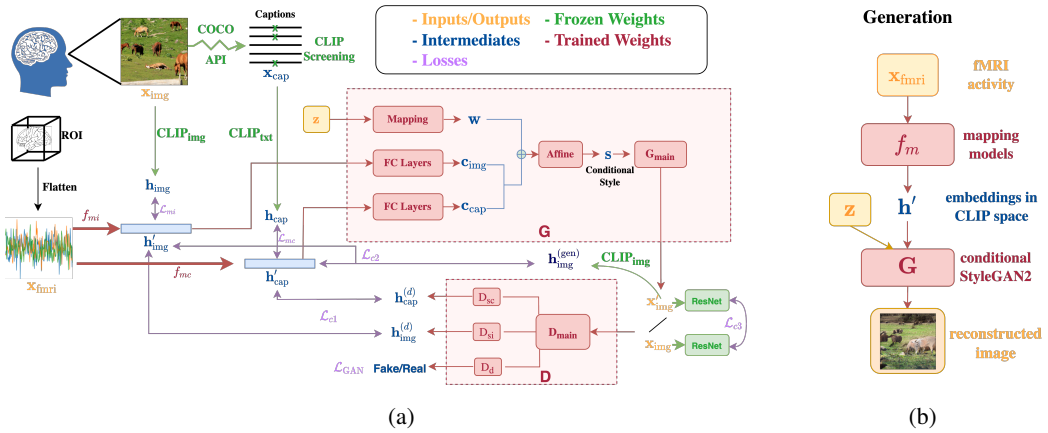
Understanding how the brain encodes external stimuli and how these stimuli can be decoded from the measured brain activities are long-standing and challenging questions in neuroscience. In this paper, we focus on reconstructing the complex image stimuli from fMRI (functional magnetic resonance imaging) signals. Unlike previous works that reconstruct images with single objects or simple shapes, our work aims to reconstruct image stimuli that are rich in semantics, closer to everyday scenes, and can reveal more perspectives. However, data scarcity of fMRI datasets is the main obstacle to applying state-of-the-art deep learning models to this problem. We find that incorporating an additional text modality is beneficial for the reconstruction problem compared to directly translating brain signals to images. Therefore, the modalities involved in our method are: (i) voxel-level fMRI signals, (ii) observed images that trigger the brain signals, and (iii) textual description of the images. To further address data scarcity, we leverage an aligned vision-language latent space pre-trained on massive datasets. Instead of training models from scratch to find a latent space shared by the three modalities, we encode fMRI signals into this pre-aligned latent space. Then, conditioned on embeddings in this space, we reconstruct images with a generative model. The reconstructed images from our pipeline balance both naturalness and fidelity: they are photo-realistic and capture the ground truth image contents well.

1 Introduction

In an effort to understand visual encoding and decoding processes, researchers in recent years have curated multiple datasets recording fMRI signals while the subjects are viewing natural images [3, 7, 32, 39]. In particular, the Natural Scene Dataset (NSD [3]) was built to meet the needs of data-hungry deep learning models, sampling at an unprecedented scale compared to all prior works while having the highest resolution and signal-to-noise ratio (SNR). In addition, all the images used in NSD are sampled from MS-COCO [20], which has far richer contextual information and more detailed annotations compared to datasets that are commonly used in other fMRI studies (e.g., Celeb A face dataset [21], ImageNet [9], self-curated symbols, grayscale datasets). This dataset, therefore, offers the opportunity to explore the decoding of complex images that are closer to real-life scenes.

Human visual decoding can be categorized into stimuli category classification [1], stimuli identification [36], and reconstruction. We focus on stimuli reconstruction in this study. Different from previous efforts in reconstructing images from fMRI [5, 10, 11, 22, 30, 32, 33], we approach the problem with one more modality, that of text. The benefits of adding the text modality are threefold: first, the brain is naturally multimodal. Research [6, 12, 23] indicates that the brain is not only capable of learning multisensory representations, but a larger portion of the cortex is engaged in multisensory processing: for example, both visual and tactile recognition of objects activate the same part of the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) (b)

Figure 1: The pipeline for reconstructing seen images from fMRI signals. 1a details different components, from collected data to the reconstructed image. The pipeline is trained in two stages: during the first stage, mapping models f_{mi}, f_{mc} are trained to encode fMRI activities into the CLIP embedding space. In the second stage, conditional generator $\mathbf{G}$ and contrastive discriminator $\mathbf{D}$ are finetuned while both f_{mi}, f_{mc} are kept frozen. 1b shows the image generation process once models are trained.

object-responsive cortex [24]. Visual-linguistic pathways along the border of the occipital lobe [25] also bring a more intertwined view of the brain's representation of these two modalities. Second, multimodal deep models tend to explain the brain better (having higher representation correlations) than the visual-only models, even when compared with activities in the visual cortex [8]. Lastly, our goal is to reconstruct complex images that have multiple objects in different categories with intricate interactions: it is natural to incorporate contextual information as an additional modality.

Instead of training a model to map all three modalities (fMRI, image, text) to a unified latent space, we propose to map fMRI to a well-aligned space shared by image and text, and use conditional generative models to reconstruct seen images from representations in that space. This design addresses the data scarcity issue of brain datasets by separating fMRI from the other two modalities. In this way, a large amount of data is readily available to learn the shared visual-language representation and to train a generative model conditioned on this representation. Furthermore, pre-trained models can be utilized to make the whole reconstruction pipeline more efficient and flexible.

Our contributions are as follows: (1) to the best of our knowledge, this is the first work on reconstructing complex images from human brain signals. It provides an opportunity to study the brain's visual decoding in a more natural setting than using object-centered images. Compared to previous works, it also decodes signals from more voxels and regions, including those outside the visual cortex, that are responsive to the experiment. This inclusion allows us to study the behavior and functionality of more brain areas. (2) We address the data scarcity issue by incorporating additional text modality and leveraging pre-trained latent space and models. For the reconstruction, we focus on semantic representations of the images while taking low-level visual features into account. (3) Our results show we can decode complex images from fMRI signals relatively faithfully. We also perform microstimulation on different brain regions to study their properties and showcase the potential usages of the pipeline.

2 Method

Recent developments in contrastive models allow more accurate embeddings of images and their semantic meanings in the same latent space. This performance is realized using massive datasets: models such as CLIP [26] and ALIGN [16] utilize thousands of millions of image-text pairs for representation alignment. In comparison, brain imaging datasets that record pairs of images and fMRI range from 1.2k to 73k samples, making it difficult to learn brain encoding and decoding models from scratch. However, we can utilize aligned embeddings obtained from pre-trained contrastive models as the intermediary and generate images conditioned on these embeddings. In our pipeline, as shown in fig. 1, we first map fMRI signals to CLIP embeddings of the observed image and its captions, then pass these embeddings to a conditional generative model for image reconstruction.

Captions	CLIP probabilities	
(1) A group of people sitting and standing on top of a sandy beach.	0.0376	
(2) A surfboard rests on the beach while people play in the waves.	0.519	✔
(3) A surfboard on the sand and people on the beach behind.	0.4302	✔
(4) A few people are hanging out and appreciating their time.	0.001286	
(5) A group of people are sitting on the beach shore.	0.0122	

Figure 2: Image caption screening through CLIP encoders. For this sample, threshold is put at half of the largest probability: 0.5×0.519. Therefore, captions (2) and (3) of the image are kept.

2.1 Caption Screening

Each image $\mathbf{x}_{\text{img}}$ in the COCO dataset has five captions $\{\mathbf{x}_{\text{cap}_1}, \cdots, \mathbf{x}_{\text{cap}_5}\}$ collected through Amazon's Mechanical Turk (AMT), and in nature, these captions vary in their descriptive ability. Fig. 2 shows a sample image with its five captions, and we can tell captions (2) and (3) are more objective and informative than caption (4) when it comes to describing the *content* of that image, thus are more helpful to serve as the image generation condition. We utilize pre-trained CLIP encoders to screen the high-quality captions since representations in the CLIP space are trained to be image-text aligned. A caption with an embedding more aligned to the image embedding is more descriptive than a less aligned one; it is also less general and more specific to this particular image because of the contrastive loss in CLIP. For the screening, we pass each image together with its five captions to the CLIP model, which outputs corresponding probabilities that the captions and image are proper pairs. We keep captions with probabilities larger than half of the highest probability. After screening out less informative captions, we have one to three high-quality captions per image.

2.2 Mapping fMRI signals to CLIP space

Each fMRI signal that reflects a specific image is a 3D data volume, and the value on position (i, j, k) is the relative brain activation on this voxel triggered by the image. We apply an ROI (region of interest) mask on this 3D volume to extract signals of cortical voxels that are task-related and have good SNRs. The signal is then flattened into a 1D vector and voxel-wise standardized within each scan session. The end results $\mathbf{x}_{\text{fmri}}$ are used by our image reconstruction pipeline. We choose to use the ROI with the widest region coverage, and the length N of $\mathbf{x}_{\text{fmri}}$ ranges from 12682 to 17907 for different brains in the NSD dataset.

