NUS National University of Singapore
Yong Loo Lin School of Medicine
MULTIMODAL NEUROIMAGING LABORATORY

Seeing Beyond the Brain

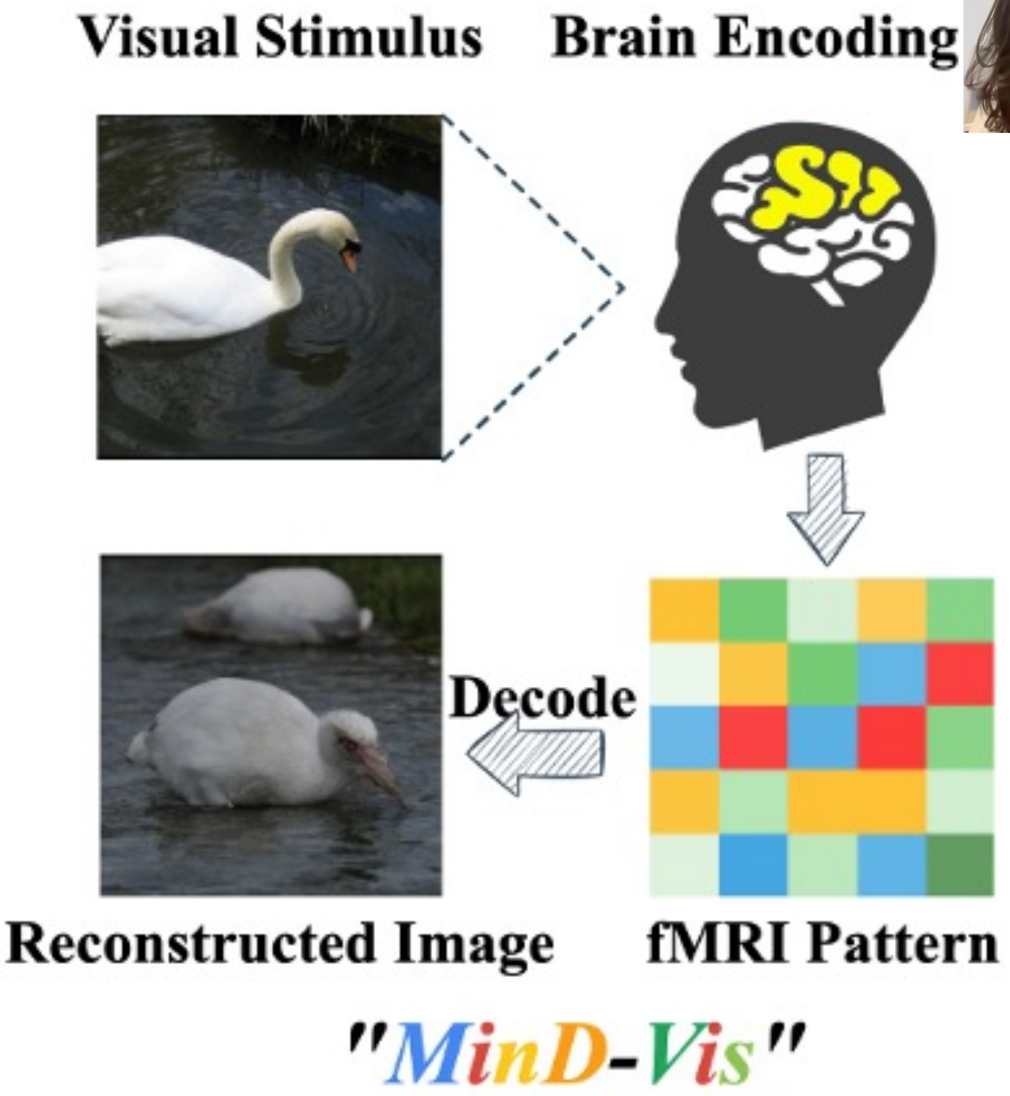
Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding

Zijiao Chen[1*] Jiaxin Qing[2*] Tiange Xiang[3]
Wan Lin Yue[1] Juan Helen Zhou[1]

[1]National University of Singapore [2]The Chinese University of Hong Kong
[3]Standford University [*]Equal Contribution

CVPR

香港中文大學工程學院
The Chinese University of Hong Kong
Faculty of Engineering
Stanford ENGINEERING
Computer Science



