This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Image-mediated fMRI-to-caption generation with visual pathway tokens and hyperbolic alignment

Hyoungshin Choi[1,2], Youkyoung Lee[1], Bo-yong Park[2,3*], and Hyunjin Park[1,2*]

[1] Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, South Korea
[2] Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, South Korea
[3] Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea
* Bo-yong Park and Hyunjin Park are corresponding authors.
{gudtls17, hyunjinp}@skku.edu

Abstract. Generating natural language captions from fMRI remains challenging due to the large modality gap between brain activity and linguistic representations. We propose a two-stage brain-to-caption framework that leverages image embeddings as an intermediary between fMRI and language. In the first stage, a vision-language pipeline learns image-to-caption generation via LLaMA3. In the second stage, fMRI activations are aligned to the image embedding space through three novel components: (1) hyperbolic embedding alignment in the Poincaré ball providing more accurate and stable fMRI-to-image matching than Euclidean alternatives; (2) Token Probability Distribution Alignment (TPDA), which distills next-token distributions from the image-to-caption branch to the fMRI-to-caption branch; and (3) visual pathway tokenization, which converts fine-grained vertex-level visual cortical inputs into query tokens to supplement whole-brain regional features. On the Natural Scenes Dataset, our method achieves a CIDEr score of 56.79, outperforming existing methods. Beyond quantitative gains, neuroscientific analysis reveals the model's regional activations align with known functional brain specialization of visual pathway areas without explicit word type supervision, suggesting neurobiologically plausible brain-to-language mappings. Our code is at https://github.com/gudtls17/BrainToCaption.

Keywords: Neural decoding · fMRI · Language decoding · Hyperbolic embedding · Visual pathway · Teacher-student.

1 Introduction

The human brain transforms visual input into rich semantic representations expressible as natural language. Recent advances in neuroimaging and machine learning have enabled decoding visual and semantic content directly from brain activity [17,12,19,39,31,37,36]. In particular, functional magnetic resonance imaging (fMRI) has been used to reconstruct perceived images or predict high-level

semantic descriptions from neural signals. Despite this progress, generating accurate natural language captions from fMRI remains a significant challenge due to the large modality gap between brain signals and linguistic representations.

In the vision-language domain, large language models (LLMs) paired with visual encoders have shown remarkable success in generating image captions [26]. This suggests that once visual information is properly extracted, LLMs can handle the image-to-language mapping. Meanwhile, prior studies have shown that visual content can be directly reconstructed from fMRI signals by exploiting the hierarchical organization of the visual cortex [36,29,35], indicating a relatively small modality gap between brain activity and visual representations. In contrast, language decoding directly from fMRI typically recovers only coarse semantic content rather than exact linguistic forms [37], reflecting a larger gap between neural signals and language. These observations motivate a two-stage approach: first bridging the image-to-language gap and then utilizing the image embedding space as an intermediate facilitator to align fMRI representations.

We identify two key challenges in prior studies: **(1) alignment geometry-** aligning fMRI and image embeddings requires a distance metric that reflects the complex, non-uniform similarity structure of brain representations and semantic concepts. **(2) fine-grained semantic recovery-** fMRI signals are typically treated in predefined regional level (average over brain vertices) [7,8,9], lacking the detailed visual information needed to generate precise linguistic expressions. For the first challenge, we adopt hyperbolic embedding alignment using the Poincaré ball model, whose geometry-aware distance penalizes misalignment more sharply than Euclidean alternatives [28,14,21]. It has been effective in modeling functional brain networks [27,2]. Regarding the second, we take two complementary approaches. For fMRI modeling, we introduce a visual pathway tokenization that separately processes vertices from visual pathway areas, providing fine-grained spatial information to supplement standard region-level whole-brain features. We also propose a knowledge distillation objective that transfers next-token distributions of image-to-caption branch to those of the fMRI-to-caption branch, further compensating for information lost in the brain-to-vision mapping. Given this, we propose a two-stage brain-to-caption framework with the following contributions:

1. We align fMRI and image embeddings in the hyperbolic space achieving more faithful representation matching than Euclidean alternatives.
2. We introduce Token Probability Distribution Alignment (TPDA), which distills next-token probability distributions from the image-to-caption branch to the fMRI-to-caption branch, improving caption quality.
3. We design a visual pathway tokenization that converts visual cortex vertices into query tokens incorporating fine-grained spatial information to supplement whole-brain regional decoding.
4. Our method outperforms existing baselines and highlights brain regions linked to language expression confirming interpretability.

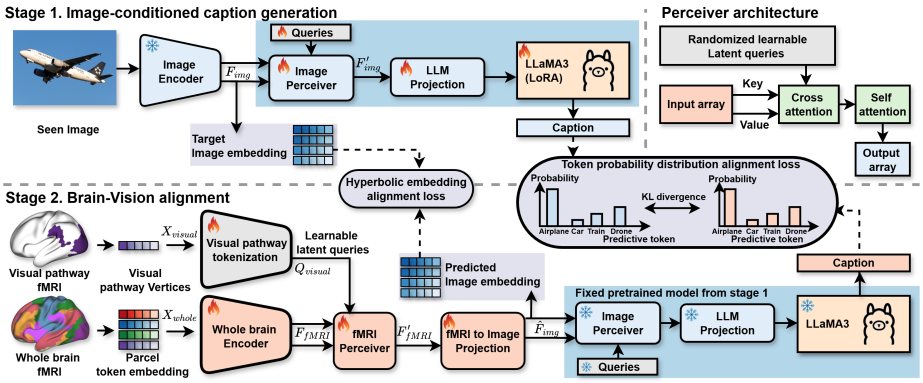

Fig. 1. Overview of the two-stage brain-to-caption framework. In stage1, a frozen image encoder extracts image embeddings, compressed by the image perceiver, and they are projected to LLaMA3 for caption generation. In stage2, whole-brain fMRI responses and visual pathway vertices are encoded separately and fused by the fMRI perceiver to predict image embeddings. Captions are generated from brain signals using these embeddings. The predicted image embeddings are aligned with target image embeddings (stage1) via hyperbolic embedding alignment loss, while TPDA loss aligns next-token distributions between the image- and fMRI-conditioned branches.

2 Method

2.1 Overview of the method

Our framework consists of two stages (**Fig. 1**). In stage1, an image viewed by the subject is passed through a frozen CLIP ViT encoder to extract image embeddings $F_{img} \in R^{N \times D_{img}}$, where N is the number of patches and D_{img} is the size of image embedding [32]. An image perceiver then compresses these embeddings into K semantic tokens using randomized learnable queries and the resulting tokens are projected to the LLM input space via an LLM projection layer. The projected tokens serve as prefix tokens for LLaMA3 [15], which is fine-tuned with low-rank adaptation (LoRA) to generate captions [18]. In stage2, the goal is to predict image embeddings from fMRI input so that the frozen stage1 pipeline can generate captions from brain signals. Whole-brain fMRI responses are parcellated into P regions of interest (ROIs) based on a given atlas and fed into a pretrained whole-brain encoder. Separately, vertex-level fMRI responses from visual pathway areas are transformed into learnable latent query tokens by the visual pathway tokenization. The fMRI perceiver then uses these learnable visual query tokens to attend over the whole-brain representations, producing predicted image embeddings $\hat{F}_{img} \in R^{N \times D_{img}}$ via an fMRI-to-image projection. The predicted embeddings are aligned with the target image embeddings (from stage1) through hyperbolic embedding alignment loss, while a TPDA loss ensures consistent caption predictions between the image-conditioned and fMRI-conditioned

branches. Finally, the predicted image embeddings are passed through the frozen image perceiver, LLM projection, and LLaMA3 to generate captions.