Our goal is to train two mapping models, f_{mi} and f_{mc} in fig. 1 (collectively denoted as f_m), that encodes $\mathbf{x}_{\text{fmri}} \in \mathbb{R}^N$ to $\mathbf{h}_{\text{img}} = \mathrm{C}_{\text{img}}(\mathbf{x}_{\text{img}}) \in \mathbb{R}^{512}$ and $\mathbf{h}_{\text{cap}} = \mathrm{C}_{\text{txt}}(\mathbf{x}_{\text{cap}}) \in \mathbb{R}^{512}$ respectively. Here C_{img}, C_{txt} are CLIP image and text encoders, and $\mathbf{x}_{\text{cap}}$ is one of the image captions chosen randomly from the vetted caption pool. We construct both f_m as a CNN with one Conv1D layer followed by four residual blocks and three linear layers. The training objective is a combination of MSE loss, cosine similarity loss, and contrastive loss on cosine similarity. We use the infoNCE definition [37] of contrastive loss, for the ith sample in a batch of size B:

$$\text{Contra}(a^{(i)}, b^{(i)}) = -\mathbb{E}_i \left[\log \frac{\exp(\cos(a^{(i)}, b^{(i)})/\tau)}{\sum_{j=1}^{B} \exp(\cos(a^{(i)}, b^{(j)})/\tau)} \right] \tag{1}$$

For the mapping model f_{mi} that encodes fMRI to image embeddings, we have $\mathbf{h}_{\text{img}}^{(i)\,\prime} = f_{mi}(\mathbf{x}_{\text{fmri}}^{(i)})$. The training objective is:

$$\mathcal{L}_{mi} = \mathbb{E}_i \left[\alpha_1 ||\mathbf{h}_{\text{img}}^{(i)\,\prime} - \mathbf{h}_{\text{img}}^{(i)}||_2^2 + \alpha_2 (1 - \cos(\mathbf{h}_{\text{img}}^{(i)\,\prime}, \mathbf{h}_{\text{img}}^{(i)})) \right] + \alpha_3 \, \text{Contra}(\mathbf{h}_{\text{img}}^{(i)\,\prime}, \mathbf{h}_{\text{img}}^{(i)}), \tag{2}$$

where $\tau, \alpha_1, \alpha_2, \alpha_3$ are non-negative hyperparameters selected through sweeps. The loss $\mathcal{L}_{mc}$ for caption embedding mapping model f_{mc} is defined similarly. Although CLIP embeddings are trained to be aligned, there are still systematic differences between image and text embeddings, with embeddings under each modality showing outlier values at a few fixed positions. In addition, we also notice the generated images emphasize either image content (object proximity, shape, etc.) or semantic features depending on which condition we use. Therefore, including both embeddings as the conditions for a generator can cover both ends, and that is why we train two mapping models for the two modalities. Since the outlier indices are fixed for each modality across images, clipping the value should not affect image-specific information. Therefore, before normalizing the ground

truth embeddings into unit vectors, we set $\mathbf{h} = \mathrm{clamp}(\mathbf{h}, -1.5, 1.5)$. This can greatly improve the mapping performance during training.

2.3 Image reconstruction with CLIP embedding conditioning

The mapping models output fMRI-mapped CLIP embeddings $\mathbf{h}'_{\text{img}}$ and $\mathbf{h}'_{\text{cap}}$ that serve as conditions for the generative model. We aim to generate images that have both naturalness (being photo-realistic) and high fidelity (can faithfully reflect objects and relationships in the observed image). Our generation model is built upon Lafite [40], a text-to-image generation model: it adapts unconditional StyleGAN2 [19, 18] to conditional image generation contexted on CLIP text embeddings.

In our generator $\mathbf{G}$, both conditions $\mathbf{h}'_{\text{img}}$ and $\mathbf{h}'_{\text{cap}}$ are injected into the StyleSpace: each of them goes through two fully connected (FC) layers and is transformed into condition codes $\mathbf{c}_{\text{img}}$ and $\mathbf{c}_{\text{cap}}$. These condition codes are max-pooled and then concatenated with the intermediate latent code $\mathbf{w} \in \mathcal{W}$, which is obtained from passing the noise vector $\mathbf{z} \in \mathcal{Z}$ through a mapping network (see fig. 1). Using a mapping network to transform $\mathbf{z}$ into an intermediate latent space $\mathcal{W}$ is the key of StyleGAN as $\mathcal{W}$ is shown to be much less entangled than $\mathcal{Z}$[34]. The conditioned style $\mathbf{s}$ is then passed to different layers of $\mathbf{G}$ as in StyleGAN2, generating image $\mathbf{x}_{\text{img}}{}'$:

$$\mathbf{s} = \mathbf{w}||\max(\mathbf{c}_{\text{img}}, \mathbf{c}_{\text{cap}}), \quad \mathbf{x}_{\text{img}}{}' = \mathbf{G}(\mathbf{s}). \tag{3}$$

We align the semantics of generated $\mathbf{x}_{\text{img}}{}'$ and condition vectors by passing $\mathbf{x}_{\text{img}}{}'$ through pre-trained CLIP encoders and apply contrastive loss (eq. (5) $\mathcal{L}_{c2}$) between them. For further alignment of the lower-level visual features, such as prominent edges, corners and shapes, we also pass the image through resnet50 and align the position-wise averaged representation obtained from $\mathrm{Layer2}$ (eq. (5)$\mathcal{L}_{c3}$).

The discriminator $\mathbf{D}$ has three heads that share a common backbone: the first head $\mathbf{D}_d$ classifies images to be real/fake, the second and the third semantic projection heads $\mathbf{D}_{si}$, $\mathbf{D}_{sc}$ map $\mathbf{x}_{\text{img}}{}'$ to $\mathbf{h}'_{\text{img}}$ and $\mathbf{h}'_{\text{cap}}$. The latter two ensure the generated images are faithful to the conditions. It is also shown that contrastive discriminators are useful for preventing discriminator overfitting and improving the final model performance [17, 15]. Applying contrastive loss (eq. (5) $\mathcal{L}_{c1}$) between the outputs from discriminator semantic projection heads and the condition vectors fed to $\mathbf{G}$ can therefore help stabilize the training. To summarize the objective function, the standard GAN loss is used to ensure the naturalness of generated $\mathbf{x}_{\text{img}}{}'$:

$$\begin{aligned} \mathcal{L}_{\mathrm{GAN}_{\mathbf{G}}} &= -\mathbb{E}_i \left[\log \sigma(\mathbf{D}_d(\mathbf{x}_{\text{img}}^{(i)\prime})) \right], \\ \mathcal{L}_{\mathrm{GAN}_{\mathbf{D}}} &= -\mathbb{E}_i \left[\log \sigma(\mathbf{D}_d(\mathbf{x}_{\text{img}}^{(i)})) - \log(1 - \sigma(\mathbf{D}_d(\mathbf{x}_{\text{img}}^{(i)\prime}))) \right], \end{aligned} \tag{4}$$

where σ denotes the Sigmoid function. Meanwhile, contrastive losses are used to align the semantics of generated images and the fMRI-mapped condition vectors that supposedly residing in the CLIP space:

$$\begin{aligned} \mathcal{L}_{c1} &= \mathrm{Contra}(\mathbf{D}_{sc}(\mathbf{x}_{\text{img}}^{(i)\prime}), \mathbf{h}_{\text{cap}}^{(i)\prime}) + \mathrm{Contra}(\mathbf{D}_{si}(\mathbf{x}_{\text{img}}^{(i)\prime}), \mathbf{h}_{\text{img}}^{(i)\prime}), \\ \mathcal{L}_{c2} &= \mathrm{Contra}(\mathrm{C}_{\text{img}}(\mathbf{x}_{\text{img}}^{(i)\prime}), \mathbf{h}_{\text{cap}}^{(i)\prime}) + \mathrm{Contra}(\mathrm{C}_{\text{img}}(\mathbf{x}_{\text{img}}^{(i)\prime}), \mathbf{h}_{\text{img}}^{(i)\prime}), \\ \mathcal{L}_{c3} &= \mathrm{Contra}(\mathrm{ResNet}(\mathbf{x}_{\text{img}}^{(i)\prime}), \mathrm{ResNet}(\mathbf{x}_{\text{img}}^{(i)})) \end{aligned} \tag{5}$$

The overall training objectives are: $\mathcal{L}_{\mathbf{G}} = \mathcal{L}_{\mathrm{GAN}_{\mathbf{G}}} + \lambda_1 \mathcal{L}_{c1} + \lambda_2 \mathcal{L}_{c2} + \lambda_3 \mathcal{L}_{c3}, \mathcal{L}_{\mathbf{D}} = \mathcal{L}_{\mathrm{GAN}_{\mathbf{D}}} + \lambda_1 \mathcal{L}_{c1}$, where $\lambda_1, \lambda_2, \lambda_3$, are non-negative hyperparameters.