Poster session: THU-PM-201

fMRI - Functional magnetic resonance imaging

- Measures the small changes in blood flow

  blood-oxygen-level-dependent (BOLD) signal

- Proxy of brain activity
- High spatial resolution

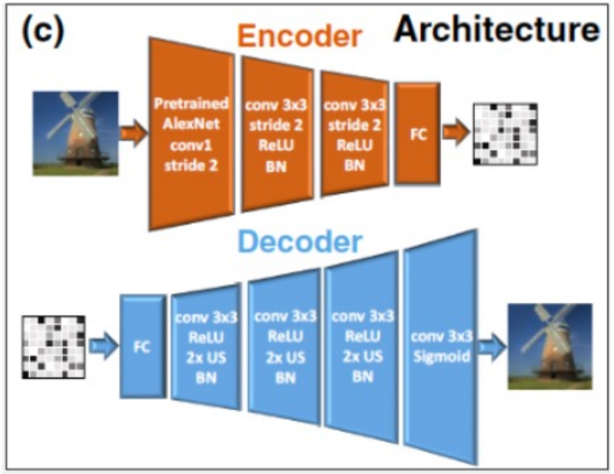
  about 1 millions voxels in a brain

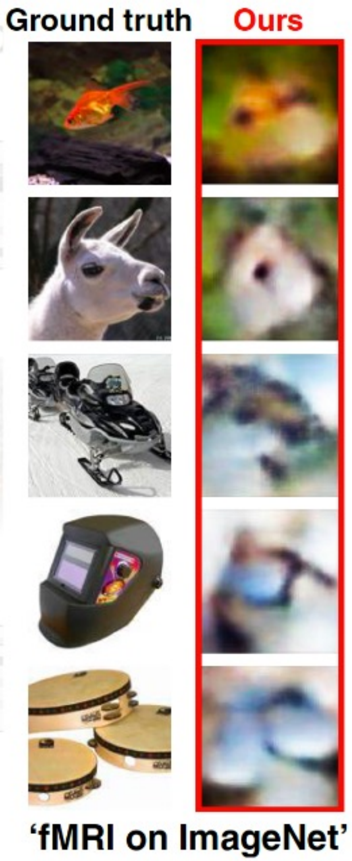
- Low temporal resolution

  TR = 1-2s

Previous methods on fMRI decoding – first reconstruction work

VAE-like architecture





Generated images are still blurry

From voxels to pixels and back: Self-supervision in natural-image reconst
fMRI, NeurIPS 2019

Gap & Solution

Gap

- **Non-linear implicit relationships** within brain activities -> highly complex

  Solution: Effective representation learner

- **Individual differences** are huge -> domain shift

  Solution: Pre-train on a large-scale dataset with only fMRI

  Pre-training dataset: Human connectome project on 1000+ subjects

- {fMRI, Image} pairs are **limited** -> few-shot learning

  Solution: Self-supervised learning with pre-text task

Two stage design

A. Self-supervised **representation learning** on large-scale fMRI dataset

B. Strong image **generation model**

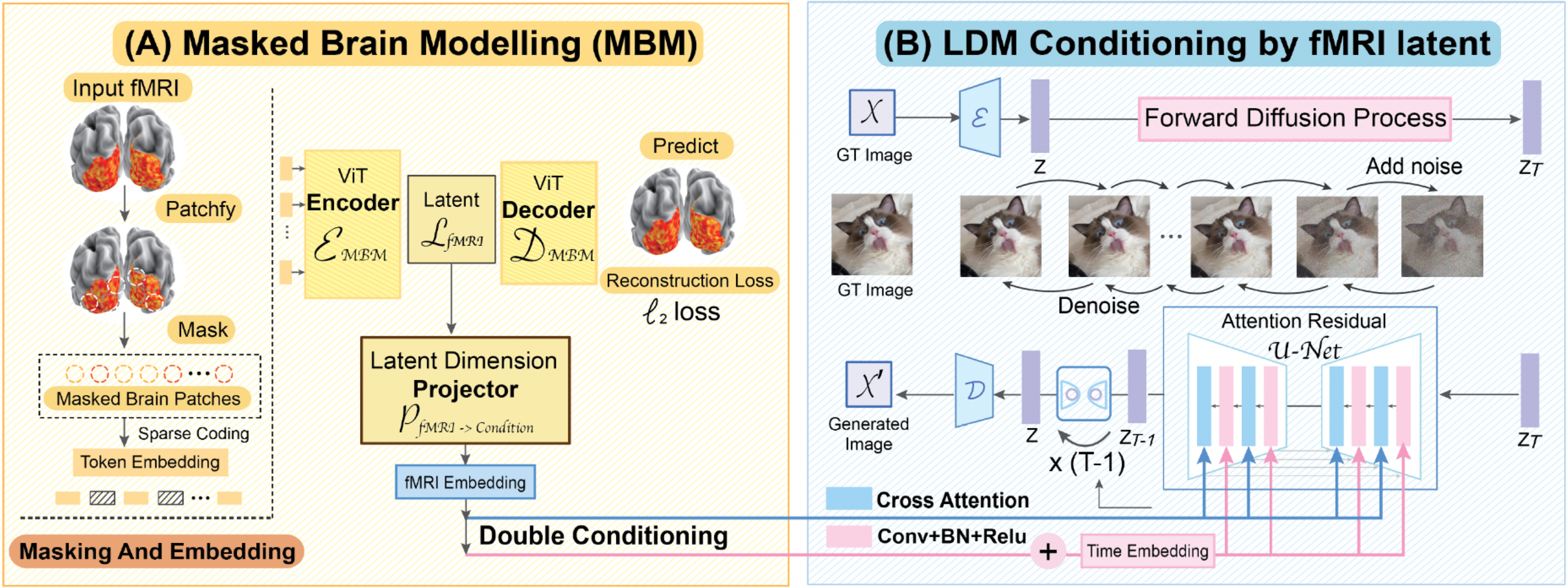
Characteristics of fMRI

- Spatial redundancy in fMRI due to regional homogeneity

- Number of voxels in VC is a lot less than images -> Difference in encoding/decoding strategy
  - Visual cortex: around 4000 voxels
  - Images: 256*256*3 = 200k voxels

