# Reconstructing perceived visual stimuli from brain fMRI signals using diffusion models

**Gabor Ioana**
**Babeș-Bolyai University**

**WeADL 2025 Workshop**

The workshop is organized under the umbrella of WinDMiL, project funded by CCCDI-UEFISCDI, project number PN-IV-P7-7.1-PED-2024-0121, within PNCDI IV

# Outline

- Task definition
- Classifications
- NSD Dataset
- Related work
- Research questions
- Our approach
- Methodology
- Results & Comparisons
- Future directions
- Conclusion

# Task Definition

- **Visual**: Viewed images
- **Brain**: Neural signals (e.g fMRI, MEG, EEG)
- **Decoding**: Classify/Retrieve/Reconstruct the visual stimulus

# General intuition

- **Extract** features from the image
- **Map** the brain information to the features
- **Predict** the features
- **Reconstruct** the image based on the predicted features

# Classifications

- Brain imaging
  - **fMRI**: High spatial resolution, slow temporal resolution
  - **EEG/MEG**: High temporal, low spatial resolution
  - fMRI preferred for decoding due to spatial specificity
- Image generation
  - **G**enerative **A**dversarial **N**etworks (Mindreader, NeurIPS 2022[4])
  - **L**atent **D**iffusion **M**odels
    - Stable Diffusion
    - Versatile Diffusion

# Classifications

- Number of subjects
  - Single subject
  - Multiple subjects
  - *Cross-Subject transfer of single subject models
- Brain mapping pipeline
  - Based on number of stages
    - Single-stage
    - Multi-stage
  - Commonly used models
    - Ridge regression
    - Multi Layer Perceptrons
    - Transformers

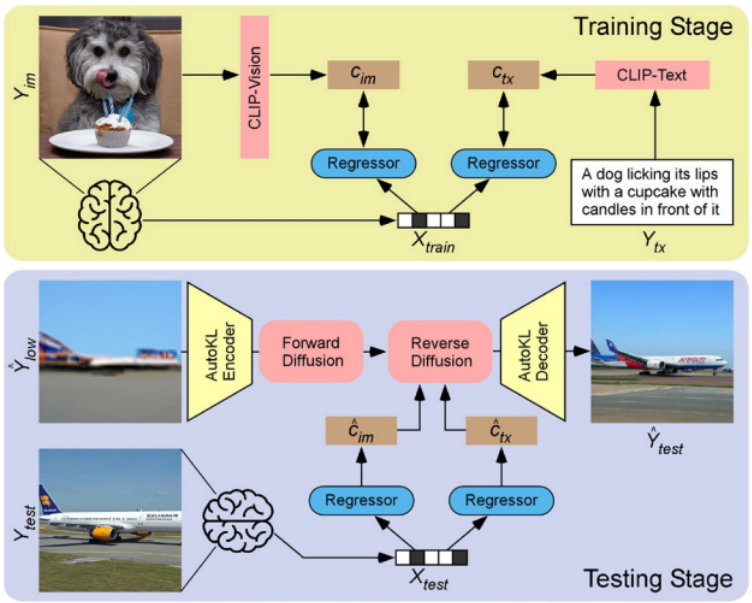
# Dataset: Natural Scenes Dataset (NSD), 2022 [1]

- Large-scale, high-resolution 7-Tesla fMRI dataset
- 8859 training, 982 test
- Microsoft Common Objects in Context [5]images
- 8 subjects, only 4 commonly studied
- Preprocessing: GLMDenoise[3], nsdgeneral ROI, flattened array of voxels

| **Subject ID** | **Voxel Vector Length** |
|---|---|
| 01 | 15724 |
| 02 | 14278 |
| 05 | 13039 |
| 07 | 12682 |