The whole generation pipeline, consisting of mapping models and GAN, is trained in two stages. First, mapping models f_{mi} and f_{mc} are trained on fMRI-CLIP embedding pairs. Next, starting from the trained mapping model weights and Lafite language-free model weights, we modify the losses and model structure and finetune the conditional generator. For the additional condition vector projection layers in $\mathbf{G}$ and semantic head in $\mathbf{D}$, we duplicate the weights in the existing parallel layers to make the model converge faster. Note that Lafite is pre-trained on the Google Conceptual Captions 3M dataset [31] then finetuned on the MS-COCO dataset, both of which are much larger than NSD. Finetuning from it allows us to exploit the natural relationships between semantics and images with sparse fMRI data. We can still utilize a two-stage training to compensate for data scarcity even if no pre-trained conditional GAN like Lafite is available, for example, when using a different generator architecture. Only this time, we should firstly train the conditional GAN on a large image dataset with noise perturbed $\mathbf{h}_{\text{img}}$ and $\mathbf{h}_{\text{cap}}$ as the pseudo input condition vectors.

3 Results

3.1 Data and experimental setup

The NSD data is collected from eight subjects. We focus on reconstructing observed scenes from a single subject's brain signals. The reasons are twofold: first, it is more accurate to utilize individual brain coordinates instead of mapping voxels into a shared space, which can result in information loss during the process. More importantly, brain encoding and perception are different among individuals. This project aims to get the best reconstruction for a single individual, thus training models on one subject's data. Nevertheless, the commonality of this encoding process among the population is an exciting topic for future explorations.

We use subject one from NSD: the available data contains 27750 fMRI-image sample pairs on 9841 images. Each image repeats up to three times during the same or different scan sessions. Note that brain responses to the same image can differ drastically during the repeats (fig. 6). The dataset is split image-wise: 23715 samples corresponding to 8364 images are used as the train set, and 4035 samples corresponding to the remaining 1477 images are used as the validation set. Therefore, our pipeline never sees the image it will be tested on during the training. We use 1pt8mm-resolution scans and only consider fMRI signals from voxels in the $\mathrm{nsdgeneral}$ ROI provided by the dataset. This ROI covers voxels responsive to the NSD experiment (voxels with high SNR) in the posterior aspect of the cortex, and contains 15724 voxels for subject one ($\mathbf{x}_{\text{fmri}} \in \mathbb{R}^{15724}$). Images are all scaled to 256×256. Additional experiment settings , including hyperparameters of two training phases, are provided in appendices A.1 and A.2. Our experiments are conducted on one Tesla V100 GPU and one Tesla T4 GPU. All codes will be publicly available.

3.2 Mapping models from fMRI to CLIP embeddings

Evaluation Criteria In the first training stage, mapping models f_{mi} and f_{mc} are trained to encode fMRI signals to CLIP embeddings. We use two criteria to evaluate the mapper performance to decide which one to use in the next stage. The first criterion is FID (Fréchet Inception Distance) [13] between generated image and ground truth using the trained mapper and a pre-trained generator. Given a Lafite model pre-trained on MS-COCO (language-free setting), we can replace its conditional vector with the outputs of our mapping models to generate images conditionally. These FIDs can indicate the starting points of the finetuning processes: the lower the FID, the better the candidate model. Secondly, we use the success rate of image "retrieval" in a batch of size 300. For the ith sample in the batch, if the cosine similarity between $\mathbf{h}^{(i)\prime}$ and $\mathbf{h}^{(i)}$ is larger compared to between $\mathbf{h}^{(i)\prime}$ and $\mathbf{h}^{(j)}, j \neq i$, then it counts as one successful forward retrieval. For backward retrieval, we count the number of correct matches of $\mathbf{h}^{(i)}$ to all $\mathbf{h}^{(j)\prime}$.

Configuration comparisons We tested different configurations on the mapping models, including: (1) Whether to place the threshold at ± 1.5 as mentioned in section 2.2; (2) When training f_{mi}, whether to perform image augmentations before passing images through the CLIP encoder; (3) When training f_{mc}, whether to use the CLIP text embedding of a fixed caption, a random valid caption, or use the average embedding of all valid captions; (4) Which loss function to use: MSE only, $\cos$ (cosine similarity) only, Contra only, $\mathrm{MSE} + \cos$, $\mathrm{MSE} + \cos + \mathrm{Contra}$; (5) Whether auxiliary networks help. We tested adding an auxiliary discriminator with GAN loss, as well as adding auxiliary expander networks with VICReg loss [4].

We found: (1) Clamping ground truth embeddings significantly increase performance; (2) Using image augmentations increase f_{mi} performance. This further indicates CLIP embeddings are more semantic related; (3) For f_{mc}, selecting a random caption from the valid caption pool each time is better than using a fixed one or using the average embedding of all valid captions; (4) Using $\mathrm{MSE} + \cos$ as the loss gives the best base models, but then finetune these base models with $\mathrm{MSE} + \cos + \mathrm{Contra}$ can further lower the starting FID for pipeline finetuning, making the training in the next stage converge faster; (5) Adding auxiliary networks and objectives will not improve the performance. In general, although $\mathbf{h}_{\text{cap}}$ and $\mathbf{h}_{\text{img}}$ are already relatively well aligned, f_{mc} can still map $\mathbf{x}_{\text{fmri}}$ closer to $\mathbf{h}_{\text{cap}}$ than $\mathbf{h}_{\text{img}}$, whereas f_{mi} maps $\mathbf{x}_{\text{fmri}}$ to an embedding that is equally close to both, while being able to capture a few extreme values in $\mathbf{h}_{\text{img}}$ (see appendix A.3 for numerical details and mapped embedding visualizations). We think this difference reflects that it is easier to map fMRI signals to a more semantic representation (from the text space) than to a visual one.

Figure 3: Mismatches are semantically close to the ground truth. Figure shows examples of incorrect matches j (red frame) in a batch of 300 in the validation set. For each ground truth image i (green frame), we pass it through CLIP encoder to get $\mathbf{h}^{(i)}$ and through f_{mc} to get $\mathbf{h}^{(i)\prime}$. The shown incorrect ones are those images with $\mathbf{h}^{(j)\prime}, j \neq i$ that is closer to $\mathbf{h}^{(i)}$ than $\mathbf{h}^{(i)\prime}$.

Table 1: FID of the pipeline under different settings.

FID↓			f_{mi}	f_{mc}	f_{mi} & f_{mc}
from supervised	without $\mathcal{L}_{c3}$	f_m frozen	37.75	41.51	—
from LF	without $\mathcal{L}_{c3}$	f_m frozen	30.83	33.78	50.59
from LF	with $\mathcal{L}_{c3}$	f_m frozen	**29.74**	33.35	49.47
from LF	with $\mathcal{L}_{c3}$	end to end	45.02	48.54	50.96

To verify fMRI-mapped embeddings $\mathbf{h}'$ are semantically well aligned with ground truth CLIP embeddings, we examined the mismatches during the image retrieval. For four incorrect retrievals, fig. 3 shows which images' $\mathbf{h}^{(j)\prime}$ are closer to the ground truth images' $\mathbf{h}^{(i)}$ than $\mathbf{h}^{(i)\prime}$. Notably, these mismatches are semantically close to the ground truth images. This indicates that the mapping models can successfully map fMRI signals into a semantically disentangled space. Embeddings in this space are suitable for providing contexts to a conditional generative model. We also tested another mapping model f_{mr} that maps fMRI signals to representations obtained from resnet50 Layer2. Unlike the CLIP embedding space, the resnet vector encodes more lower-level visual features. We see a jump in the image retrieval rate when we combine the representations obtained from f_{mi}, f_{mc} with f_{mr} (table 3). However, the generative model is difficult to train when taking in two conditions from distinct embedding spaces. Therefore, we add the low-level vision constraint into the contrastive loss $\mathcal{L}_{c3}$ instead.

3.3 Conditional image generation

Quantitative results In the second training stage, we finetune the conditional StyleGAN2.[1] There is no standard metric to measure image reconstruction quality from fMRI signals for complex images. Since previous works focused on reconstructing simpler images, the metrics typically involve pixel-wise MSE or correlation measures. However, when it comes to complex images, it seems more reasonable to use a perceptual metric, such as FID, which is based on Inception V3 [35] activations and is widely used in GAN. We also detail another metric, n-way identification accuracy, that reflects more of the fidelity and uniqueness of the generated images, in appendix A.4. We perform the ablation studies on the pipeline to answer the following questions: (1) Which mapping model trained in stage one leads to the best final performance? f_{mc} or f_{mi} or using both? (2) Which pre-trained GAN leads to the best final performance? For this, we compare using Lafite pre-trained on either the langue-free (LF) setting or the fully supervised setting. (3) Whether including the contrastive loss $\mathcal{L}_{c3}$ between lower-level visual features can further improve the performance of a semantic-based generative model? Finally, we tested (4) whether finetune the whole pipeline end-to-end or freezing the mapping models is better? The new mapping model losses are set to $\mathcal{L}'_m = \mathcal{L}_m + \lambda_4 \mathcal{L}_{\text{GAN}_\mathbf{G}}$ if trained end-to-end.