- Both generation consistency and flexibility are desired
  - Consistency: For a fixed stimulus, we wish the generated images to have the same semantic meanings
  - Flexibility: Due to individual differences, each person's response to this visual stimulus is different, and we also hope that the model has a certain degree of variance and flexibility

MinD-Vis Overview



Stage A: Pre-train on fMRI only with SC-MBM

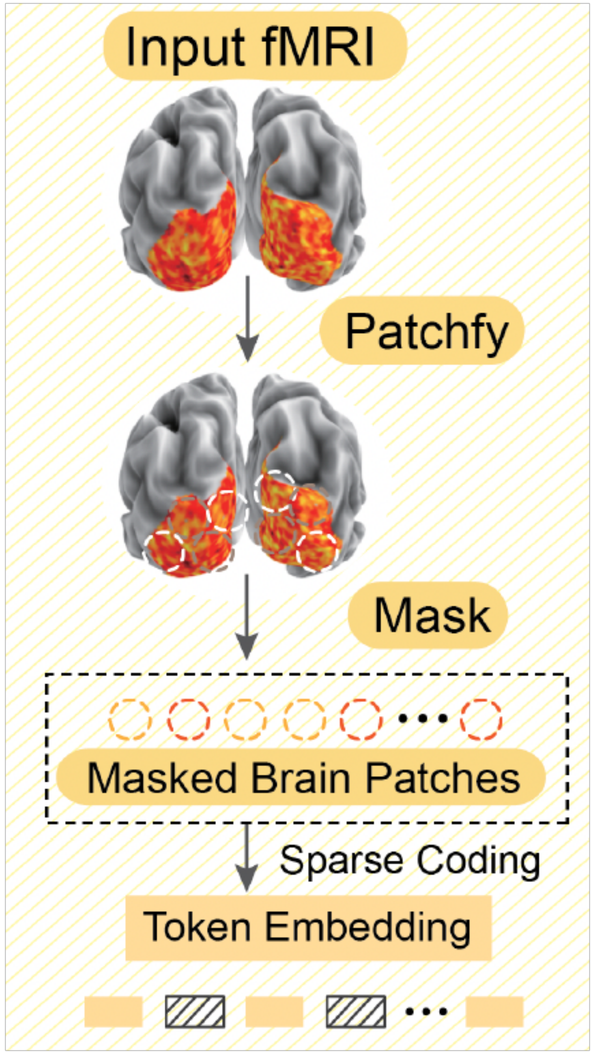
- Patchify
- Random mask
- Tokenize to large embedding
- Recover to masked patches

Stage B: Integration with LDM through double conditioning

- Project the fMRI latent using latent dimension projectot
  - fMRI latent -> cross-attention heads
  - fMRI latent + time embedding -> residual blocks
- Latent diffusion model finetune
- Image latent -> Image

Stage A: Masked Brain Modelling (MBM)

Architecture: Masked Autoencoder with Vision Transformer backbone



on HCP dataset (fMRI only)

Masking and embedding

Input: (# of subject, # of channel, # of voxels)

Steps:
1. Patchify -> (# of subject, # of patch, patch size), record position of each patch