Table: 1D voxel vector lengths for each subject

# Related work




BrainDiffuser diagram

- **BrainDiffuser, 2023, Nature[6]**: Two-stage ridge regression, first VDVAE stage, finalized with Versatile Diffusion
- **BrainGuard, 2025, AAAI[7]**: Federated learning approach, emphasizes privacy and collaborative learning
- **MindFormer, 2024[2]**: IP Adapter & ViT-inspired architecture
- Trade-off: Model complexity vs decoding performance

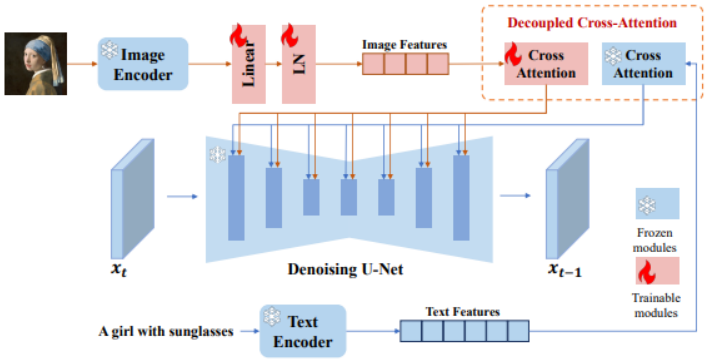
# Research questions

- How can we improve the scalability and computational efficiency of the approach?
- Do current methods extract redundant or non-informative feature embeddings?

# Our approach

- In the literature, CLIP embeddings are commonly used, typically 257 visual patched embeddings and 77 textual embeddings. We investigate the performance when decoding with only 16 target embeddings, obtained from a pretrained IPAdapterPlus[8] model. This approach has been first used by Mindformer [2], but has not yet been explored for other model architectures.
- We perform experiments both on single-subject ridge regression models and on a multi-subject federated learning framework (adapted from BrainGuard).

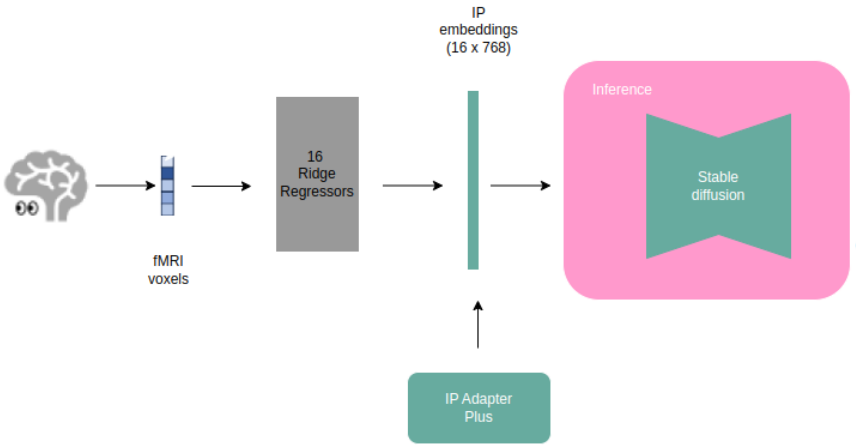
# Methodology: IP Adapter




IP Adapter architectural diagram [8]

- Enables image-based prompting for text-to-image diffusion models
- The authors released multiple variations: IPAdapter, IPAdapterPlus, IPAdapterFull, IPAdapterPlusXL
- Our experiments use IPAdapterPlus, which offers 16x768 positive embeddings, having a good tradeoff between quality and size