Results are reported in table 1. We observed the following: (1) in terms of FID, using $\mathbf{h}'_{\text{img}}$ obtained from f_{mi} as the generator condition is better than using the $\mathbf{h}'_{\text{cap}}$ from f_{mc} or using two conditional heads. On the other hand, f_{mc} and the two-head setting achieve as good or even better performance as f_{mi} does in terms of n-way identification accuracy. In addition, if training time or resource is the

[1] Codes are adapted from https://github.com/NVlabs/stylegan2-ada-pytorch, https://github.com/drboog/Lafite

(a) Ground truth stimuli (top row) and generated images conditioned on fMRI (bottom row).

(b) Generated images from three different fMRI scans responding to the same stimulus (green frames).

Figure 4: Images generated by our pipeline given input fMRI signals.

concern, using two heads and pre-trained LF-Lafite with only condition feeding interface changes and cloned weights in the new branches can already give reasonably good results. (2) Training the pipeline on LF-Lafite is much better than on the fully supervised Lafite. This result is expected for the generator conditioned on $\mathbf{h}'_{\text{img}}$ since the supervised version is conditioned on CLIP text embeddings. However, the same discrepancy exists for the generator conditioned on $\mathbf{h}'_{\text{cap}}$. This may reflect the flexibility of pre-trained generators to adapt to the slight changes in the embedding space. It also shows the crucial impact of a pre-trained model on final performance when training data is limited. (3) The addition of low-level visual feature constraint $\mathcal{L}_{c3}$ is beneficial for the model performance, especially faithfulness. It also seems to have more effects on single-head models than the two-head one. (4) For the end-to-end pipeline training, we test performance with $\lambda_4 = [0.1, 1, 10]$, all of which give worse performance than keeping the mapping model weights frozen (reported values are from $\lambda_4 = 1$). In particular, we found that $\mathbf{h}'$ tends to collapse to having nonzero values at only a few positions if the mappers are finetuned together with GAN.

Qualitative results We show several generated images in fig. 4. Although the generator takes in both the noise vector $\mathbf{z}$ and fMRI-mapped embeddings, the results vary much more with the latter condition, while $\mathbf{z}$ only contributes to variations on some minor details. In general, the generated images capture both semantics and visual features relatively well, even on complex images containing interactions of multiple objects. Since each stimulus is repeated up to three times to the subject, we have multiple fMRI scans corresponding to the same image. The semantic differences in the generated images conditioned on these multiple scans could potentially reveal brain processing discrepancies of the same stimulus. For example, the three generations for the second image in fig. 4b emphasize respectively: (1) the overall scene and the fence, (2) people with green suits, and (3) overhead flags and the fence; these might reflect the variations in the subject's attention or interpretations of that image. Eyetracking data can be further examined to study attention's effect on generated images.

It is challenging to perform one-to-one comparisons with previous deep image reconstruction works since the images in the MS-COCO dataset have much higher complexities than artificial shapes, faces, or images containing a single centered object (like in ImageNet). We show results from a few best models for reconstructing images from fMRI in fig. 5a. There is also a recent survey [27] covering more models and results if readers are interested. As our dataset is different, we search for similar images in the NSD validation set and show our generations in fig. 5b. Compared to other methods, our pipeline can generate more photo-realistic images that reflect objects' shapes and backgrounds well. It also utilizes more voxel activities than previous works (15724 voxels versus a few hundred). More importantly, it is able to reconstruct the relationships of different components when the images are more complex. As natural scenes around us are rarely isolated objects and always information-laden, we think reconstructing images through semantic alignment and conditioning is more beneficial and realistic than focusing on lower-level visual features.

Microstimulation In neuroscience, microstimulation refers to the electrical current-driven excitation of neurons and is used to identify the functional significance of a population of neurons. Here,

(a) Image reconstruction results from fMRI in previous works.

(b) Image reconstruction results from fMRI by our pipeline. Four ground truth images are green framed.

Figure 5: Comparisons between previous works and our pipeline. We are using the recent NSD dataset that involves more complex scenes. However, for comparison purposes, we choose four similar images from NSD, each containing a single object "plane", and show our reconstructions from fMRI signals in fig. 5b

we "microstimulate" the input fMRI signals of voxels in different brain ROIs, aiming to identify the roles of individual regions. In the NSD dataset, there are four floc (functional localizer) experiments targeting regions responsible for faces, bodies, places, and words. A typical standardized fMRI signal has a value range around $[-4, 4]$. For the experiment, we locate the corresponding task-specific voxels based on ROI masks and increase the voxel activities to 10 while keeping the activities in unrelated voxels unchanged (see appendix A.5 for visual results). We observe the emergence of bodies or words when we increase the voxel activities in "bodies" or "words" ROIs. For voxels in "places" ROIs, elevating the signals will result in mesh-like patterns in the background, and this is true across different images. For "faces" ROIs, the generated images under elevated facial area signals seem to contain many small repeated patterns/perturbations. Interestingly, this appears to result from FFA (fusiform face area) signal changes since increasing only OFA (occipital face area) regions' activity does not result in similar patterns. Overall, increasing a specific task ROI's signal across fMRI samples results in CLIP embedding changes in similar positions. This means the disentangled space of CLIP embedding aligns well with how the human brain processes visual cues.

Apart from task-specific ROIs, we also changed brain region activities based on their roles in the visual processing hierarchy. We use the streams mask in the dataset to identify early visual cortex ROIs, intermediate ROIs, and higher-level ROIs. We then zero out voxels at each level. Our observations are: (1) when silencing the early visual cortex, objects and the whole scene are prone to be in dull colors, and objects tend to have sharp shapes. Meanwhile, the mapped embedding in the CLIP space will constantly have a lower value at almost all positions compared to mapped from unchanged signals . (2) Silencing the higher-level ROIs has the opposite effect: more colors, more shapes, and crowded scenes. This is reasonable since the lower-level visual regions will bring up all the details when they lack high-level control. This time, the embeddings in the CLIP space have values consistently higher than normal. Finally, (3) silencing the intermediate ROIs seems to have the least visual impact or CLIP embedding changes among the three. We performed the above microstimulation experiments on our pipeline with existing ROIs; however, it is potentially helpful for testing new ROI definitions and hypotheses.

4 Further discussions

Prior to our current pipeline design, we experimented with a DALL-E-like structure [29] since we can view the image reconstruction problem as signal-to-signal translation. In particular, we applied VQVAE [38] on both fMRI and image to represent them as discrete latent codes and train a Transformer model to autoregressively generate text and image tokens from fMRI tokens. However, it was challenging to train the Transformer-based model to converge with limited fMRI-image data. Incorporating the caption as text tokens to serve as the bottleneck between fMRI and image tokens while utilizing pre-trained models on the text and image modality did not help either. We think this suggests the need to introduce a semantic medium to avoid direct translations between fMRI and image, as well as a solution to data scarcity.

We address both issues with the semantic space of CLIP embeddings. First, CLIP space is semantically informative and visually descriptive: for example, we can use image-text CLIP embedding alignment probabilities to screen captions. Mapping fMRI signals to representations in this latent space will retain rich information about the image that needs to be reconstructed. Second, the pre-training of the generative model can be separated entirely from fMRI data, meaning it can utilize much larger datasets than the one we use. However, there is a trade-off between generating a semantically similar scene and faithfully reconstructing each pixel. Although trained with additional contrastive loss targeting low-level visual features, the generated images by our pipeline are still leaning towards the former. We consider this a reasonable choice since brains are more likely to perceive the image as a whole rather than identifying each pixel, especially with multiple objects in the scene. Nevertheless, this results in worse reconstructions for images with fine details but less semantic, such as single faces. The reconstruction of complex images with better aligned low-level visual features is worth further studies.

There are many more areas to explore. First, our study focuses on reconstructing a single subject's brain signals. Applying the model to different subjects and observing the differences when generating the same image would be interesting. Since the data contains behavioral measures like valence and arousal towards each image, one can test if the generated images reflect personalized attention and perceptions. Second, other latent spaces can be examined. Although CLIP is one of the best-aligned computation models for the brain, other multimodal models like TSM [2] seem to have a better alignment [8] with the visual cortex. In addition, other conditional generative models, such as diffusion models, can be explored. In particular, DALL-E 2 [28] generates images conditioned on CLIP embeddings through diffusion, and it also provides an alternative solution to the differences exhibited in the image the text CLIP embeddings by learning a *Prior* model. Third, given the additional text modality, our pipeline opens up new opportunities to study visual imagery even without ground truth images. For example, one can either use mapping models trained on given fMRI-image pairs and pre-trained generators to reconstruct imagined scenes, or study the mapping between brain signals and the text embeddings of the mental images' descriptions. Lastly, we focus on the decoding (brain-to-image) process, but the encoding (image-to-brain) process of complex images is equally important and exciting.