2. Token embedding -> (# of subject, # of patch, embedding dimension), through a conv layer

3. Random masking -> e.g. make 75% of the embedding zero

Output: Tokenized patches

Reconstruction

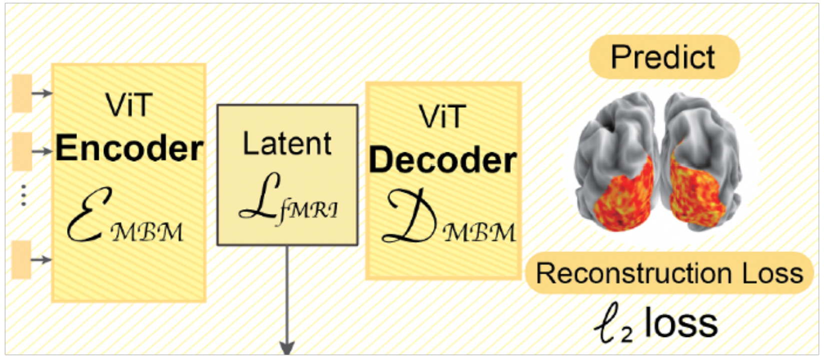
Input: Tokenized patches

Steps:
1. Token embedding -> ViT encoder -> Latent representation

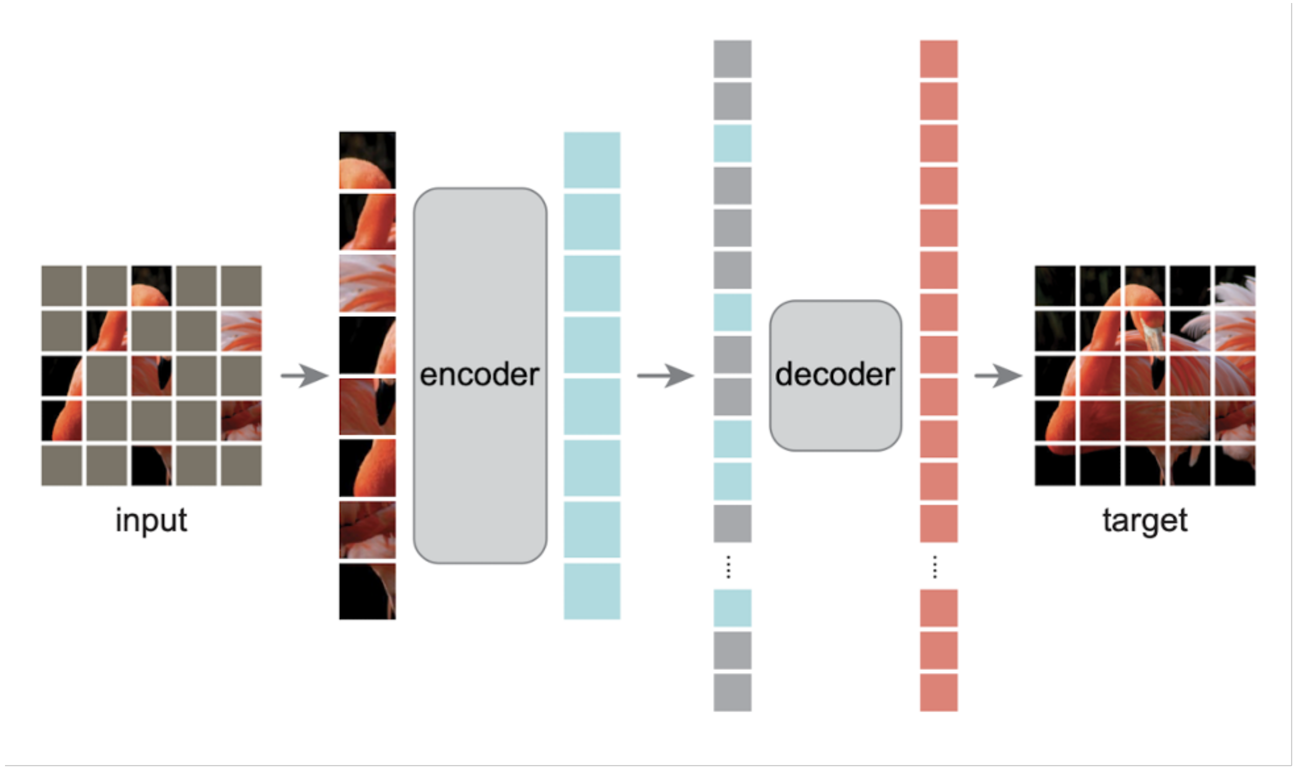
2. Latent representation -> ViT decoder -> Reconstructed brain patches

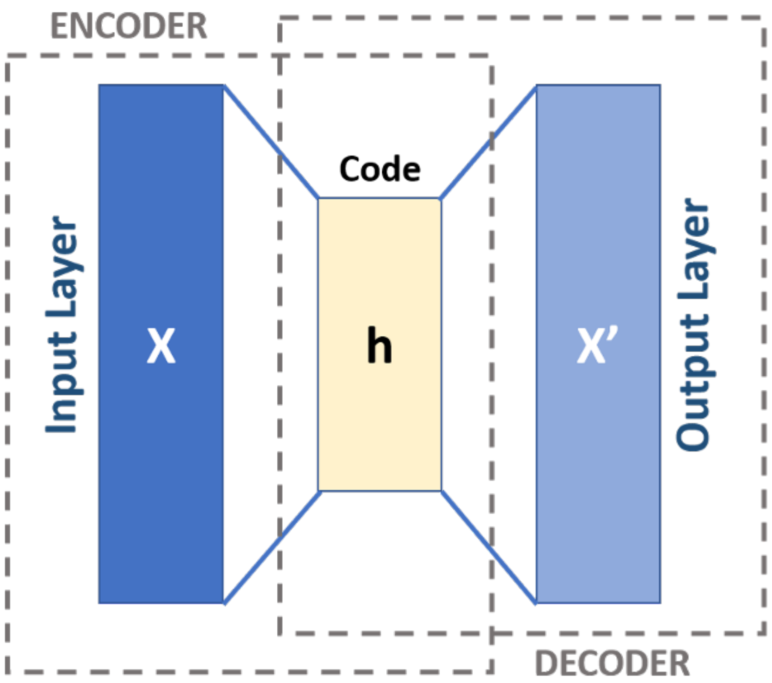
3. Calculate loss: L2 (reconstructed patches, original patches)

