Human-Aligned Image Models Improve Visual Decoding from the Brain

Nona Rajabi[1], Antônio H. Ribeiro[2,*], Miguel Vasco[1,*], Farzaneh Taleb[1], Mårten Björkman[1], Danica Kragic[1]

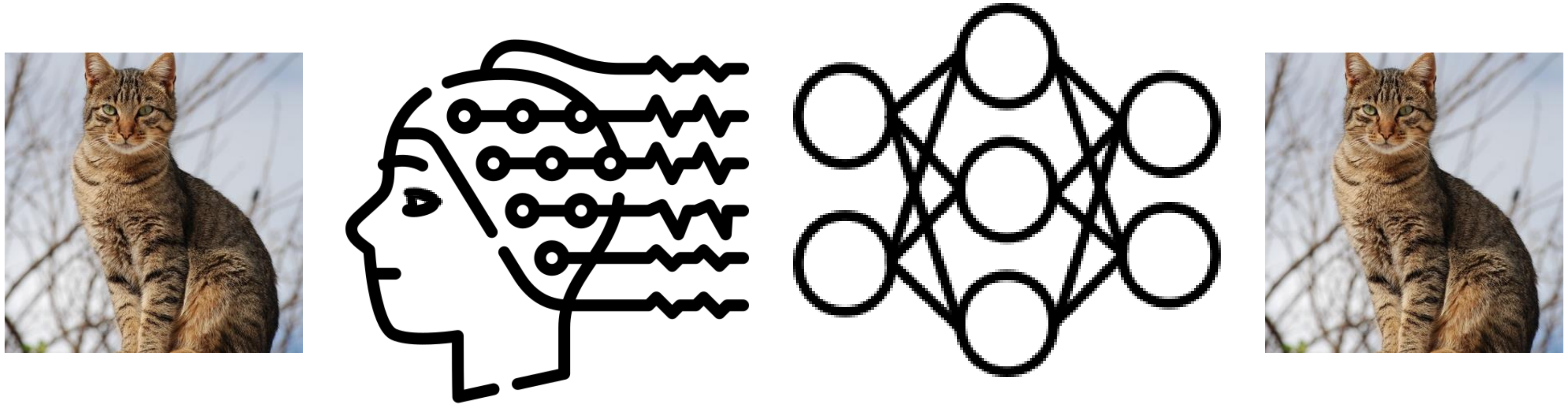
[*] Equal contribution

[1] KTH Royal Institute of Technology, Stockholm, Sweden

[2] Uppsala university, Uppsala, Sweden

