Multi-Expert Representation Learning with Dynamic Routing for Brain Captioning

Zhilin Guo[1,†], Weihao Xia[2,†(✉)], Mingdeng Cao[3], Chenliang Zhou[1], and Cengiz Oztireli[1]

[1] University of Cambridge, Cambridge, UK
[2] Harbin Institute of Technology (Shenzhen), Shenzhen, China
[3] University of Tokyo, Tokyo, Japan
† Z. G. and W. X. contributed equally to this work.
* Correspondence: weihaox@outlook.com

Abstract. Recent advances in vision-language models have accelerated multimodal brain decoding, particularly brain captioning, which aims to translate neural activity into natural language descriptions. However, most existing methods align brain signals with either textual embeddings or single-stream visual representations, overlooking the hierarchical and context-dependent organization of the human visual cortex. In this work, we propose a multi-expert routing framework with progressive alignment optimization for brain captioning. We introduce task-specialized vision experts that encode complementary perceptual attributes across multiple abstraction levels. To reflect neural variability across cortical regions and subjects, we develop a brain-conditioned dynamic routing mechanism that adaptively modulates expert contributions based on neural activity. To address noise and inter-subject variability, we formulate a progressive dual-objective alignment strategy that combines optimal transport for global distribution matching with robust regression for feature-level refinement. Extensive experiments demonstrate consistent improvements in both single-subject and multi-subject settings. These results suggest that structured multi-expert supervision combined with dynamic routing provides a principled framework for multimodal brain decoding.

Keywords: Brain Captioning · Hierarchical Visual Representation · Multimodal Representation Learning

1 Introduction

Decoding natural language descriptions from brain activity, commonly referred to as brain captioning, represents a central challenge in multimodal brain decoding [22, 7, 28, 33, 31]. Recent advances in vision-language models (VLMs) [13, 12] have substantially improved performance by aligning neural signals with textual or visual representations and leveraging large language models for caption generation. Despite this progress, most existing approaches rely on static single-stream alignment, mapping brain activity into either textual embeddings or visual features. Such monolithic supervision overlooks the multi-level and

context-dependent structure of cortex visual processing, making it difficult to model region- and subject-specific variability in brain-vision alignment.

Neuroscientific evidence indicates that visual processing in the cortex is distributed and hierarchically structured [6, 19, 27]. Early visual areas encode low-level geometric properties such as edges and depth, while higher-order regions represent object categories and semantic abstractions. This organization parallels task decomposition in computer vision, where specialized models are trained for depth estimation, surface normal prediction, segmentation, object detection, and optical character recognition. In parallel, recent vision-language research has demonstrated that integrating multiple pre-trained task experts can substantially improve data efficiency and reasoning performance. Prismer [13] demonstrates that pooling diverse vision experts, while keeping their parameters frozen, enables strong performance with limited training data. Rather than relying on a monolithic representation, the multi-expert paradigm aggregates complementary domain knowledge from heterogeneous sources, leading to more expressive and robust multimodal representations. Despite the conceptual correspondence, current brain captioning methods rarely leverage such task-specialized representations as structured intermediate supervision and aggregate visual features in a static manner without considering neural variability across cortical regions and subjects.

To address these limitations, we propose MERPA (**M**ulti-**E**xpert **R**outing with **P**rogressive **A**lignment), a multi-expert alignment framework for brain captioning. Instead of decoding brain signals into a single visual embedding space, MERPA aligns neural activity with structured tokens derived from multiple frozen, task-specialized vision experts spanning geometric and semantic levels. To reflect context-dependent cortical contributions, we introduce a brain-conditioned dynamic routing mechanism that adaptively modulates expert importance based on neural activities. Furthermore, neural signals are inherently noisy and exhibit substantial inter-subject variability, posing challenges for stable cross-modal alignment. We therefore design a progressive alignment optimization strategy that integrates optimal transport for global distribution matching with robust regression for feature-level refinement. The training objective gradually shifts from structural alignment to point-wise robustness, improving convergence stability and robustness to inter-subject variability. We evaluate MERPA on the Natural Scenes Dataset [1] (NSD) under both single-subject and multi-subject settings. Despite not using external caption annotations as supervision during alignment, MERPA achieves superior performance across standard captioning metrics under both single-subject and multi-subject settings. These findings suggest that structured multi-expert supervision combined with dynamic routing constitutes a principled and robust framework for multimodal brain decoding.