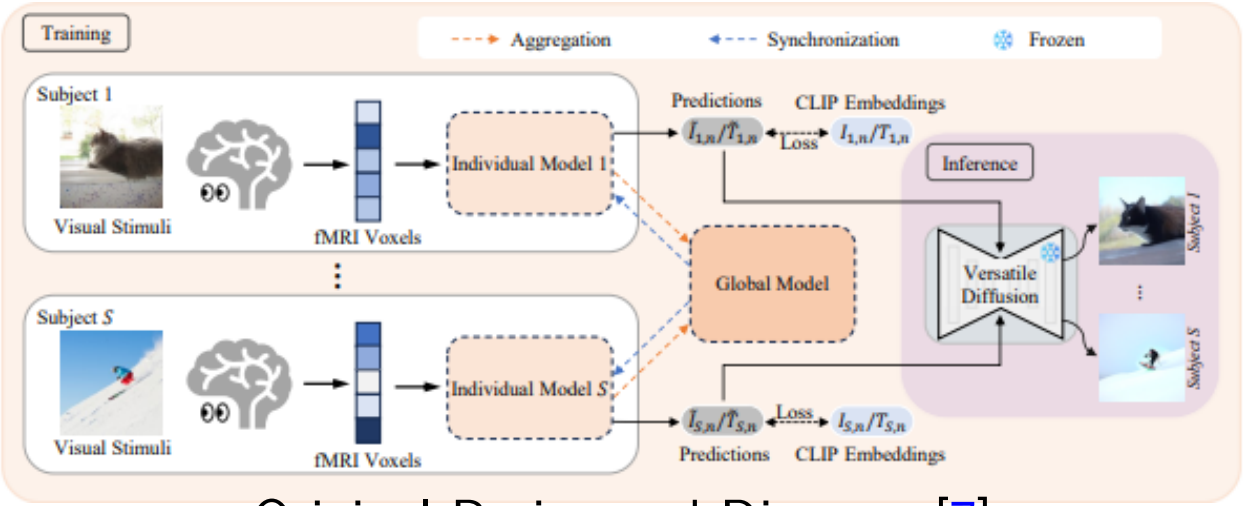
# Methodology : Ridge regression framework architecture



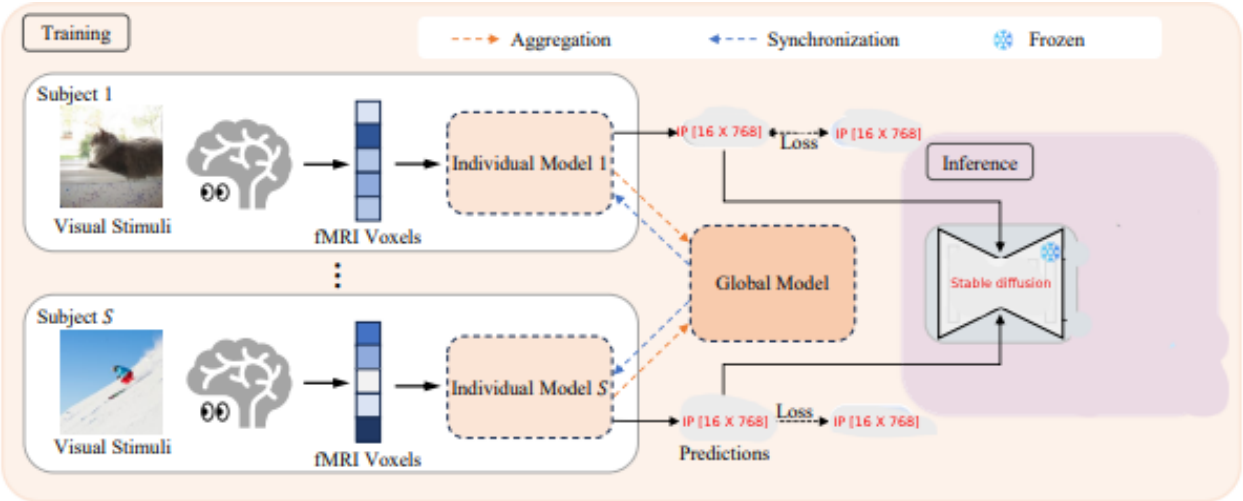

Simple Ridge Regression Framework

- One stage pipeline
- 16 Ridge Regressors

# Methodology : BrainGuard-inspired framework architecture




Original Brainguard Diagram [7]