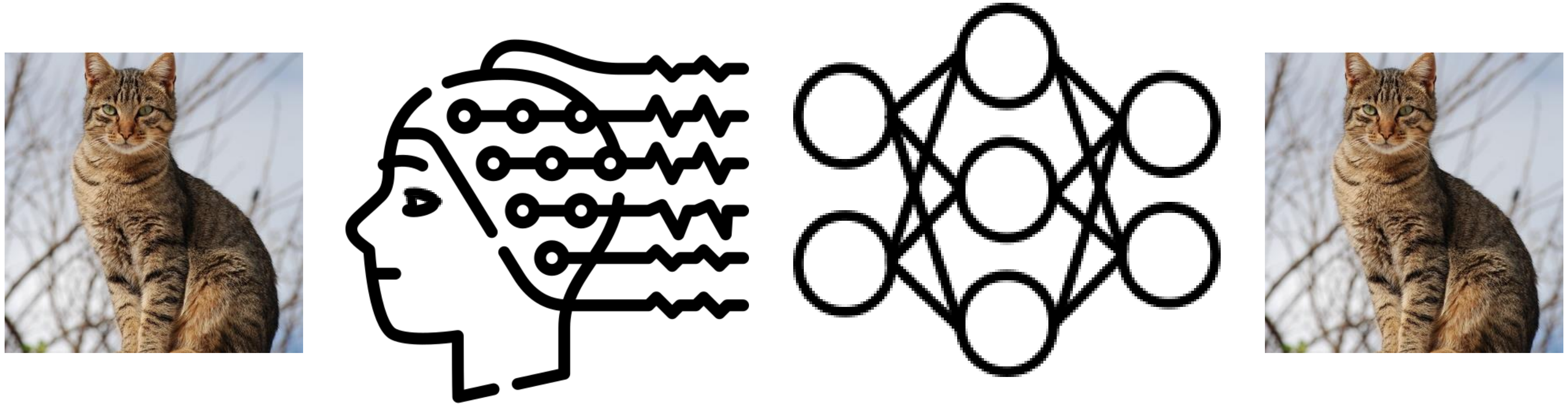
Visual Decoding from the Brain

Retrieve or reconstruct the observed or imaginged visual image from the corresponding brain activity



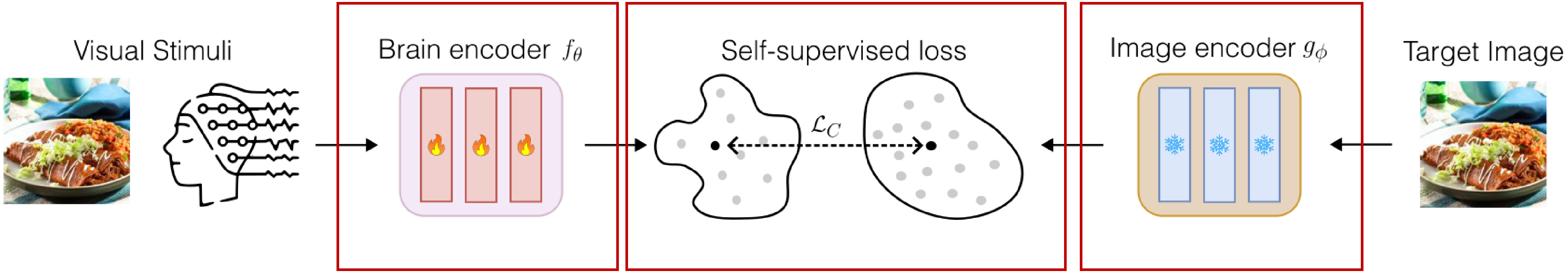
Visual Decoding from the Brain

Retrieve ~~or reconstruct~~ the observed ~~or imaginged~~ visual image from the corresponding brain activity