Our **main contributions** are as follows: (1) We present MERPA, a structured multi-expert alignment framework that introduces task-specialized intermediate supervision into brain captioning. Our method decomposes multimodal alignment into expert-specific representation spaces, enabling more semantically grounded and interpretable brain-language mapping beyond conventional static feature aggregation. (2) We propose a dynamic routing mechanism to adaptively model

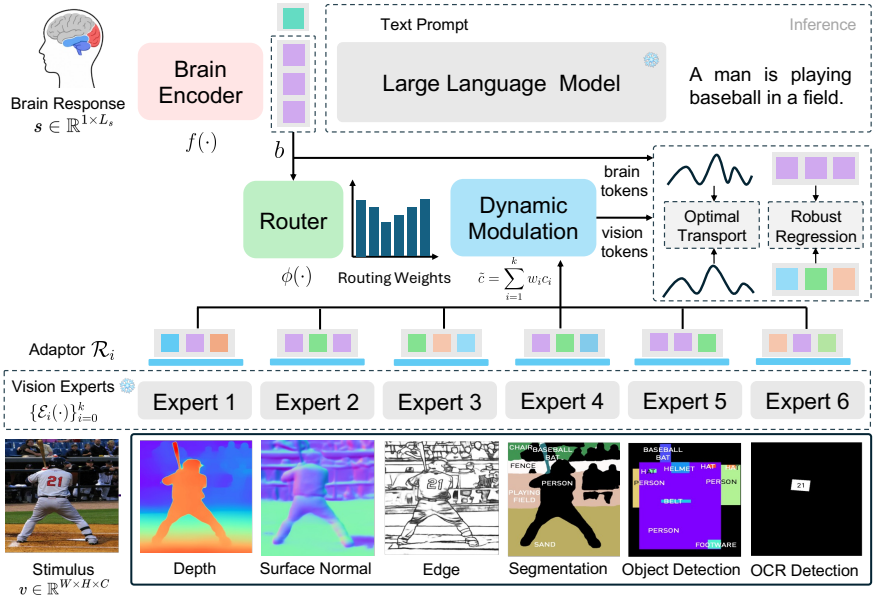

Fig. 1. The overview of MERPA. During training, frozen visual experts extract multi-expert visual tokens from stimulus images as alignment targets, while brain responses are mapped into decoder-compatible brain tokens. The routing module predicts brain-conditioned expert weights to modulate brain tokens. The vision experts and captioning decoder remain frozen; only the brain encoder and routing module are optimized without caption supervision. During inference, predicted brain tokens replace image-derived visual tokens and are fed to the frozen captioning decoder for caption generation.

neural variability across cortical regions and subjects. This design provides a structured way of handling inter-subject and inter-region heterogeneity in brain signals. (3) We demonstrate that MERPA achieves consistent improvements over state-of-the-art baselines on NSD under both single-subject and multi-subject settings without external caption supervision during training.

2 Method

Fig. 1 illustrates the overview of MERPA for brain captioning, comprising three main components: neuro-inspired expert modeling (Sec. 2.1), brain-conditioned dynamic routing (Sec. 2.2), and progressive alignment optimization (Sec. 2.3).

2.1 Hierarchical Vision Expert Modeling

To construct structured visual supervision for brain alignment, we introduce a set of task-specialized vision experts that capture complementary perceptual attributes across multiple abstraction levels. Rather than relying on a single

unified visual embedding, we decompose visual representations into heterogeneous components corresponding to geometric and semantic cues.

Specifically, we employ six frozen vision experts spanning both low-level geometric cues and high-level semantic understanding. For low-level perceptual properties, we adopt DPT [21] for depth estimation, NLL-AngMF [3] for surface normal prediction, and DexiNed [18] for edge detection, capturing geometric structure and fine-grained boundaries. For high-level semantic representations, we utilize Mask2Former [5] for semantic segmentation, UniDet [34] for object detection, and CharNet [14] for optical character recognition, providing object-level, region-level, and textual semantic cues. This expert set forms a hierarchical decomposition of visual perception, captures complementary attributes across abstraction levels rather than arbitrarily combining task models, and provides structured supervision for brain-to-vision alignment. Given an input image v, each expert $\mathcal{E}_i$ produces a task-specific prediction y_i:

$$y_i = \mathcal{E}_i(v), \quad i = 1, \ldots, k, \tag{1}$$

where k denotes the number of experts.