With current brain signal recording devices, the negative social impact of this work is minimal: portable devices like EEG have poor spatial resolutions, making them unlikely to provide enough image-related details; On the other hand, fMRI scanners are used under highly controlled settings with designed procedures, therefore unlikely to have subject-unapproved privacy violations. However, when new devices that can address these issues become readily available, regulations would be needed on collecting and inspecting user data, since they potentially reveal sensitive information that users are unwilling to share through neural decoding. With pre-trained components, the pipeline may also misinterpret brain signals or be hacked to generate from manipulated inputs (no matter how unlikely it is) and produce over-confident false reconstructions because of the training data distributions. Several tricks may alleviate this issue, for example, training an input discriminator and placing it before the entire pipeline to filter out suspicious inputs. Or, using a parallel pipeline targeting pixel-level reconstruction as a check: if the two systems agree with each other above a certain threshold, the reconstruction results are accepted, otherwise discarded. Future pipeline improvements should also focus on exploring high-performing models pre-trained on large (thus more generalizable) and unbiased datasets (more in appendix A.8).

5 Conclusion

The paper proposes a pipeline to reconstruct complex images observed by subjects from their brain signals. With more objects and relationships presented in the image, we bring in an additional text modality to better capture the semantics. To achieve high performance with limited data, we utilize pre-trained semantic space that aligns visual and text modalities. We first encode fMRI signals to this visual-language latent space and use a generative model conditioned on the mapped embeddings to reconstruct the images. We also introduce additional contrastive loss to incorporate low-level visual features into this semantic-based pipeline. As a result, the reconstructed images by our method are both photo-realistic and, most of the time, can faithfully reflect the image content. This brain signal to image decoding pipeline opens new opportunities to study human brain functions through strategic input alterations and can even potentially be helpful for human-brain interfaces.

References

[1] Y. Akamatsu, R. Harakawa, T. Ogawa, and M. Haseyama. Perceived image decoding from brain activity using shared information of multi-subject fmri data. *IEEE Access*, 9:26593–26606, 2021.

[2] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.

[3] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[4] A. Bardes, J. Ponce, and Y. LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[5] R. Beliy, G. Gaziv, A. Hoogi, F. Strappini, T. Golan, and M. Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Y. Cao, C. Summerfield, H. Park, B. L. Giordano, and C. Kayser. Causal inference in the multisensory brain. *Neuron*, 102(5):1076–1087, 2019.

[7] N. Chang, J. A. Pyles, A. Gupta, M. J. Tarr, and E. M. Aminoff. BOLD5000: A public fMRI dataset of 5000 images. *arXiv preprint arXiv:1809.01281*, 2018.

[8] B. Choksi, M. Mozafari, R. Vanrullen, and L. Reddy. Multimodal neural networks better explain multivoxel patterns in the hippocampus. In *Neural Information Processing Systems (NeurIPS) conference: 3rd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM 2021)*, 2021.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] T. Fang, Y. Qi, and G. Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, 33:13038–13048, 2020.

[11] G. Gaziv, R. Beliy, N. Granot, A. Hoogi, F. Strappini, T. Golan, and M. Irani. Self-supervised natural image reconstruction and rich semantic classification from brain activity. *bioRxiv*, 2020.

[12] A. A. Ghazanfar and C. E. Schroeder. Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, 10(6):278–285, 2006.

[13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[14] T. Horikawa and Y. Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15, 2017.

[15] J. Jeong and J. Shin. Training gans with stronger augmentations via contrastive discriminator. *arXiv preprint arXiv:2103.09742*, 2021.

[16] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[17] M. Kang and J. Park. Contragan: Contrastive learning for conditional image generation. *Advances in Neural Information Processing Systems*, 33:21357–21369, 2020.

[18] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020.

[19] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[22] M. Mozafari, L. Reddy, and R. VanRullen. Reconstructing natural scenes from fMRI patterns using bigbigan. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

[23] A. Pasqualotto, M. L. Dumitru, and A. Myachykov. Multisensory integration: Brain, body, and world. *Frontiers in Psychology*, 6:2046, 2016.

[24] P. Pietrini, M. L. Furey, E. Ricciardi, M. I. Gobbini, W.-H. C. Wu, L. Cohen, M. Guazzelli, and J. V. Haxby. Beyond sensory images: Object-based representation in the human ventral pathway. *Proceedings of the National Academy of Sciences*, 101(15):5658–5663, 2004.

[25] S. F. Popham, A. G. Huth, N. Y. Bilenko, F. Deniz, J. S. Gao, A. O. Nunez-Elizalde, and J. L. Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11):1628–1636, 2021.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[27] Z. Rakhimberdina, Q. Jodelet, X. Liu, and T. Murata. Natural image reconstruction from fMRI using deep learning: A survey. *Frontiers in neuroscience*, 15:795488, 2021.

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

[29] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[30] Z. Ren, J. Li, X. Xue, X. Li, F. Yang, Z. Jiao, and X. Gao. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228:117602, 2021.

[31] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[32] G. Shen, K. Dwivedi, K. Majima, T. Horikawa, and Y. Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, page 21, 2019.

[33] G. Shen, T. Horikawa, K. Majima, and Y. Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.

[34] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[36] S. Takada, R. Togo, T. Ogawa, and M. Haseyama. Question answering for estimation of seen image contents from multi-subject fmri responses. In *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, pages 712–713. IEEE, 2020.

[37] A. Van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.

[38] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[39] R. VanRullen and L. Reddy. Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications biology*, 2(1):1–10, 2019.

[40] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See section 4
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See section 4
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See appendix A.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See tables 4 and 5 in appendix A.4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See section 3.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See section 3.3.
   (b) Did you mention the license of the assets? [No] The code and the data are proprietary.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Additional codes are provided in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The dataset paper containing the details is cited.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The dataset paper containing the details is cited.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Data

fMRI data fMRI activities differ when the same individual sees the same image at different times (fig. 6). Although we use *activities* and *signals* interchangeably throughout the paper, what we mean are fMRI *betas* in the NSD dataset. Betas are not direct measurements of BOLD (blood-oxygenation-level dependent) changes, but the inferred activities from BOLD signals through GLM (general linear models). The reason for using betas instead of direct measurements is that image stimuli are shown consecutively to the subjects without prolonged delay, and activities triggered by the previous image can interfere with the next one if there is no proper separation. Authors of NSD proved the effectiveness of their GLM approach with much improved SNR in the betas over raw measurements [3].

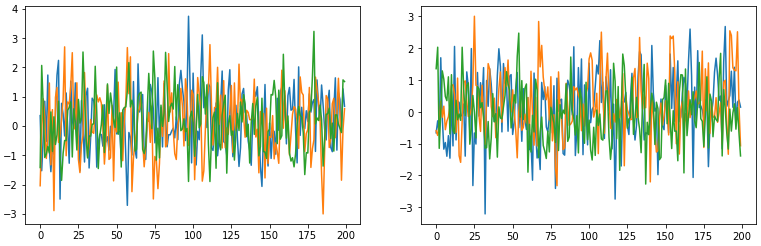


Figure 6: fMRI activities responding to two images, each repeating three times. The figure only shows the activities of the first 200 voxels for visualization purposes.

Image augmentation during training Based on conclusions from StyleGAN2-ADA [18], we perform the following image augmentations before passing images into the CLIP encoder when training the fMRI-CLIP mapping model:

- perform random sized crop with a scale between 0.8 to 1.
- perform horizontal flip with probability $p = 0.5$.
- perform $\text{ColorJitter}(0.4, 0.4, 0.2, 0.1)$ with $p = 0.4$.
- perform grayscale with $p = 0.2$.
- perform Gaussian blur with $p = 0.5$ and kernel size 23.
- perform random masking with 0.3 masking ratio.