Output: whole brain voxels



Masked Autoencoders

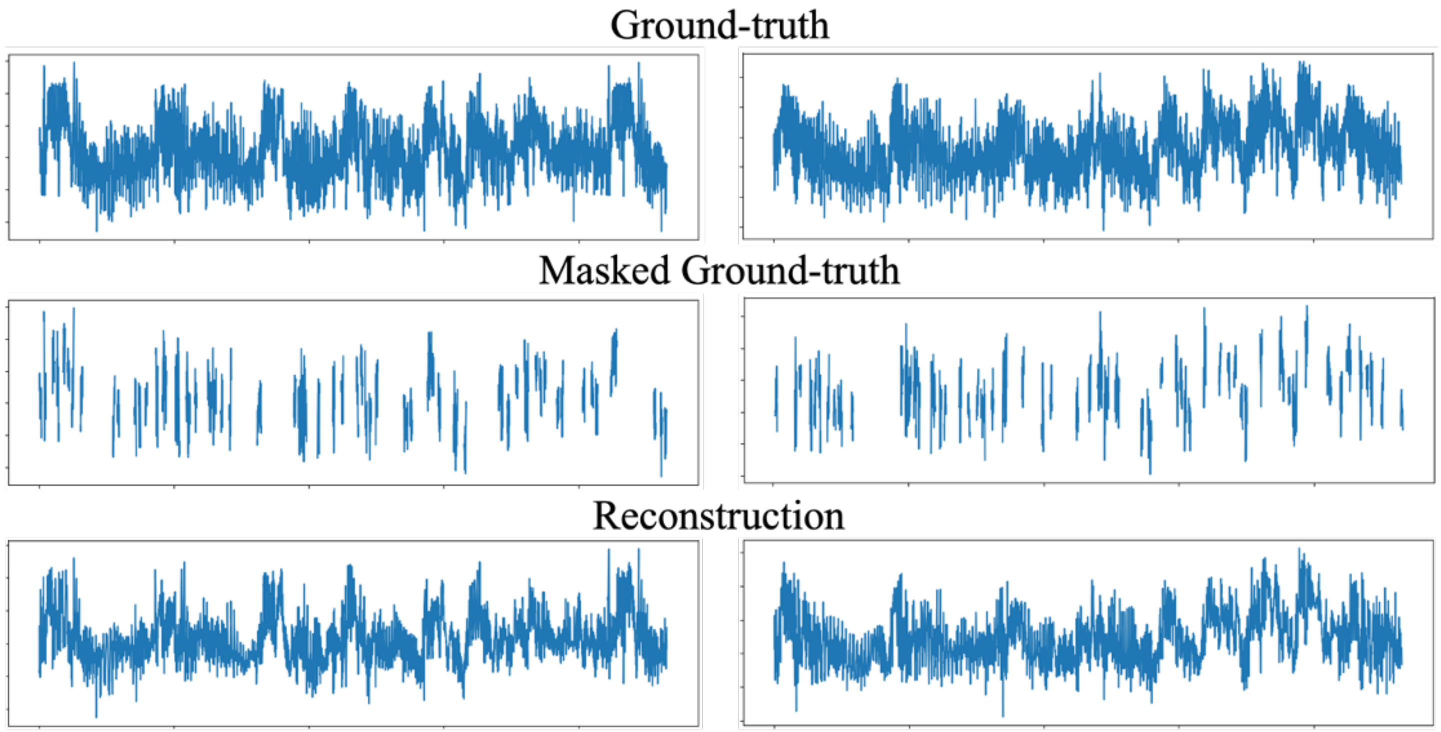




Encoder: maps the input into Code (h) - lower-dimensional representation of the input
Decoder: maps the Code (h) followed by the encoder and reconstructs the input.

(He 2022, CVPR)

Result for Stage A



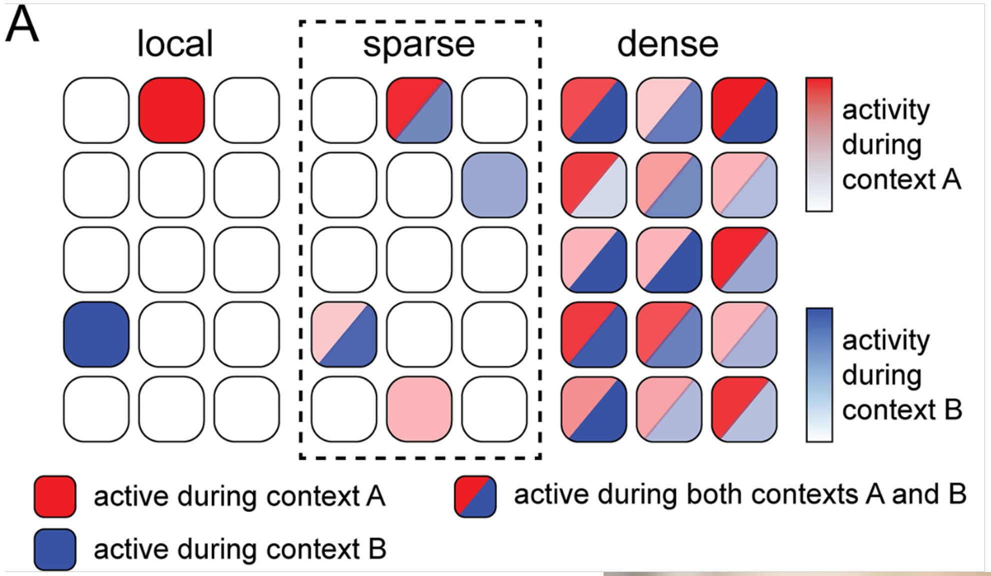
Note

- The quality of the reconstructed brain voxels are not directly related to the generation result
- We only use the latent representation in the next step