2.2 Constructing fMRI embeddings

We define fMRI input as an activation map (beta map) extracted from raw 4D timeseries data computed on the cortical surface [13]. For region-level whole-brain encoding, the cortical surface is parcellated into P ROIs based on the given atlas. We define the i-th parcel token as the collection of vertex-level activations within the i-th ROI. Since each ROI contains varying number of vertices, we apply adaptive average pooling to produce a fixed embedding dimension D_{fMRI}, yielding a parcel token embedding $X_{whole} \in R^{P \times D_{fMRI}}$. For fine-grained visual pathway encoding, we extract vertex-level activations from brain regions directly involved in visual information processing [16]. Early visual cortex, dorsal and ventral stream, fusiform face area, parahippocampal place area, and lateral occipital complex were selected, totaling V=12,535 vertices, yielding $X_{visual} \in R^{V}$.

2.3 Encoders, Perceivers, and LLaMA3

Image Encoder. We use a frozen CLIP ViT-L/14 as the image encoder [32]. We extract embeddings from the second-to-last layer to retain localized spatial properties [26], yielding target image embedding F_{img} in stage1.

Whole-brain Encoder. We adopt a surface vision transformer (SiT) [10], which applies self-attention across cortical parcels, as the whole-brain encoder. The encoder transforms parcel token embedding X_{whole} into fMRI embedding $F_{fMRI} \in R^{P \times D'_{fMRI}}$, where embedding dimension D_{fMRI} is changed to D'_{fMRI}.

Visual pathway tokenization. Visual pathway vertices X_{visual} are projected into a set of query tokens via an MLP. To ensure stable query initialization, the projected features are combined with learnable base queries Q_{base} through a residual connection using a learnable scaling factor α. This allows the model to rely on base queries during early training and progressively incorporate fine-grained visual pathway information as α increases.

$$Q_{visual} = Q_{base} + \alpha \cdot \mathrm{MLP}(X_{visual}), \quad Q_{visual} \in R^{N \times D'_{fMRI}} \tag{1}$$

This design compresses high-dimensional vertex features into N query tokens matching the image patch count and generates query tokens guiding cross-attention over whole-brain features. With these queries, visual pathway information selectively attends to relevant whole-brain regions in the fMRI perceiver.

Perceiver. Both perceivers adopt a cross-attention architecture, compressing input embeddings into a fixed number of tokens using latent queries [20]. This compression filters redundant information and distills semantically relevant content into compact prefix tokens suitable for the LLM. The image perceiver compresses image embedding F_{img} into K semantic tokens using randomly initialized learnable queries resulting in $F'_{img} \in R^{K \times D_{img}}$. In contrast, the fMRI perceiver uses visual query tokens Q_{visual} from Eq. (1), enabling visual pathway areas information to guide the selection of relevant whole-brain features resulting in $F'_{fMRI} \in R^{N \times D'_{fMRI}}$. The fMRI perceiver output is then projected to the image embedding dimension to produce predicted image embedding $\hat{F}_{img} \in R^{N \times D_{img}}$, which are used for alignment and caption generation. Similarly, the image perceiver output is projected to serve as prefix tokens for LLaMA3.

2.4 Loss function

Hyperbolic embedding alignment. Both predicted embeddings ($\hat{F}_{img}$) and target image embeddings (F_{img}) are projected to the tangent space via separate hyperbolic projection layers, then mapped onto the Poincaré ball with the exponential map [28,21,27,2]. The alignment loss minimizes the pairwise hyperbolic distance in the Poincaré ball with curvature c.

$$d_H(x, y) = \frac{1}{\sqrt{c}} \text{arcosh}(1 + \frac{2\|x - y\|^2}{(1 - c\|x\|^2)(1 - c\|y\|^2)}) \tag{2}$$

$$L_{align} = \frac{1}{N} \Sigma_{i=1}^{N} d_H(\hat{z}^i_{img}, z^i_{img}) \tag{3}$$

Where $\hat{z}^i_{img}$ and z^i_{img} are the hyperbolic embeddings of the i-th predicted and target image patch embeddings, respectively.