Adapted Brainguard Diagram

# Methodology : BrainGuard-inspired framework architecture

Synchronization strategy, as described in the original paper

- Retention for foundational layers
- Global alignment for intermediate layers
- Adaptive tuning for advanced layers
  - **Dynamic Fusion Learner [9]** Module
  - used on the advanced *m* layers of the module
  - trained dynamically, based on the similarity between the weights of the advanced layers of the local and the global models

# Methodology : BrainGuard-inspired framework architecture

Network architecture, for both clients and global:

- **Input:** High-dimensional vector $x \in \mathbb{R}^{\text{in\_dim}}$
- **MLP Stack:** $n$ residual blocks of

$$\texttt{Linear} \rightarrow \texttt{GELU} \rightarrow \texttt{Norm} \rightarrow \texttt{Dropout}$$

- **Flatten:** Reshape to $[B, h]$
- **Head:** $\texttt{Linear}(h \rightarrow \text{latent\_size}) \Rightarrow$ image embedding
- **Projector:** Deep nonlinear MLP:

$$\begin{aligned} &\texttt{LN} \rightarrow \texttt{GELU} \rightarrow \texttt{Linear}(768 \rightarrow h) \\ &\rightarrow \texttt{LN} \rightarrow \texttt{GELU} \rightarrow \texttt{Linear}(h \rightarrow h) \\ &\rightarrow \texttt{LN} \rightarrow \texttt{GELU} \rightarrow \texttt{Linear}(h \rightarrow 768) \end{aligned}$$

- **Output:** Projected embedding $\in \mathbb{R}^{\text{latent\_size}}$

# Results - ridge regression

| Subject | PixCorr↑ | SSIM↑ | Alex (2)↑ | Alex (5)↑ | Incept.↑ | CLIP↑ | EffNet-B↓ | SwAV↓ |
|---|---|---|---|---|---|---|---|---|
| subj1 | 0.244 | 0.303 | 93.3% | 97.9% | 92.9% | 91.7% | 0.733 | 0.396 |
| subj2 | 0.242 | 0.331 | 90.5% | 96.4% | 90.0% | 86.1% | 0.770 | 0.427 |
| subj5 | 0.221 | 0.328 | 89.7% | 96.4% | 91.2% | 87.9% | 0.757 | 0.421 |
| subj7 | 0.218 | 0.327 | 87.9% | 94.5% | 88.9% | 85.5% | 0.789 | 0.444 |
| average | 0.231 | 0.322 | 90.35% | 96.3% | 90.75% | 87.8% | 0.744 | 0.422 |

Table: Quantitative results for individual subjects using the 8 evaluation metrics.

# Results - BrainGuard adaptation

| No. local iterations | PixCorr↑ | SSIM↑ | Alex (2)↑ | Alex (5)↑ | Incept.↑ | CLIP↑ | EffNet-B↓ | SwAV↓ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.263 | 0.335 | 0.951 | 0.985 | 0.959 | 0.939 | 0.611 | 0.334 |
| 5 | 0.256 | 0.330 | 0.949 | 0.983 | 0.957 | 0.937 | 0.614 | 0.336 |
| 10 | 0.254 | 0.335 | 0.948 | 0.982 | 0.955 | 0.939 | 0.618 | 0.335 |
| 20 | 0.252 | 0.331 | 0.946 | 0.983 | 0.956 | 0.937 | 0.618 | 0.339 |