Sparse Coding with SC-MBM

Biological inspired design in MBM

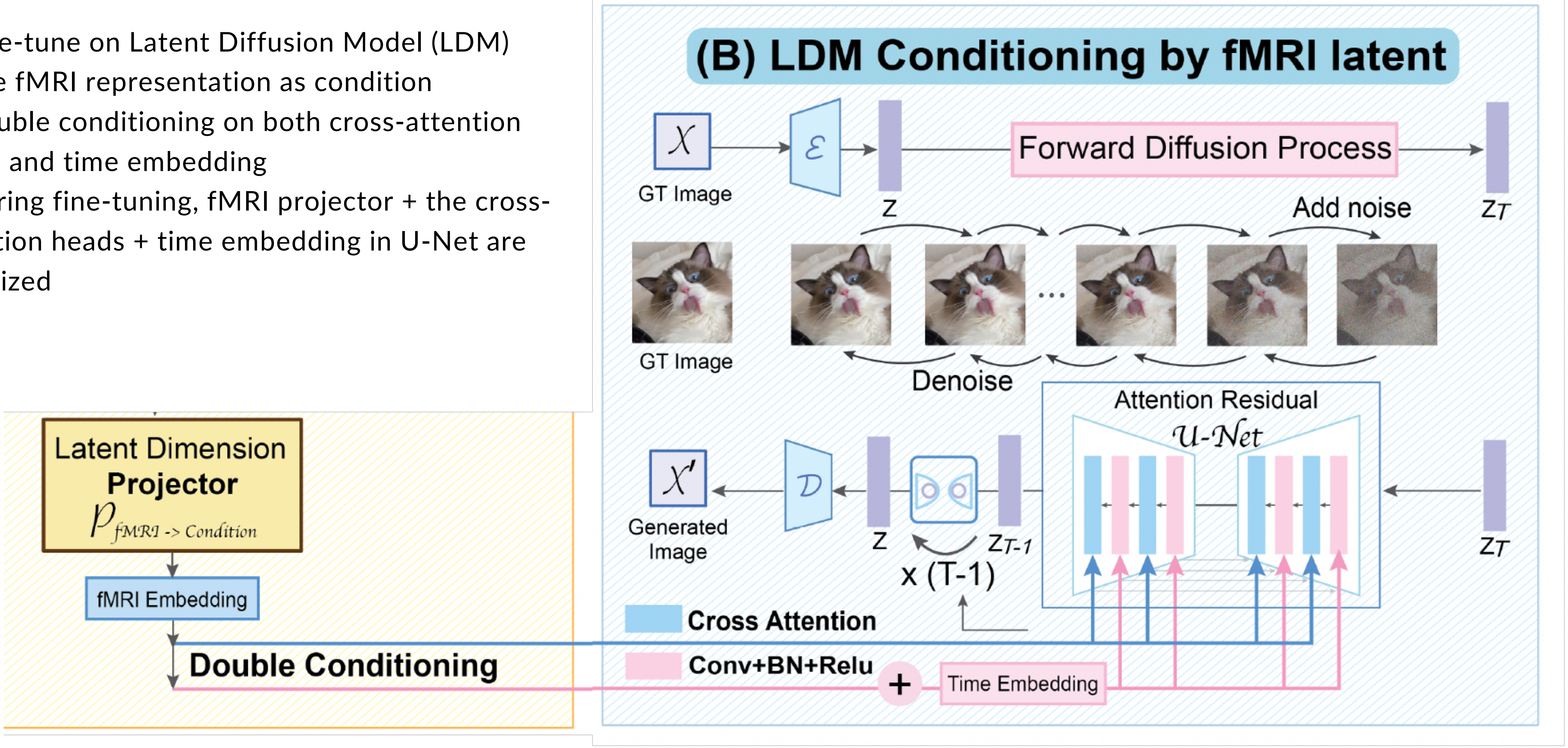
- Visual stimuli are sparsely encoded in the primary visual cortex, increasing information transmission efficiency and reducing redundancy
- Sparse coding is an efficient way for vision encoding, both in the brain and in computer vision
- In SC-MBM, fMRI data are divided into patches
- Each patch is encoded into a high-dimensional vector space with a size much larger than the original data space
  - i.e. large embedding-to-patch-size ratio
  - for fMRI: 1024/16 = 64
  - for image: 1024/(16*16*3) = 1.333 or 768/(14*14*3) = 1.3, depending on the architecture