Token Probability Distribution Alignment. LLaMA3 generates captions autoregressively from the prefix tokens. In stage2, both the image-conditioned and fMRI-conditioned branches produce next-token probability distributions at each caption-generating step. TPDA aligns these distributions using KL divergence, with the image-conditioned branch serving as the teacher.

$$L_{TPDA} = KL(p(y_t|\hat{F}_{img}) \| p(y_t|F_{img})) \tag{4}$$

Where y_t denotes the predicted token at position t. The total training objective for stage2 is $L = \lambda_{align} L_{align} + \lambda_{TPDA} L_{TPDA}$.

3 Experiments

3.1 Datasets and experimental settings

We use the Natural Scene Dataset (NSD) [1], a 7T fMRI recordings of subjects viewing MS COCO images [24]. Following previous studies [17,12,19], we focus

Table 1. Caption prediction results. **Bold** indicates the best performance. Underline indicates the second best. C: CIDEr, B: BLEU, M: METEOR, R-L: ROUGE-L.

Model	C	B-1	B-2	B-3	B-4	M	R-L
SDRecon(CVPR23) [36]	13.83	36.21	17.11	7.22	3.43	10.03	25.13
OneLLM(CVPR24) [17]	22.99	47.04	26.97	15.49	9.51	13.55	35.05
BrainCap(Arxiv23) [12]	41.30	55.96	36.21	22.70	14.51	16.68	40.69
BrainChat(ICASSP25) [19]	26.10	52.30	29.20	17.10	10.70	14.30	**45.70**
UMBRAE(ECCV24) [39]	51.93	57.63	38.02	25.00	16.76	18.41	42.15
MindLLM(ICML25) [31]	43.04	58.05	37.95	24.40	16.14	16.62	42.03
Ours (K=8)	53.48	61.68	42.40	29.52	20.93	21.80	44.67
Ours (K=32)	50.86	60.97	41.96	29.24	20.78	21.55	44.25
Ours (K=16)	**56.79**	**62.37**	**44.15**	**31.26**	**22.73**	**22.30**	45.44

on the first subject, who completed all experimental trials. The dataset provides standardized splits containing 8,859 training images with 24,980 fMRI trials (up to three repetitions per image) and 982 test images with 2,770 fMRI trials. For images with multiple presentations, the corresponding fMRI activation maps were averaged. Brain regions were parcellated into 1,000 ROIs using the Schaefer atlas [34]. We use a frozen CLIP ViT-L/14 and a pretrained SiT as image and whole-brain encoders, with both image and fMRI perceivers compressing embeddings into K=16 semantic tokens. We set curvature c=0.1 in Poincaré ball. LLaMA3-8B is fine-tuned with LoRA (r=8). Stage1 and 2 are trained for 20 epochs with batch sizes 32 and 16, using λ_{align}=λ_{TPDA}=1. Embedding dimensions are D_{img}=1024, D_{fMRI}=64, and D'_{fMRI}=384. Caption evaluation uses standard CIDEr, BLEU, METEOR, and ROUGE-L metrics [38,30,3,25].

3.2 Experimental results

Comparison with existing methods. Table 1 presents the quantitative caption decoding results. Our model achieves the best performance across all metrics except ROUGE-L, outperforming the strongest baseline UMBRAE suggesting hyperbolic alignment and visual pathway tokenization provide effective fMRI-to-language bridging. We also evaluate the effect of the number of prefix tokens K. Using K=16 prefix tokens achieves the best balance between compressing redundant information and retaining sufficient semantic content.