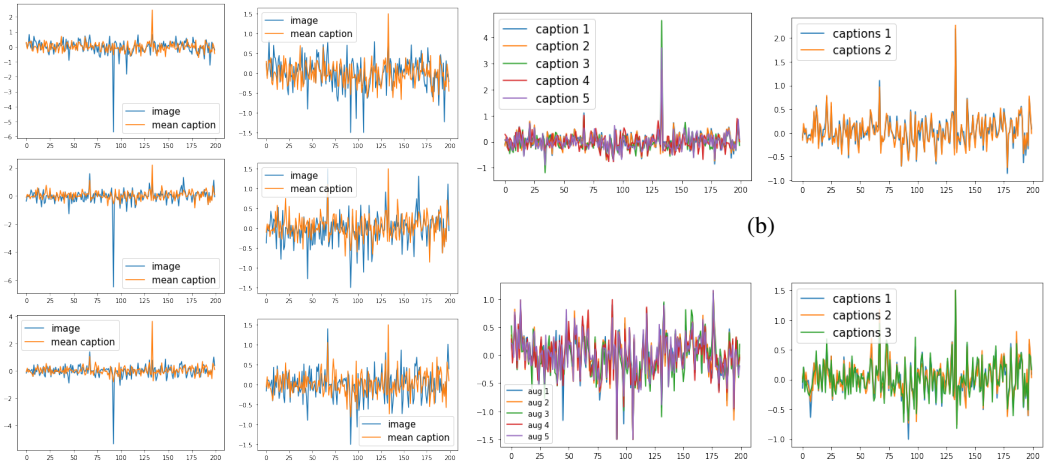
We test mapping models trained with and without the above augmentations, and found augmentations can improve fMRI to CLIP image embedding mapping performance (details are in table 2).

CLIP embeddings and thresholding See fig. 7 for the visualizations of CLIP embeddings that show image and text embedding differences, effects of thresholding, image augmentation, and random caption selection.

A.2 Experiment hyperparameters

The following hyperparameters are used in our experiments:

- $\tau = 0.5$ in eq. (1) for all the contrastive losses.
- for fMRI-CLIP mappers f_{mi}, f_{mc} (losses are in eq. (2)), the models are first trained with $\alpha_1 = 0.4, \alpha_2 = 0.6, \alpha_3 = 0$, then finetuned with $\alpha_1 = 0.2, \alpha_2 = 0.3, \alpha_3 = 0.5$.
- mappers are trained with batch size 32 (on a single GPU) when not including contrastive loss, and batch size 128 when including contrastive loss or using VICReg loss. Learning rate is 0.0004.
- $\lambda_1 = 5, \lambda_2 = 10, \lambda_3 = 10$ for the losses of conditional StyleGAN2.
- conditional StyleGAN2 is trained with batch size 16 $\times$ number of GPUs (in our case $B = 32$ since we used two GPUs). Learning rate is 0.0025.



(a) (b) (c)

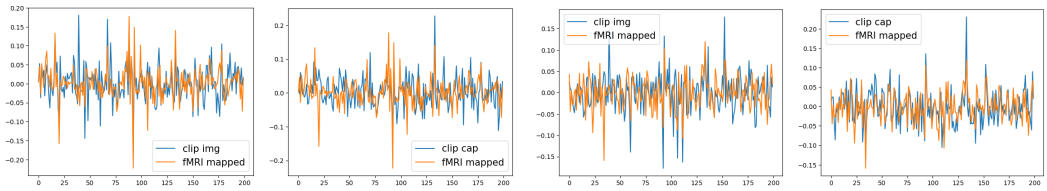
Figure 7: CLIP vector visualizations and thresholding. 7a: before (left column) v.s. after (right column) thresholding at ± 1.5 to remove outliers. There are **systematic differences between CLIP image embeddings** and text embeddings; the outliers typically occur at the same positions for each modality. 7b: the caption screening process can make the kept caption embeddings more aligned. (b)1 and (b)2 are from the same sample, only difference is the screening process. 7c: (thresholded) embeddings of the same image with different augmentations; embeddings of same image's different screened captions. All embeddings are shown the first 200 values for visualization purposes.

A.3 Results for the fMRI-CLIP mapping models f_m

Mapping models f_{mi} and f_{mc} are trained under different settings detailed in section 3.2, here we list the numerical results of the summarized findings in table 2. Simply put, forward retrieval checks the correct match of "*which ground truth CLIP embedding is the closest to the fMRI-mapped one?*" while the backward retrieval checks "*which fMRI-mapped embedding is the closest to the ground truth CLIP one?*". When multiple losses are involved, we use hyperparameter settings as in A.2.

Fig. 8 visualizes the mapping results of the best setting (models trained with threshold, image augmentation, use a random valid caption each time, pre-trained with MSE+cos loss then finetuned with MSE+cos+Contra loss).

Combining the mapped embeddings from multiple mappers boosts the retrieval performance, especially the backward one (as shown in table 3). To use multiple mapping models, we first calculate a $B \times B$ batch similarity matrix between the mapped embeddings for each model. Then we combine the similarity matrices with a weighted sum (weights are obtained through grid search) and perform image retrievals based on this combined similarity matrix. The mapping model f_{mr} that encodes fMRI to ResNet embeddings has a correct forward retrieval 6 and backward retrieval 50. But when its similarity matrix is combined with mapped-CLIP embedding similarity matrices, the performance is far above that of both ResNet and CLIP embeddings.



(a) $f_{mi}(\mathbf{x}_{\text{fmri}})$ with $\mathbf{h}_{\text{img}}$ and $\mathbf{h}_{\text{cap}}$ (b) $f_{mc}(\mathbf{x}_{\text{fmri}})$ with $\mathbf{h}_{\text{img}}$ and $\mathbf{h}_{\text{cap}}$

Figure 8: Embeddings mapped from fMRI signals overlay on ground truth CLIP embeddings. fig. 8a shows the results of image embedding mapping model f_{mi}, and fig. 8b shows the results of caption embedding mapping model f_{mc}. For visualization purposes, the figures only show the first 200 values of the length-512 vectors.

Table 2: Starting FID without generator finetuning (pre-trained LF-Lafite is used here) and correct retrievals in a batch of size 300 using embeddings obtained from f_{mi} and f_{mc}. In the top table, models are trained with MSE+cos loss. In the bottom table, defaults are: with threshold, with image augmentation, using random caption. For the two options with auxiliary modules, the model is finetuned from MSE + cos model since training from scratch gives much worse results. FID evaluations are omitted if the retrieval performance of a setting is strictly worse than its competitors.

		threshold	no threshold	**image aug**	no image aug	fixed caption	**random caption**	average caption embedding
f_{mi}	FID ↓	73.46	—	73.46	—	n/a	n/a	n/a
	Retrieval (forward) ↑	21	13	21	19	n/a	n/a	n/a
	Retrieval (backward) ↑	49	25	49	46	n/a	n/a	n/a
f_{mc}	FID ↓	75.24	—	n/a	n/a	—	75.24	79.36
	Retrieval (forward) ↑	14	11	n/a	n/a	13	14	15
	Retrieval (backward) ↑	**64**	45	n/a	n/a	39	**64**	43

		MSE	cos	Contra	MSE + cos	MSE + cos + Contra (from scratch)	**MSE + cos + Contra (from MSE + cos)**	Auxiliary GAN	Auxiliary expander (VICReg)
f_{mi}	FID ↓	—	—	—	73.46	—	**68.14**	—	—
	Retrieval (forward) ↑	5	12	25	21	27	**29**	25	19
	Retrieval (backward) ↑	16	34	50	49	50	**51**	42	35
f_{mc}	FID ↓	—	—	—	75.24	—	**53.68**	—	—
	Retrieval (forward) ↑	4	10	27	14	30	**33**	24	9
	Retrieval (backward) ↑	19	31	42	**64**	43	45	38	37

Table 3: Correct image retrievals in a batch of size 300 when combining different models.

Multiple models	$f_{mi} + f_{mc}$	$f_{mi} + f_{mc} + f_{mr}$
Retrieval (forward)	32	24
Retrieval (backward)	73	147

A.4 Additional quantitative results (generator)

In addition to using FID as a metric, we also perform 2-way identification for images reconstructed by models under different settings, and n-way identification of generated images with $n = 2, 5, 10, 50$ under the best setting (finetuned from LF, with $\mathcal{L}_{c3}$, with mapping models f_m frozen). For n-way identification, we reconstruct an image from the fMRI signal for each sample in the validation set. For each generated image, we compare it with a set of n randomly selected images, including the ground truth one. Then based on the cosine similarity of their Inception V3 embeddings (before FC layers, the length-2048 vector), we identify which image the generated one corresponds to. This process is repeated ten times because of the randomness of the n-sample selection. Results are reported in tables 4 and 5. The n-way identification accuracy of the two-head setting (f_{mi} & f_{mc}) is slightly better most of the time (table 5), but this advantage is minor when taking the standard deviations into account. Note that when performing n-way identification, previous image reconstruction works are

Table 4: 2-way identifications accuracy of the pipeline under different settings.

