We thank all reviewers for the valuable and constructive comments.

To Reviewer #1

What stimulus set was being used to train the GAN, and for augmentation? The fMRI dataset includes 1200 training images (150 categories) and 50 test images with the corresponding fMRI recordings. The GAN was trained with the 1200 training images in the fMRI dataset. In augmentation, we sampled extra 200 images from the ImageNet dataset. The images used for augmentation have no overlap with both training and test images from the fMRI dataset.

Is Figure 2(a) displaying results for test stimuli or training stimuli? Figure 2(a) displays the results for test stimuli. We will make it clear in the final version. Thanks.

How did the hyperparameter values being selected? In the training of GAN, 100 fMRI recordings of the training set were kept as a validation set for hyperparameter selection. Then we used the whole training set (containing the samples in validation set) to train the model with the selected hyperparameters. We will make it clear in the final version.

Why use a DNN for the semantic features instead of a linear regression, as for the visual ones? Why Tanh instead of ReLU? We have evaluated both DNN and linear regression for semantic decoding and did not put it in the paper for space limit. In this experiment, the DNN model obtained better performance (53.9% v.s. 49.1% in classification accuracy). About the activation function, using Tanh function can converge stably in this experiment. We agree that functions like ReLU can also work in this model. We will explore different models for semantic decoding in the future study.

About the evaluation and comparison criterion. Thanks for your constructive suggestions. We did not conduct other comparison methods because the codes of some papers were not available during the experiment. We are trying to supplement some results of the Friedman test in the final version as suggested.

For the other suggestions/issues, we will revise the paper accordingly. Many thanks for your valuable comments.

To Reviewer #2

About the contributions. Thanks for confirming our contributions. The previous study [17] provides a good basis and support for our approach. We will clarify the contributions in the introduction, and add extra discussions in the final version. In addition, we will add necessary details to the unclear parts in the article.

To Reviewer #4

Why using the U-Net structure in the generator? The advantage of using U-Net lies in that it both processes high-level information (passed through in the bottleneck) and preserves low-level structures (by skip connections). For image reconstruction, we need to combine both high-level (category) and low-level (shape) features, thus U-Net is a suitable choice.

Regarding the necessity of the HVC signals and loss of information. We agree that theoretically LVC information should contain the information from HVC. However, in the process of recording fMRI signals, the LVC signals cannot be guaranteed to contain all the useful information. Experiment 3.4 compares the performance of decoding semantic information from different cortical areas, which shows that fMRI signals from HVC contain more reliable high-level information (see Figure 3(a)). Therefore, the fMRI signals from HVC should be much helpful for the reconstruction. In reconstruction, our approach did not explicitly model other features (e.g. color). However, part of the information might be implicitly encoded in the weights of the GAN model during the training phase. In the future work, we will try to explicitly decode more critical information for improvement.

Why and how the images' classes were merged? In the original dataset there are only 8 images within one category (totally 150 categories). Training the DNN model with so few samples within one category is difficult. To facilitate the model's training, the number of images in each category was increased by merging similar classes. We manually grouped similar images from subdivide categories (e.g. gorilla and chimpanzee) into new coarser categories. We will add a supplementary material to explain the details.

Was the testing set withheld for the training of the decoders as well? Yes. The test set was strictly withheld in the training of the shape decoder, semantic decoder, and the GAN model. Thanks.

About details of implementation. We will include a supplemental material to describe the preprocessing of fMRI data and network implementation details. And we will explain in detail what is not clear in the paper. For the other suggestions/issues, we will revise the paper accordingly. Many thanks for your valuable comments.