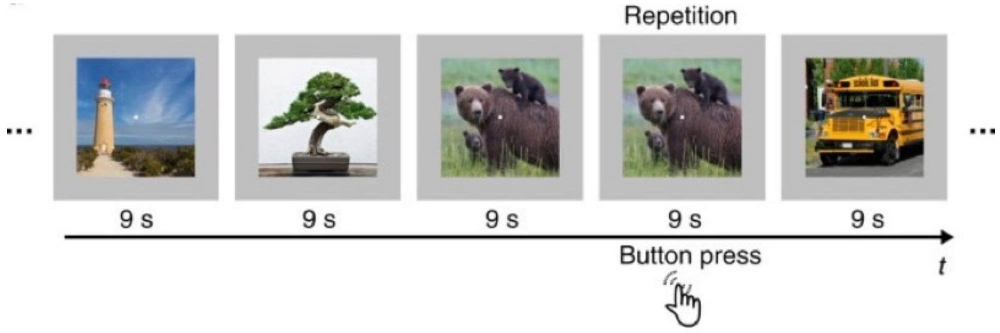
Stage B: Conditional Latent Diffusion Model

on GOD+BOLD5000 dataset (paired {fMRI, image})

1. Fine-tune on Latent Diffusion Model (LDM)
2. Use fMRI representation as condition
3. Double conditioning on both cross-attention heads and time embedding
4. During fine-tuning, fMRI projector + the cross-attention heads + time embedding in U-Net are optimized



fMRI Data Collection



Dataset #1 Generic of Decoding

- Training: {Image, fMRI} pair * 1200
- Testing: {Image, fMRI} pair * 50

- Image: Natural Image from ImageNet
- fMRI: fMRI scan from 5 participants
  - Selected voxels from visual cortex
- Training set and testing set don‘t have overlapping category

(T Horikawa, 2017 Nat Comm)

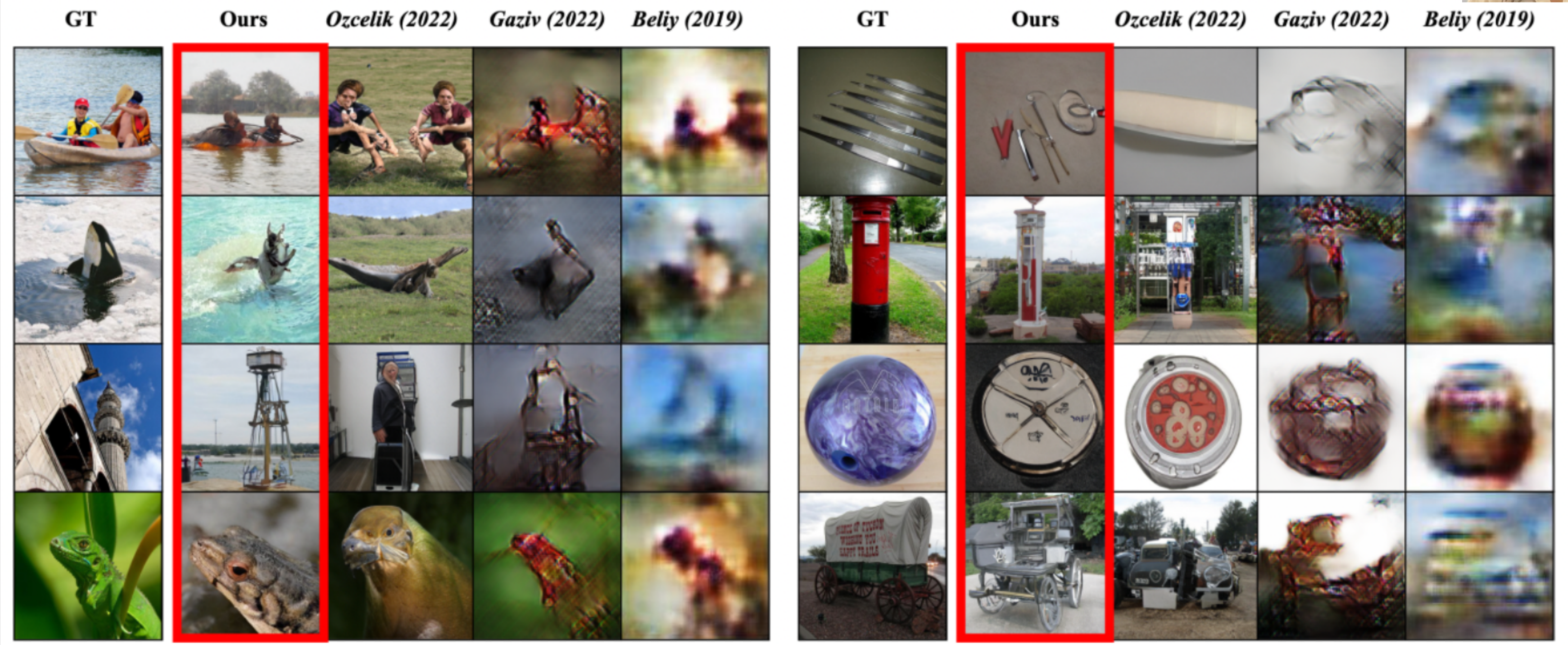
Dataset #2 BOLD5000

- Training: {Image, fMRI} pair * 4916
- Testing: {Image, fMRI} pair * 113

- Image: Natural Image from ImageNet, SUN dataset, COCO dataset
- fMRI: fMRI scan from 4 participants
  - Selected voxels from visual cortex
- Training set and testing set have some overlapping categories

(N Chang, 2019 Scientific Data)

Results – Compare with Benchmarks



- Ozcelik is GAN-based method
- Gaziv and Beliy are autoencoder-based methods

Result - Generation Consistency

Higher consistency - model reliability (as diffusion model is a probablistic model)

Figure 7. **Generation Consistency of MinD-Vis.** Images generated by our method were consistent across different samplings trials, sharing similar low-level features and semantics.