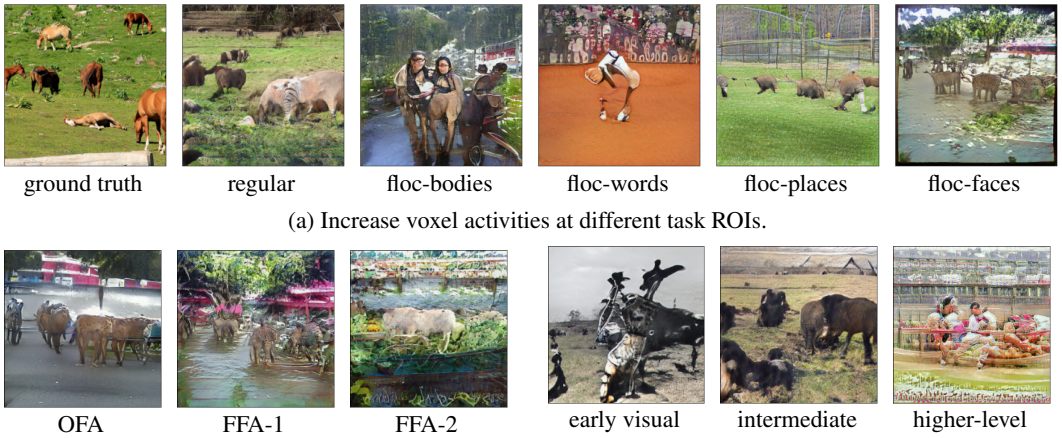
accuracy (%)			f_{mi}	f_{mc}	f_{mi} & f_{mc}
from supervised	without $\mathcal{L}_{c3}$	f_m frozen	72.6 ± 6.14	68.6 ± 5.22	—
from LF	without $\mathcal{L}_{c3}$	f_m frozen	73.0 ± 4.40	73.2 ± 4.49	76.2 ± 5.89
from LF	**with $\mathcal{L}_{c3}$**	**$\mathbf{f_m}$ frozen**	76.8 ± 4.16	$\mathbf{78.2 \pm 5.47}$	78.0 ± 4.47
from LF	with $\mathcal{L}_{c3}$	end to end	51.4 ± 5.59	50.8 ± 5.43	50.2 ± 5.31

Table 5: n-way identification accuracy (%) with $n = 2, 5, 10, 50$.

n	2	5	10	50
f_{mi}	76.8 ± 4.16	55.2 ± 3.23	41.9 ± 6.09	24.9 ± 3.98
f_{mc}	$\mathbf{78.2 \pm 5.47}$	56.4 ± 3.32	42.2 ± 4.33	25.6 ± 4.05
f_{mi} & f_{mc}	78.0 ± 4.47	$\mathbf{57.3 \pm 3.63}$	$\mathbf{44.0 \pm 6.05}$	$\mathbf{25.8 \pm 3.82}$

typically tested on a validation set that contains 50 images of 50 different categories [14]. However, there are multiple objects involved in each image in the complex images we aim to reconstruct; it is not straightforward to separate them into different categories and pick one from each. Therefore, we leave the validation set as is (1477 image-fMRI pairs in total), and there will be overlapping categories in it; for example, several images contain scenes of animals in a natural environment.

A.5 Visual results from microstimulation experiments



(a) Increase voxel activities at different task ROIs.

(b) Increase voxel activities at different face areas.

(c) Set voxel activities to 0 at different processing levels

Figure 9: Images generated in microstimulation experiments. In 9a9b, voxel activities at multiple task ROIs are increased before passed into the pipeline. In 9c, voxel activities at various visual processing stages are silenced.

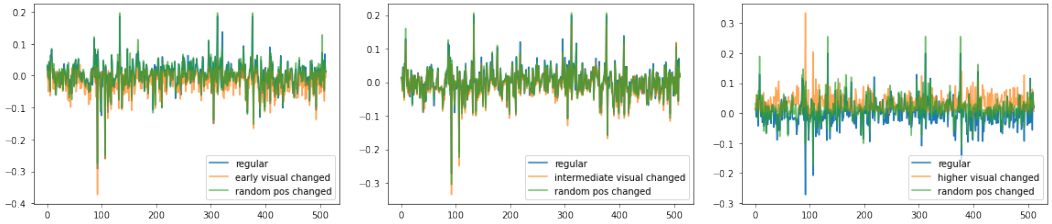


Figure 10: fMRI-mapped embeddings in the CLIP space ($\mathbf{h}'$). Each figure contains (i) an embedding mapped from a regular fMRI signal, (ii) an embedding mapped from the fMRI signals with voxel activities in earlier-visual ROIs (left)/ intermediate ROIs (middle) / higher-level ROIs (right) set to zero, (iii) an embedding mapped from the fMRI signal with voxels at random positions (same number of voxels as (ii)) set to zero. Setting activities of the earlier-visual cortex to zero lowers overall embedding vector values, while setting activities of higher-level ROIs has the opposite effect. We can also perform the reverse masking: only keep voxel activities at earlier-level visual/ intermediate / higher-level ROIs, then the effects are reversed.

Fig 9 shows generated images under different microstimulation experiments. Fig 10 shows the results regarding changes of mapped fMRI embeddings in the CLIP space when perturbing voxels in different visual cortex levels.

Figure 11: Generated images from interpolation of two fMRI scans. Step number is set to 10.

A.6 Input interpolations and the potential extension to movie reconstruction

In addition to reconstructing observed images, we found utilizing the CLIP space can also result in a smooth transition when decoding from interpolations of two fMRI scans (fig. 11). Combined with the ability to capture complex semantics, this pipeline can be helpful for movie reconstruction from brain signals. Temporal constraints can also be added, which could, in turn, benefit the reconstruction of each frame.

A.7 CLIP space as the intermediary

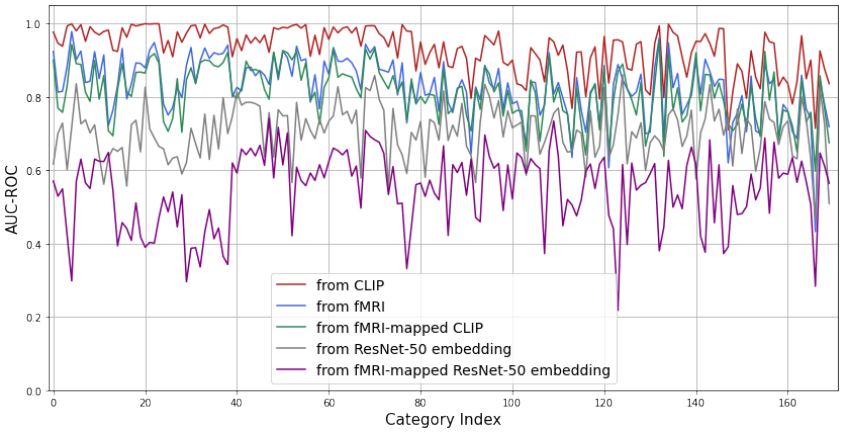


Figure 12: Category-wise AUC-ROC of multi-label classifiers that predicts from four different signal / embedding sources. The first 80 categories are "things categories" and the last 91 are "stuff categories" in COCO.

Table 6: Numerical AUC-ROC values of the classifiers presented in fig. 12.

AUC-ROC	"things" categories	"stuff" categories	Overall	Performance w.r.t. fMRI (%)
CLIP	0.9718 ± 0.0266	0.8973 ± 0.0639	0.9318 ± 0.0624	112.36
fMRI	0.8704 ± 0.0557	0.7937 ± 0.0824	0.8293 ± 0.0807	100.00
fMRI-mapped CLIP	0.8468 ± 0.0604	0.7817 ± 0.0733	0.8119 ± 0.0748	97.90
ResNet-50	0.7061 ± 0.0736	0.7032 ± 0.0719	0.7044 ± 0.0725	84.94
fMRI-mapped ResNet-50	0.5410 ± 0.1106	0.5520 ± 0.0941	0.5469 ± 0.1020	65.95

In this section, we show that multimodal embedding space, particularly the CLIP space, is beneficial for brain signal decoding. To this end, we trained a set of multi-label category classifiers to classify if a certain object category exists in the image based on the following inputs: (1) image-triggered fMRI $\mathbf{x}_{\text{fmri}} \in \mathbb{R}^{15724}$; (2) image CLIP embeddings $\mathbf{h}_{\text{img}} \in \mathbb{R}^{512}$; (3) CLIP embeddings mapped from image-triggered fMRI $\mathbf{h}'_{\text{img}} \in \mathbb{R}^{512}$; (4) image ResNet embeddings $\mathrm{ResNet}(\mathbf{x}_{\text{img}}) \in \mathbb{R}^{2048}$ (obtained from $\mathrm{Layer4}$, the final block before fully connected layers). All classifier models consist of 3 linear layers with ReLU activations in between, and finished with a $\mathrm{Sigmoid}$ activation. For fMRI signals, we use (2048, 512) as the hidden dimension; for CLIP embeddings (setting (2) and (3)), we use (384, 256) as hidden dimensions; and for the ResNet embedding, we use (512, 256) as hidden dimensions. The final output covers 171 classes, including 80 things categories (bounded objects, like "person", "car"), and 91 stuff categories (mostly unbounded objects, like "tree", "snow"). [2] Binary cross-entropy loss is used for each class to predict its existence in the input image.

[2] Please refer to https://github.com/nightrome/cocostuff/blob/master/labels.txt for the full category list.