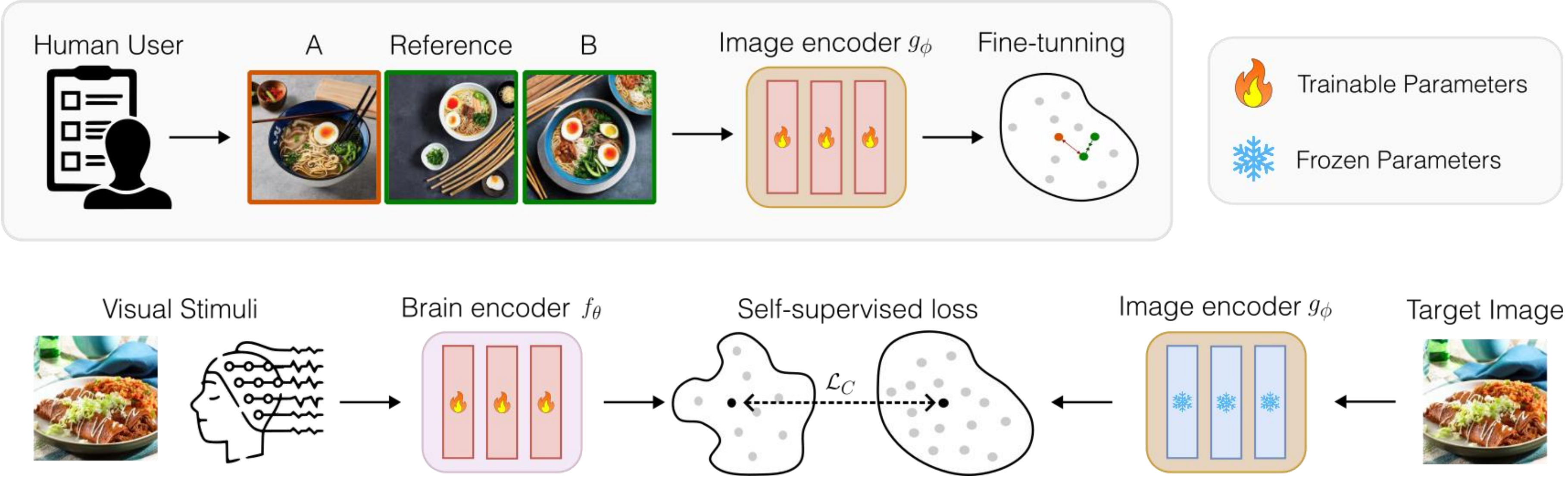
State-of-the-Art

- Current SOTA methods have three main components:
  - A brain-signal encoder: $x \rightarrow f_\theta(x) = v$
  - A pretrained image encoder: $b \rightarrow g_\theta = w$
  - A self-supervised loss function (InfoNCE loss)



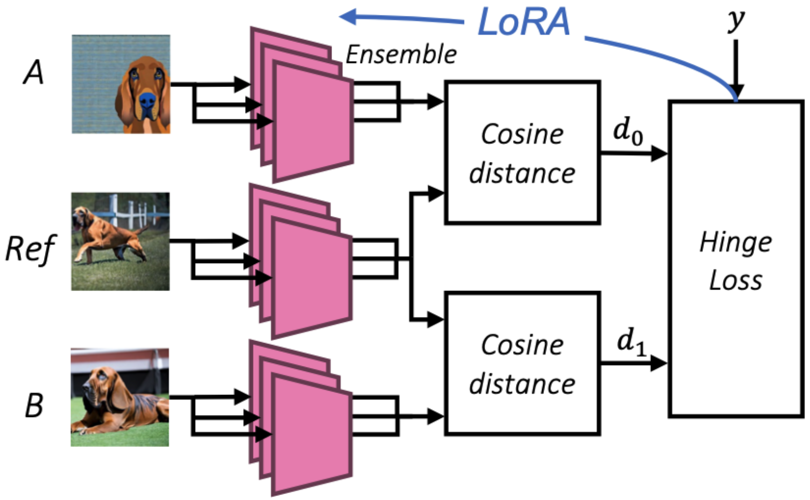
Proposed Method

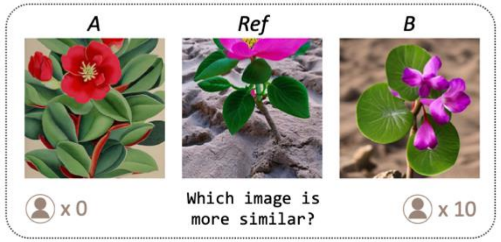
We propose to use human-aligned image representation models for visual decoding from brain signals



Human-Aligned Image Models

- Finetune image encoder models to better align with human perception [1, 2].



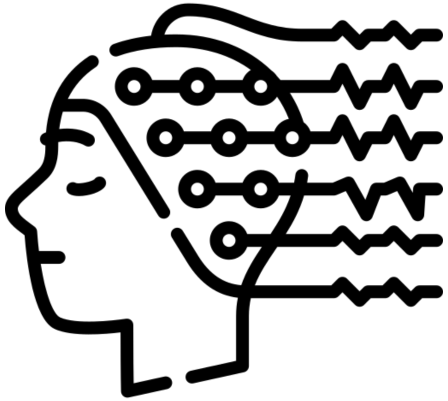


2AFC Data Collection

[1] Fu, Stephanie, et al. "DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data." *Advances in Neural Information Processing Systems* 36 (2023): 50742-50768.
[2] Sundaram, Shobhita, et al. "When does perceptual alignment benefit vision representations?." *Advances in Neural Information Processing Systems* 37 (2024): 55314-55341.