Table: Quantitative results based on the frequency of synchronization

# Results - BrainGuard adaptation

| Number of layers | PixCorr↑ | SSIM↑ | Alex (2)↑ | Alex (5)↑ | Incept.↑ | CLIP↑ | EffNet-B↓ | SwAV↓ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.251 | 0.330 | 0.947 | 0.982 | 0.954 | 0.942 | 0.616 | 0.339 |
| 2 | 0.262 | 0.332 | 0.949 | 0.983 | 0.952 | 0.939 | 0.622 | 0.341 |
| 3 | 0.247 | 0.328 | 0.943 | 0.982 | 0.954 | 0.942 | 0.612 | 0.337 |
| 4 | 0.248 | 0.330 | 0.945 | 0.983 | 0.958 | 0.951 | 0.608 | 0.335 |
| 5 | 0.249 | 0.331 | 0.948 | 0.982 | 0.957 | 0.942 | 0.613 | 0.338 |

Table: Quantitative results based on the number of layers synchronized

# Results - BrainGuard adaptation

| Size of hidden layer | PixCorr↑ | SSIM↑ | Alex (2)↑ | Alex (5)↑ | Incept.↑ | CLIP↑ | EffNet-B↓ | SwAV↓ |
|---|---|---|---|---|---|---|---|---|
| 64 | 0.091 | 0.248 | 0.759 | 0.854 | 0.798 | 0.802 | 0.837 | 0.489 |
| 128 | 0.134 | 0.293 | 0.839 | 0.927 | 0.883 | 0.884 | 0.741 | 0.426 |
| 256 | 0.188 | 0.320 | 0.891 | 0.956 | 0.918 | 0.918 | 0.683 | 0.389 |
| 2048 | 0.263 | 0.334 | 0.951 | 0.985 | 0.959 | 0.937 | 0.611 | 0.334 |

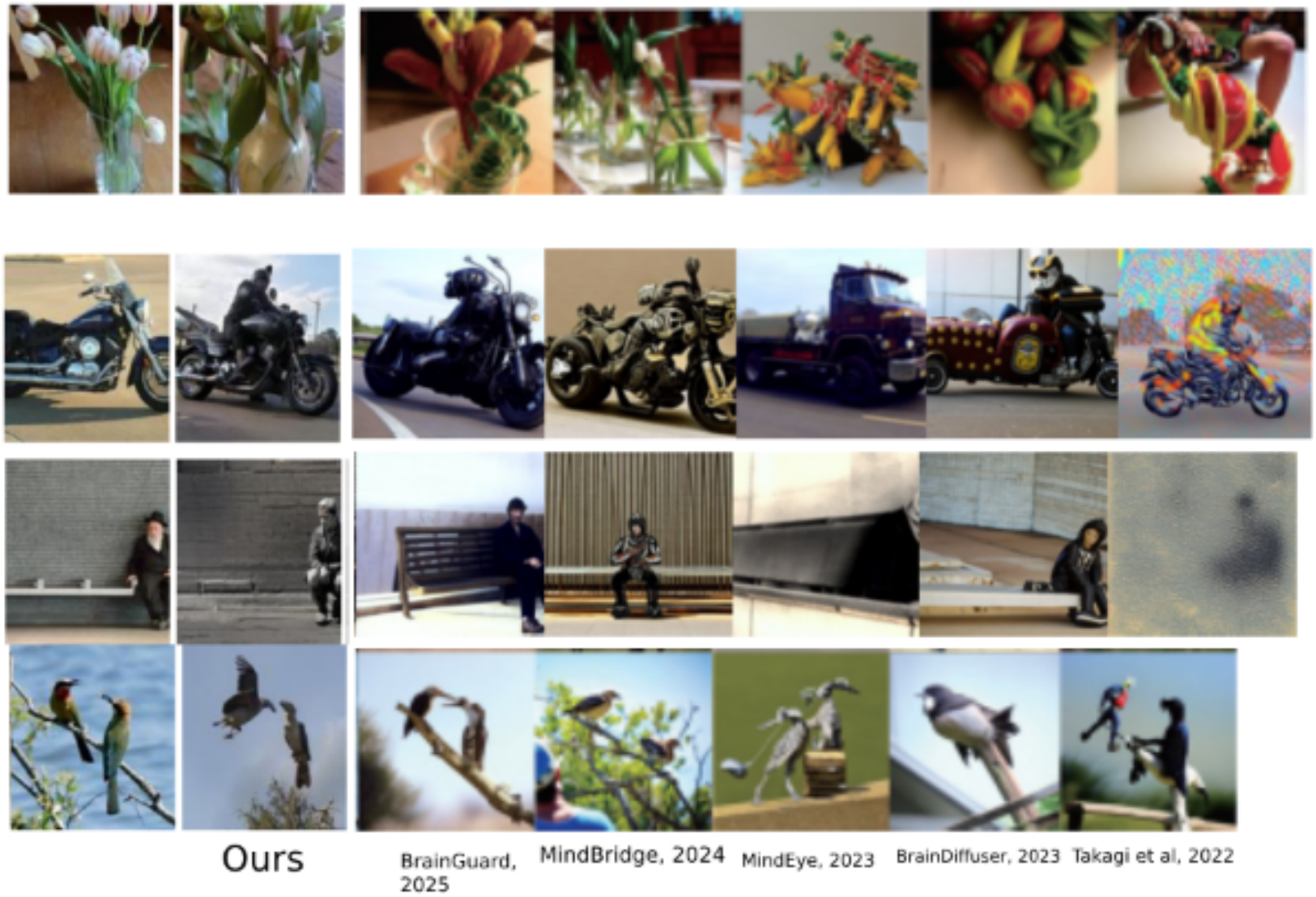
Table: Quantitative results based on the hidden layer size