Fig 12 shows the category-wise AUC-ROC. The result demonstrates that CLIP embeddings contain the most object-level information about the image out of all the input sources. Following it, fMRI signals are also surprisingly very predictive, considering they carry a lot of noise. The performance discrepancy between settings (2) and (3) is minimal, meaning mapping fMRI signals into the CLIP space retains most of the fMRI signals' information: this provides strong support for the validity of our design. Lastly, ResNet embeddings perform poorly compared with other input sources. Therefore, even with a perfect mapping model, projecting fMRI signals into this space will lose information about the image since the expressiveness of the embedding is bounded by the lower performer. In addition, we note that both CLIP embeddings and fMRI have poorer performance on stuff categories than on things categories, whereas ResNet embeddings do not. This can indicate brain signals align better with the multimodal CLIP space than with single-modality ResNet space. Previous brain signal decoding work utilizing pre-trained generators all relied on image-only embedding spaces (ResNet-50 [22], VGG19 [33]), and we believe moving to a multimodal latent space is a crucial step towards better brain signal decodings.

A.8 Using pre-trained models

Figure 13: Image generated by Lafite pre-trained on the CC3M dataset without finetuning on COCO or NSD. Ground truth stimuli (top row) and generated images conditioned on fMRI (bottom row).

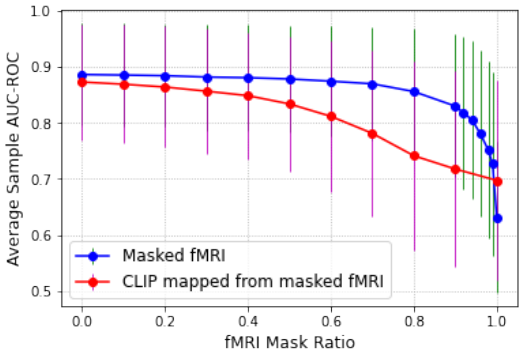


Figure 14: Multi-label classifier (defined in A.7)'s average sample-wise AUC-ROC changes when masking input fMRI at different ratios. For a masked voxel, we set its value to 0.

Our pipeline relies on two pre-trained components. The first and the most crucial one is the CLIP encoder that provides the latent space where we project fMRI signals. The second is a conditional GAN (Lafite) that generates images, which could be swapped for other generators. In what follows, we will discuss these two components separately.

CLIP One exciting aspect of CLIP is the size of its training dataset, which consists of 30 million Flicker images that should cover most of the natural image statistics. This coverage is also proved by subsequent works that generate images guided by CLIP embeddings through their abilities to perform generations in various styles. In addition, as we observed in A.7, CLIP embeddings can retain around 98% of object-level information in fMRI with a very well-aligned performance across categories.

Albeit its incredible expressive power, CLIP does have a much lower dimensionality than the original signal: no matter how faithful, it is a compression. By the nature of compression, CLIP only retains the most crucial information and removes most of the redundancies in the original signal. Indeed, if we mask fMRI at different ratios, from 0 to 1, and perform the multi-label classification (the same task as in A.7) using (1) masked fMRI or (2) CLIP mapped from masked-fMRI, we will notice a very drastic difference in the performance drop rate. As shown in fig. 14, prediction performance from fMRI only drops drastically after the masking ratio becomes larger than 0.9, indicating brain redundancies to represent the objects. In contrast, if we map the masked fMRI into the CLIP space and use these embeddings for prediction, the performance drop is almost at a constant rate. This discrepancy makes the CLIP space more vulnerable to adversarial attacks than the fMRI space since a small change would cause the generated images to derail from the ground truth. In addition, CLIP embeddings also carry more biases than fMRI, as its mean AUC-ROC is much larger even with all-masked inputs. One should consider these traits of CLIP embeddings when applying this system and design defense mechanisms accordingly.

Lafite As for the generator, we utilize a conditional GAN pre-trained on the MS-COCO dataset (containing 328K images), from which NSD drew its experiment images. This naturally provides an alignment in the data distribution. Although MS-COCO images are about everyday objects, humans, and scenes, the data statistics could vary when we move to other settings. Therefore, future studies are needed to extend current generators to one trained on broader sources (e.g., DALL-E 2, mentioned in section 4, used 650M images sampled from CLIP and DALL-E training data). This should minimize the dataset biases, although one should not interpret results without considering the training/testing discrepancies.

To show that our concept works across different generators, but dataset biases indeed play an important role, we test our pipeline with a Lafite pre-trained on the Google Conceptual Captions 3M dataset (CC3M, consisting of 3.3 million images) as the generator *without any extra finetuning*. We used our trained f_{mc} as the mapping function. The results are shown in fig. 13. All generated images have the watermark where CC3M sources its images. In addition, when trying to generate out-of-distribution images, the quality decreases in terms of photo-realism. Nonetheless, semantic alignments are still shown in these reconstructions. We also want to note that pre-trained models provide excellent bases for finetuning. For example, Lafite finetuned its COCO model on the CC3M model within three hours, compared to four days to reach the same performance if training from scratch. Therefore, if the pipeline is known to be used on certain types of images, a small-scale dataset and some light training should greatly help the model to fit into the desired data distribution.

A.9 More examples

As mentioned in section 2.2, we found that the generated results conditioned on embeddings of different modalities tend to emphasize different aspects: more visual (colors, shape, etc.) if conditioned only on $\mathbf{h}'_{\text{img}}$, and more semantic if conditioned only on $\mathbf{h}'_{\text{txt}}$. This could reflect the slight difference between the latent space of the two modalities. We show the examples conditioned on either one of these two conditions, or both, in fig. 15.

The pipeline tends to fail under the following conditions: (1) the image only contains close-up details without too much semantic information; (2) the presented scene is semantically novel (e.g., a big banana-shaped decoration hanging in the middle of the room). The model also tends to: (3) generate based on data biases: adding windows to indoor scenes, adding people to food scenes, generating colored images when the inputs are black-and-white, etc.; (4) change or ignore the background; (5) Mix-up colors (assigning colors in the scene to a wrong object); (6) generate the wrong number of objects/people. We showcase these failures together with more other generated images in fig. 16. Given that the model is confident (in terms of GAN's discriminator output staying at the same level) when generating results based on training data biases, future extensions should focus on exploring generators pre-trained on a much larger dataset, as discussed in A.8.

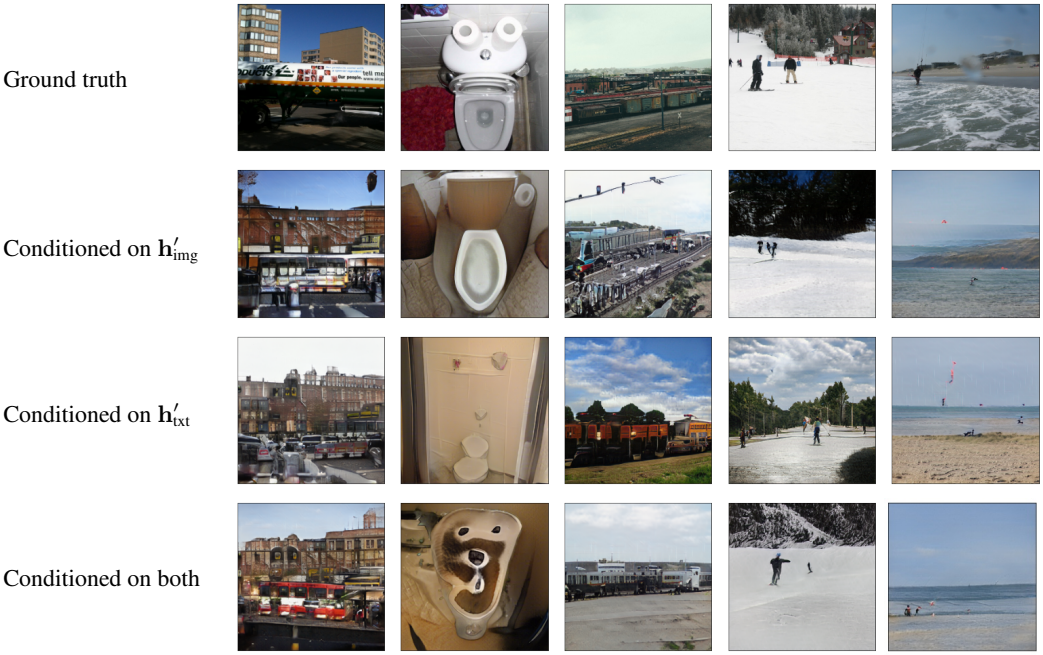


Figure 15: Generated images conditioned on fMRI-mapped CLIP image embedding $\mathbf{h}'_{\text{img}}$, fMRI-mapped CLIP text embedding $\mathbf{h}'_{\text{txt}}$, or both.

Figure 16: More examples showcasing model successes and failures. For each two-row group, the top row shows the ground truth images, and the bottom row shows the reconstructions.