Captioning results. Fig. 2 illustrates the impact of visual pathway tokenization, a major component of our method, on captions generated from whole-brain fMRI. Without it, the model captures the general scene context but misses specific objects and attributes (e.g., "broccoli" for "pizza", "airplanes" for "birds", missing the train color "red" and "yellow"). Conversely, tokenization recovers fine-grained visual details such as object identity, gender, and scene attributes, demonstrating that vertex-level visual pathway information is essential for precise caption generation and visual pathway perturbing is clinically relevant.

Fig. 2. Qualitative comparison of generated captions with and without visual pathway tokenization. Reds are incorrect and greens are correct predictions.

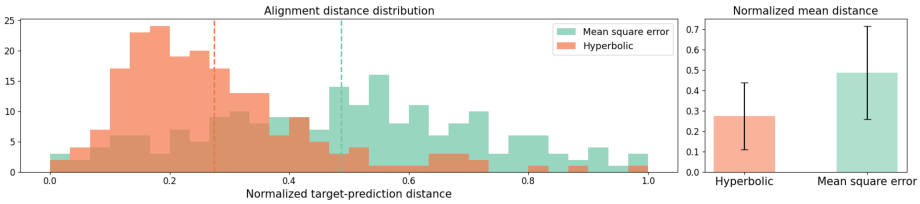

Fig. 3. Comparison of alignment methods. Distribution of distances between target and predicted image embeddings with and without applying hyperbolic alignment.

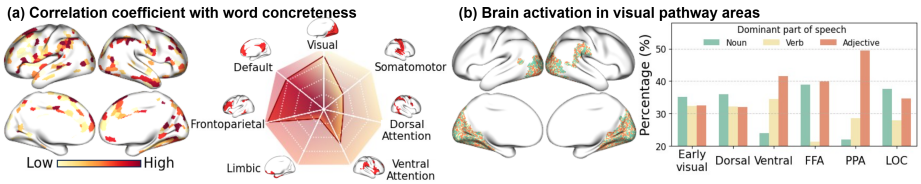

Fig. 4. Neuroscientific analysis of brain region attributions contributing to caption generation. (a) Correlation between whole-brain parcel activations and word concreteness summarized across functional communities. (b) Most activated visual pathway vertices (left) during word token prediction. Bar plot (right) shows the proportion of dominant parts of speech (noun, verb, adjective) predicted from these areas.

Interpretation of alignment strategy and brain region attribution. Fig. 3 shows that hyperbolic alignment reduces the normalized mean distance between target and predicted image embeddings by 43.6% compared to Euclidean mean squared error (MSE) with substantially smaller variance, indicating that alignment in Poincaré ball is not only closer but also more stable across samples. **Fig. 4(a)** shows that correlation with word concreteness is strongest in the visual, frontoparietal, and default mode networks, consistent with their estab-

Table 2. Ablation studies about loss and model architecture. **Bold** indicates the best performance. Underline indicates the second best. MSE: mean square error, CE: cross entropy, TPDA: Token Probability Distribution Alignment.

Alignment loss	Caption loss	C	B-1	B-2	B-3	B-4	M	R-L
MSE	CE	53.34	61.03	41.76	28.56	19.87	21.45	43.84
Hyperbolic	CE	54.84	61.96	42.81	29.67	20.94	21.77	44.32
MSE	TPDA	55.31	62.19	43.13	30.08	21.23	22.00	44.96
Ours (Hyperbolic + TPDA)		**56.79**	**62.37**	**44.15**	**31.26**	**22.73**	**22.30**	**45.44**
Method		**C**	**B-1**	**B-2**	**B-3**	**B-4**	**M**	**R-L**
w/o visual pathway tokenization		45.27	59.41	39.82	26.91	18.78	20.11	42.70
w/o whole-brain encoder		50.55	61.12	41.85	28.48	19.77	21.26	43.86
w/o stage1		50.22	61.90	42.61	29.29	20.49	20.69	43.97
w/o image prediction		52.84	61.26	42.23	29.23	20.66	21.66	44.43
CLIP ViT last layer		52.63	60.02	41.38	28.86	20.52	21.84	44.24
Ours (Second-to-last layer)		**56.79**	**62.37**	**44.15**	**31.26**	**22.73**	**22.30**	**45.44**