# Results - Comparison

| Methods | PixCorr↑ | SSIM↑ | Alex (2)↑ | Alex (5)↑ | Incept.↑ | CLIP↑ | EffNet-B↓ | SwAV↓ |
|---|---|---|---|---|---|---|---|---|
| Mind-Reader, NeurIPS 2022 | | | | | 78.2% | | | |
| Mind-Vis, CVPR 2023 | 0.080 | 0.220 | 72.1% | 83.2% | 78.8% | 76.2% | 0.854 | 0.491 |
| Takagi et al. CVPR 2023 | | | 83.0% | 83.0% | 76.0% | 77.0% | | |
| Gu et al., MIDL 2023 | 0.150 | 0.325 | | | | | 0.862 | 0.465 |
| Brain-Diffuser, Nature 2023 | 0.254 | 0.356 | 94.2% | 96.2% | 87.2% | 91.5% | 0.775 | 0.423 |
| MindEye, NeurIPS 2023 | 0.309 | 0.323 | 94.7% | 97.8% | 93.8% | 94.1% | 0.645 | 0.367 |
| MindBridge, CVPR 2024 | 0.148 | 0.259 | 86.9% | 95.3% | 92.2% | 94.3% | 0.713 | 0.413 |
| MindFormer, arxiv 2024 | 0.243 | 0.345 | 93.5% | 97.6% | 94.4% | 94.4% | 0.648 | 0.350 |
| MindEye2, ICML 2024 | **0.322** | **0.431** | **96.1%** | **98.6%** | 95.4% | 93.0% | **0.619** | **0.333** |
| Psychometry, CVPR 2024 | 0.295 | 0.328 | 94.5% | 96.8% | 94.9% | 95.3% | 0.632 | 0.361 |
| BRAINGUARD, AAAI 2025 | 0.313 | 0.330 | 94.7% | 97.8% | **96.1%** | **96.4%** | 0.624 | 0.353 |
| ours (RR), 2025 | 0.231 | 0.322 | 90.35% | 96.3% | 90.75% | 87.8% | 0.744 | 0.422 |
| ours (BG), 2025 | 0.241 | 0.326 | 93.92% | 97.72% | 95.10% | 93.98% | **0.619** | 0.343 |