Result - Replication Dataset

Figure 8. **Replication Dataset (BOLD5000).** It achieved similar quantitative results as the GOD dataset. 50-way top-1 identification accuracy: 34%; FID: 1.2 (Subject 1).

Result - Extra Feature Decoded

Pros or Cons?

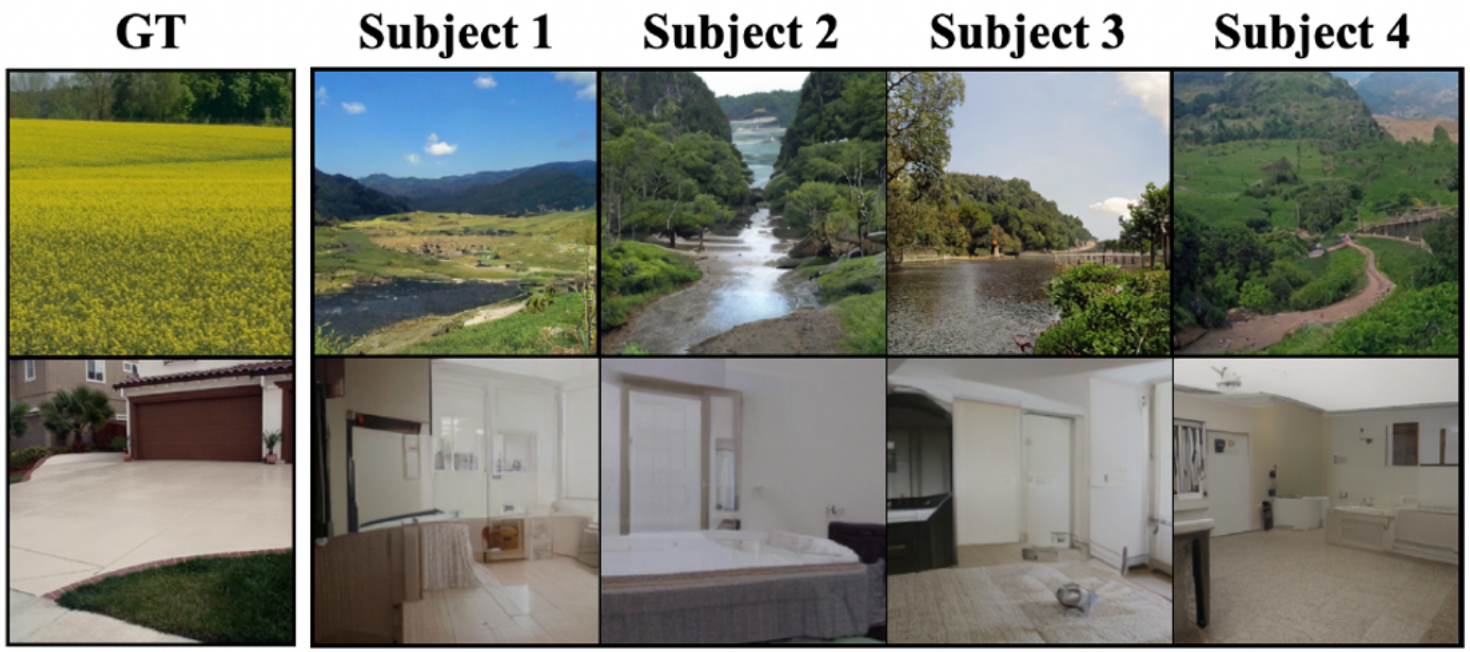
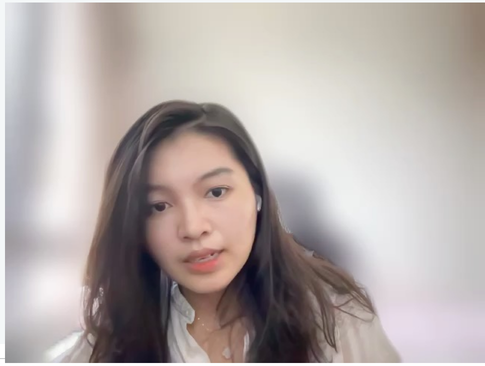


Figure 9. **Extra Features Decoded.** Imagery-related details can be decoded with our method. *e.g.* the river and blue sky were decoded with natural scenery stimulus (top row); similar interior decorating of indoor environments was decoded when a house was presented (bottom row).

Failure Cases

Possible reasons?



- Stimuli-unrelated thoughts
- These feature not common in the training set

  -> harder to decode
- Example: sock & sheep

  Animals are more common than clothings in the training set

  A semantic like “furry” ismore likely to be decoded as animals rather than clothes

Limitation

MinD-Vis

- Lacks of strong pixel-level guidance
- No interpretation of the features learned by SC-MBM
- The generation variance is larger than deterministic models

General decoding field

- Focus on individual-level decoding
- Focus on task specific region only (e.g. visual cortex)

THANKS
FOR
LISTENING