lished roles in visual processing and semantic cognition [4,23,33,6,5]. **Fig. 4(b)** shows the most activated visual pathway regions during word token prediction. Early visual and dorsal stream regions are more activated for nouns, while ventral stream and parahippocampal place area (PPA) are activated for adjectives, and fusiform face area (FFA) and lateral occipital complex (LOC) show strong activation for both nouns and adjectives. Notably, our model receives no explicit word type supervision, yet these activation patterns align with known regional functional specialization [6,11,22]. This suggests that our caption generation reflects the functional organization of the visual pathway, with distinct regions contributing to different parts-of-speech. These results are reproducible in other subjects as well (mean $r=0.478$, all false discovery rate $p<0.001$). This neurobiologically grounded mapping between brain regions and language components may inform brain-computer interfaces for enhanced communication.

Ablation study. Table 2 shows ablation results for loss functions and model architecture. In the upper section, replacing MSE with hyperbolic alignment slightly improves CIDEr and replacing cross-entropy with TPDA yields a larger gain. Combining both losses achieves the best performance, indicating the two objectives are complementary as hyperbolic alignment provides better embedding-level matching, while TPDA enforces consistency at the token prediction level. Removing visual pathway tokenization causes the largest performance drop, confirming fine-grained visual cortical information is essential for generating precise captions, while removing the whole-brain encoder also leads to a notable decline, suggesting global brain context complements the visual pathway signal (lower section). Without stage1 pretraining, establishing an image-to-language mapping proves critical for effective fMRI-to-caption transfer. Removing the image prediction objective and using the last CLIP ViT layer instead of the second-to-last layer embeddings results in moderate degradation, supporting our design choices of explicit image embedding prediction and localized feature extraction.

4 Conclusion

We presented a two-stage brain-to-caption framework that aligns fMRI with image embeddings for caption generation. Our contributions are fundamentally novel compared to existing method in that we introduce hyperbolic embedding alignment in the Poincaré ball, token probability distribution alignment for caption-level distillation, and visual pathway tokenization leveraging vertex-level visual cortical information as query tokens. Experiments on NSD demonstrate the best performance with CIDEr 56.79, with ablation studies confirming each model component's contribution. Neuroscientific analysis reveals that our framework's regional attributions align with known functional brain specialization, suggesting neurobiologically plausible brain-to-language mappings that may inform brain-computer interfaces for enhanced communication. Future work plans include task extension (eg, VQA or retrieval) using the image embedding.

Acknowledgments. This study was supported by SMC-SKKU Future Convergence Research Program (SMO125123), AI Graduate School Support Program (Sungkyunkwan University) (RS-2019-II190421), National Research Foundation (RS-2024-00408040/RS-2025-2025303), ICT Creative Consilience program (RS-2020-II201821), and the IITP program funded by the Korean Government (RS-2026-25528384/2022-0-00448/RS-2022-II220448).

Disclosure of Interests. The authors declare no competing interests relevant to the content of this article.

References

1. Allen, I., et al.: A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience **25**(1), 116–126 (2022)
2. Baker, D., et al.: Hyperbolic graph embedding of meg brain networks to study brain alterations in individuals with subjective cognitive decline. IEEE journal of biomedical and health informatics **28**(12), 7357–7368 (2024)
3. Banerjee, et al.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
4. Bi, A., et al.: Object domain and modality in the ventral visual pathway. Trends in cognitive sciences **20**(4), 282–290 (2016)
5. Brysbaert, M., Warriner, et al.: Concreteness ratings for 40 thousand generally known english word lemmas. Behavior research methods **46**(3), 904–911 (2014)
6. Cavina-Pratesi, et al.: Separate processing of texture and form in the ventral stream: Evidence from fmri and visual agnosia. Cerebral Cortex **20**(2), 433–446 (2009)
7. Choi, H., Kim, J., Chung, J., Park, B.y., Park, H.: Pmil: Prompt enhanced multimodal integrative analysis of fmri combining functional connectivity and temporal latency. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 543–553. Springer (2025)