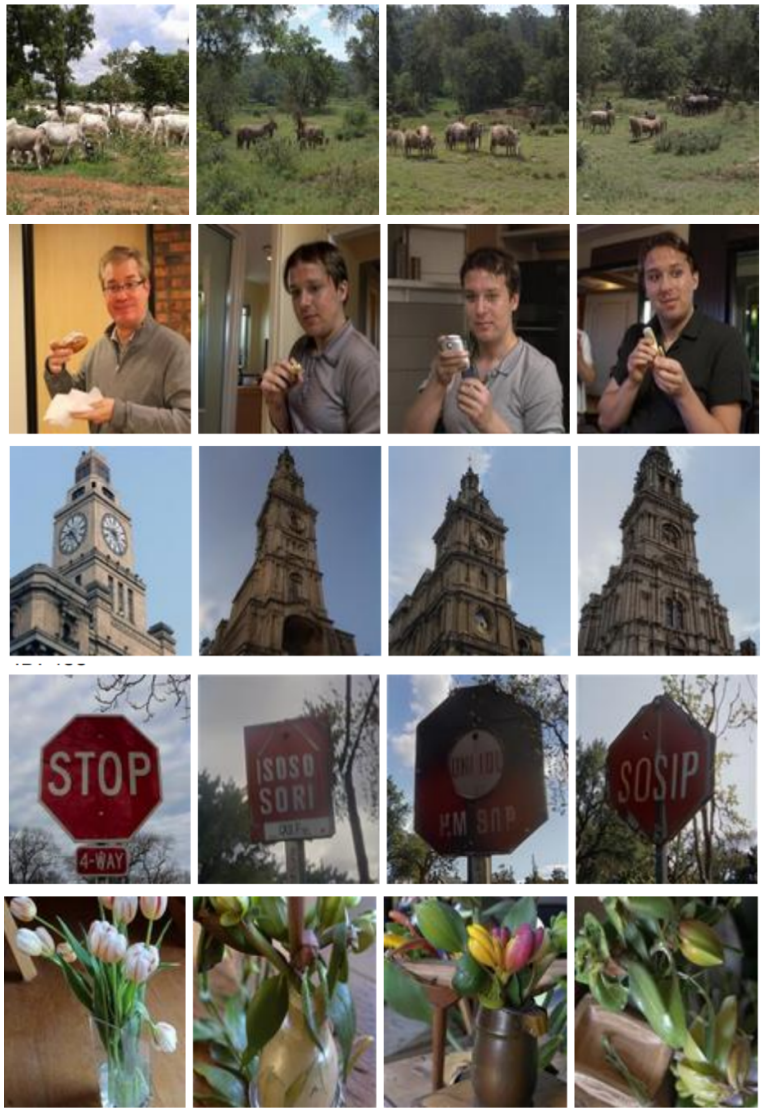
Table: Quantitative comparison results on NSD test dataset between multiple models