Datasets

- We tested our hypothesis using three brain-image paired datasets:
  - Things EEG2 (Gifford et al., 2022)
  - Things MEG (Hebart et al., 2023)
  - Natural Scene fMRI Dataset (Allan et al., 2022)

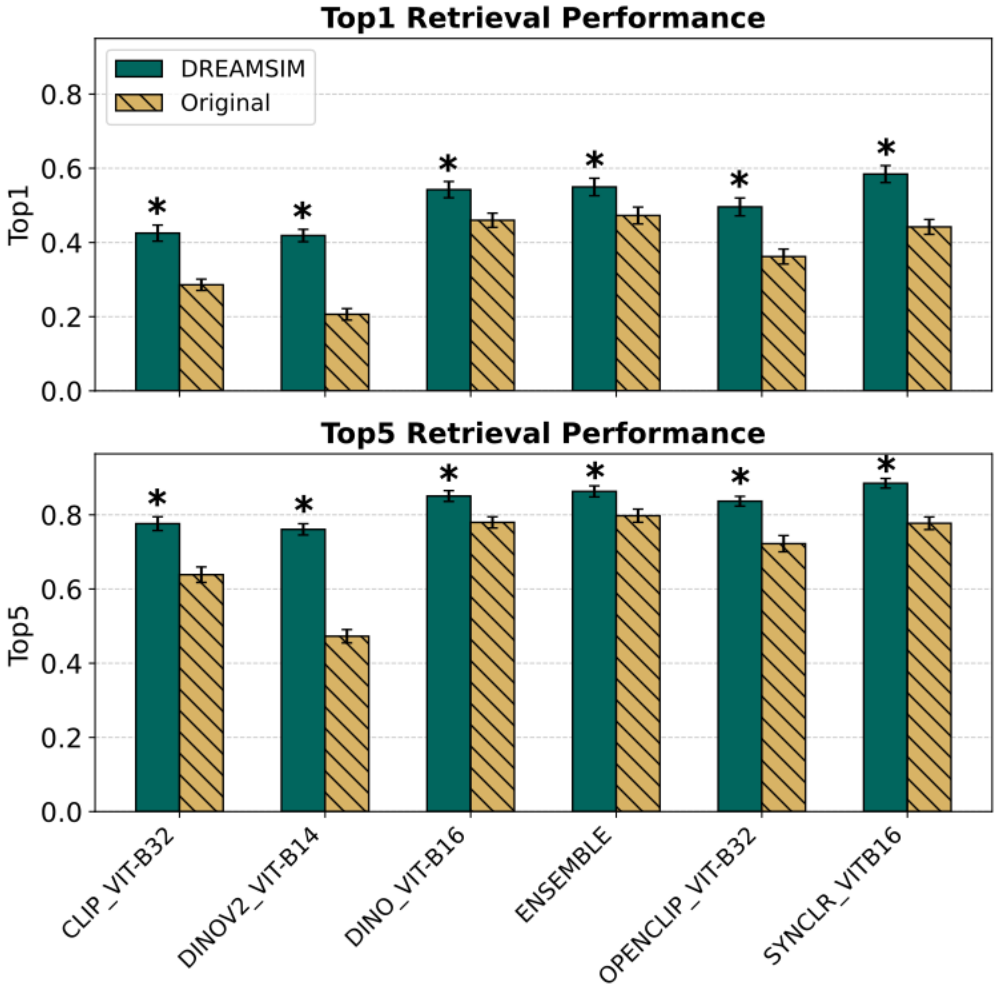


Sample Results (EEG)

- Trained the model with the multimodal InfoNCE loss (CLIP loss) (Radford et al., 2021)

$$\mathcal{L}_C = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(\mathrm{sim}(\mathbf{w}_i,\mathbf{v}_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\mathbf{w}_i,\mathbf{v}_j)/\tau)} + \log\frac{\exp(\mathrm{sim}(\mathbf{v}_i,\mathbf{w}_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\mathbf{v}_i,\mathbf{w}_j)/\tau)}\right]$$

- Evaluated the model's performance using top-1 and top-5 image retrieval accuracy from a 200 unseen image set