8. Choi, H., Kim, S., Lee, J.e., Park, B.y., Park, H.: Integrating meta-analysis in multi-modal brain studies with graph-based attention transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–425. Springer (2025)
9. Choi, H., Park, Y., Lee, J.e., Kim, S., Park, B.y., Park, H.: Spatiotemporal characterization of the functional mri latency structure with respect to neural signaling and brain hierarchy. Advanced Science **12**(43), e04956 (2025)
10. Dahan, E.C., et al.: Surface vision transformers: Attention-based modelling applied to cortical analysis. In: Proceedings of The 5th International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 172, pp. 282–303. PMLR (06–08 Jul 2022)
11. Epstein, N., et al.: The parahippocampal place area: recognition, navigation, or encoding? Neuron **23**(1), 115–125 (1999)
12. Ferrante, M., et al.: Brain captioning: Decoding human brain activity into images and text (2023)
13. Fischl, A.M., et al.: High-resolution intersubject averaging and a coordinate system for the cortical surface. Human brain mapping **8**(4), 272–284 (1999)
14. Ganea, T., et al.: Hyperbolic neural networks. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
15. Grattafiori, et al.: The llama 3 herd of models. arXiv:2407.21783 (2024)
16. Grill-Spector, J., et al.: Developmental neuroimaging of the human ventral visual cortex. Trends in cognitive sciences **12**(4), 152–162 (2008)
17. Han, X., et al.: Onellm: One framework to align all modalities with language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26584–26595 (June 2024)
18. Hu, E.J., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
19. Huang, H., et al.: Brainchat: Interactive semantic information decoding from fmri using large-scale vision-language pretrained models. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
20. Jaegle, et al.: Perceiver: General perception with iterative attention. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 4651–4664. PMLR (2021)
21. Khrulkov, et al.: Hyperbolic image embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
22. Kourtzi, Z., Kanwisher, N.: Representation of perceived object shape by the human lateral occipital complex. Science **293**(5534), 1506–1509 (2001)
23. Lanzoni, L., Jefferies, E., et al.: The role of default mode network in semantic cue integration. NeuroImage **219**, 117019 (2020)
24. Lin, C.L., et al.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
25. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
26. Liu, Y.J., et al.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023)
27. Longhena, et al.: Hyperbolic embedding of brain networks detects regions disrupted by neurodegeneration in alzheimer's disease. Phys. Rev. E **111**, 044402 (2025)
28. Nickel, D., et al.: Poincaré embeddings for learning hierarchical representations. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
29. Nishimoto, J.L., et al.: Reconstructing visual experiences from brain activity evoked by natural movies. Current biology **21**(19), 1641–1646 (2011)

30. Papineni, W.J., et al.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
31. Qiu, W., et al.: MindLLM: A subject-agnostic and versatile model for fMRI-to-text decoding. In: Forty-second International Conference on Machine Learning (2025)
32. Radford, I., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
33. Roxbury, D.A., et al.: An fmri study of concreteness effects in spoken word recognition. Behavioral and Brain Functions **10**(1), 34 (2014)
34. Schaefer, B.T., et al.: Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. Cerebral cortex **28**(9), 3095–3114 (2018)
35. Shen, Y., et al.: Deep image reconstruction from human brain activity. PLOS Computational Biology **15**(1), 1–23 (01 2019)
36. Takagi, S., et al.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14453–14463 (June 2023)
37. Tang, A.G., et al.: Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience **26**(5), 858–866 (2023)
38. Vedantam, et al.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
39. Xia, J.H., et al.: Umbrae: Unified multimodal brain decoding. In: European Conference on Computer Vision. pp. 242–259. Springer (2024)