# Qualitative evaluation - Comparison




Comparison with other works

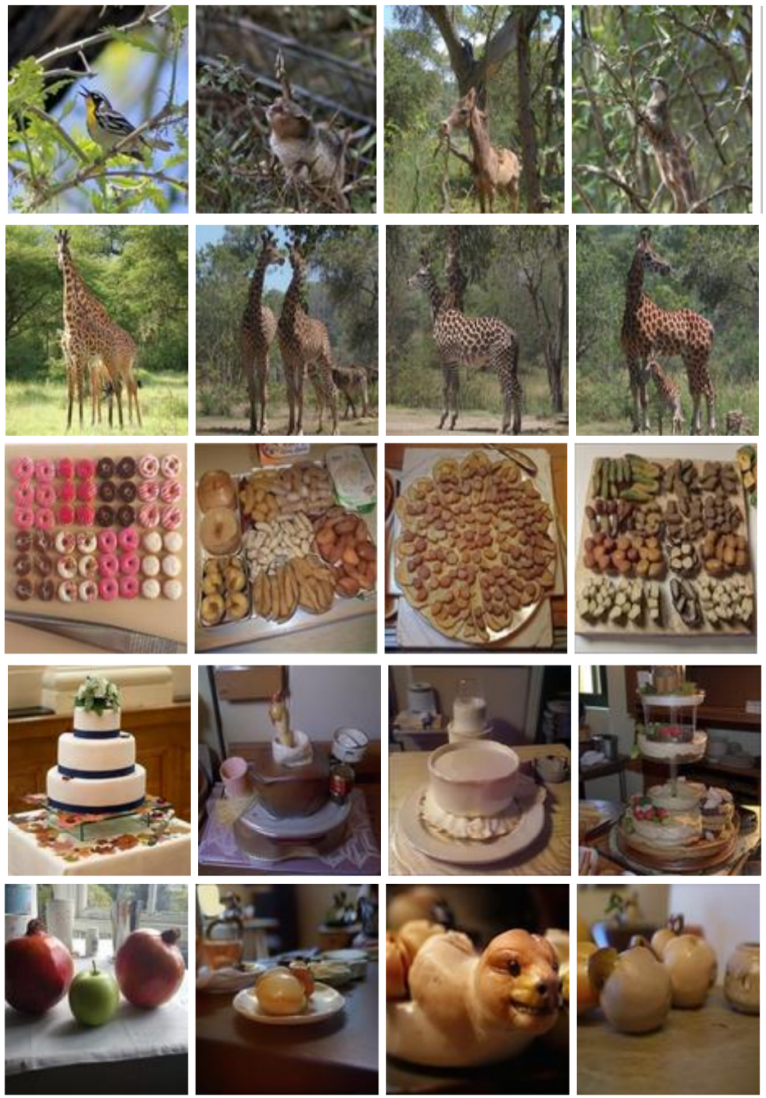
# Qualitative evaluation - Good pictures




Good Pictures

- Accurate reconstruction of colors
- Semantic meaning conserved

# Qualitative evaluation - Bad pictures



Bad Pictures

- Unrelated shapes : e.g. dog faces
- Irregularities regarding pictures with multiple small elements: e.g. donuts

# Future Directions

- Generalization across subjects & improved scalability
- Video decoding from brain signals
- Decoding imagined scenes

# Conclusion

- Presented efficient brain-to-image pipelines
- Incorporated IP Adapter into both:
  - a single subject framework
  - a cross subject federated learning framework
- Obtained competitive results with decreased model size

# References I


Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al.

A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence.

*Nature neuroscience*, 25(1):116–126, 2022.

Inhwa Han, Jaayeon Lee, and Jong Chul Ye.

Mindformer: Semantic alignment of multi-subject fmri for brain decoding.

*arXiv preprint arXiv:2405.17720*, 2024.

Kendrick N Kay, Ariel Rokem, Jonathan Winawer, Robert F Dougherty, and Brian A Wandell.

Glmdenoise: a fast, automated technique for denoising task-based fmri data.

*Frontiers in neuroscience*, 7:247, 2013.

# References II


Sikun Lin, Thomas Sprague, and Ambuj K Singh.

Mind reader: Reconstructing complex images from brain activities.

*Advances in Neural Information Processing Systems*, 35:29624–29636, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick.

Microsoft coco: Common objects in context.

In *Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

Furkan Ozcelik and Rufin VanRullen.

Natural scene reconstruction from fmri signals using generative latent diffusion.

*Scientific Reports*, 13(1):15666, 2023.

# References III


Zhibo Tian, Ruijie Quan, Fan Ma, Kun Zhan, and Yi Yang.

Brainguard: Privacy-preserving multisubject image reconstructions from brain activities.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14414–14422, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang.

Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models.

*arXiv preprint arXiv:2308.06721*, 2023.

Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan.

Fedala: Adaptive local aggregation for personalized federated learning.

In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11237–11244, 2023.