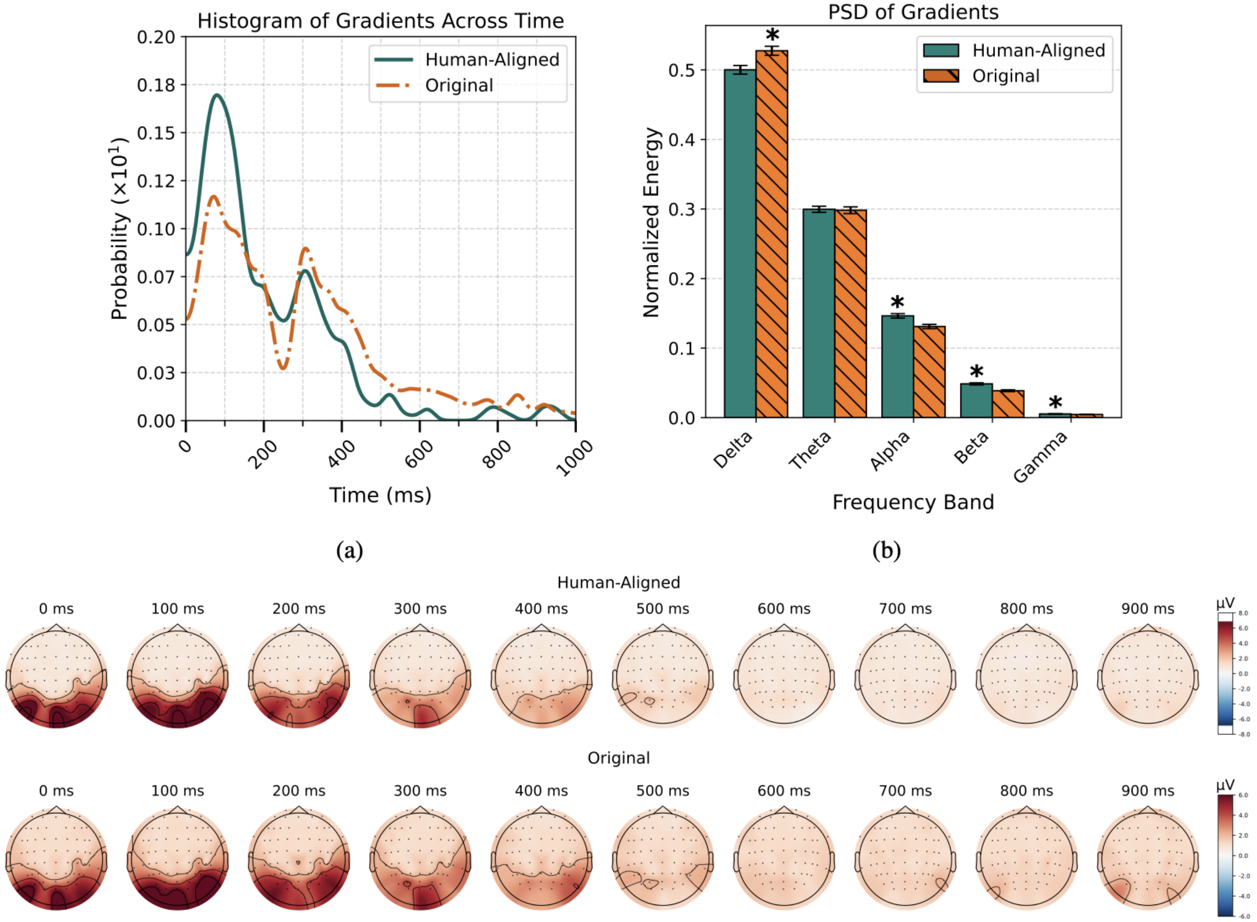


Biological Interpretation

a) Human-aligned models attend more to earlier timepoints.

b) Human-aligned models attend more to higher frequencies.

c) Both models attend to similar electrode locations on the brain



(a)

(b)

(c)

Summary

- Human-aligned models consistently and significantly improve visual retrieval from the brain.
- An extensive empirical study
- Interpreting the gradients of the model

Link to the full paper:

nonar@kth.se
www.linkedin.com/in/nonar