To unify heterogeneous outputs, we apply task-specific post-processing and reshape each prediction into a tensor $y_i \in \mathbb{R}^{H \times W \times C_i}$, where H and W denote spatial dimensions and C_i is task-specific channel size. These features are then projected into a shared embedding space using a multimodal adaptor module $\mathcal{R}$, which maps variable-length expert outputs into a fixed number of visual tokens:

$$c_i = \mathcal{R}_i(y_i) \in \mathbb{R}^{N \times L}. \tag{2}$$

This design produces structured visual tokens across multiple perceptual levels as alignment targets for brain decoding. Unlike monolithic textual or visual representations in prior studies, the proposed expert decomposition preserves task-specific features at different abstraction levels, providing richer supervisory signals and enabling flexible, context-dependent modulation during subsequent alignment. The experts remain frozen throughout training to ensure parameter efficiency and retain domain-specific knowledge from pre-trained models.

2.2 Brain-Conditioned Dynamic Expert Routing

Given multi-expert visual tokens $\{c_i\}_{i=1}^{k}$ extracted from the stimulus image v (Sec. 2.1), we construct an ROI-conditioned alignment target to capture heterogeneous cortical specializations. Rather than statically averaging expert representations [30], we introduce a brain-conditioned routing mechanism that computes expert weights conditioned on ROI-specific neural responses. For each ROI, routing coefficients are predicted from its neural representation and used to construct a convex combination of expert tokens. This adaptive construction yields region-sensitive visual targets, enabling different cortical areas to emphasize perceptual attributes that are consistent with their functional roles.

Let $s \in \mathbb{R}^{1\times L_s}$ denote an ROI-specific brain signal from a subject and $f(\cdot)$ be the brain encoder that maps s to brain tokens:

$$b = f(s) \in \mathbb{R}^{N\times L}. \tag{3}$$

where N is the number of tokens and L is the embedding dimension.

To integrate knowledge from multiple vision experts, we introduce a dynamic router to modulate expert contributions. The routing function $\phi(\cdot)$ is implemented as a lightweight multi-layer perceptron (MLP) followed by softmax normalization:

$$w = \mathrm{Softmax}(\phi(b)) \in \mathbb{R}^k, \tag{4}$$

where $w_i \geq 0$ and $\sum_{i=1}^{k} w_i = 1$.

The final visual alignment target is constructed as a convex combination of expert tokens, weighted by the brain-conditioned routing coefficients:

$$\tilde{c} = \sum_{i=1}^{k} w_i \, c_i, \quad \tilde{c} \in \mathbb{R}^{N\times L}. \tag{5}$$

This routing strategy enables context-dependent expert modulation while preserving fixed token dimensionality. Since routing weights are conditioned on neural representations, the effective visual target distribution varies across stimuli, subjects, and cortical activation patterns, providing adaptive supervision for cross-modal alignment. The model is trained jointly across multiple brain regions and subjects. For simplicity, region- and subject-specific notations are *omitted*.

2.3 Progressive Dual-Objective Alignment Optimization

Given brain tokens $b \in \mathbb{R}^{N\times L}$ and the corresponding routed visual tokens $\tilde{c} \in \mathbb{R}^{N\times L}$ (Sec. 2.2), we optimize the brain encoder by aligning neural representations with image-derived expert tokens in a shared representation space.

Structural Distribution Alignment. To preserve global geometric consistency between modalities, we adopt optimal transport (OT) as the optimization objective [32]. Let μ and ν denote empirical distributions over tokens in b and $\tilde{c}$, respectively. The structural distribution alignment objective is defined as:

$$\mathcal{L}_{\mathrm{OT}} = \min_{\gamma \in \Pi_m(\mu,\nu)} \sum_{i,j} \gamma_{ij} \, d(b_i, \tilde{c}_j) \, , \tag{6}$$

where γ is the transport plan under the partial OT formulation, $d(\cdot,\cdot)$ is cost function, and $\Pi_m(\mu,\nu)$ denotes the set of admissible couplings with transported mass $m = \lambda \min(\|\mu\|_1, \|\nu\|_1)$, $\lambda \in (0, 1]$. By restricting the transported mass, the alignment focuses on the most informative cross-modal correspondences while reducing the influence of redundant or noisy tokens. This formulation preserves global distributional structure without enforcing rigid one-to-one matching.

Robust Feature Refinement. While OT encourages global structural consistency, neural signals are inherently noisy and subject to inter-subject variability. To improve local stability, we introduce the following robust regression loss:

$$\mathcal{L}_{\mathrm{RL}} = \frac{1}{N}\sum_{n=1}^{N}\rho(b_n - \tilde{c}_n)\,, \tag{7}$$

where $\rho(\cdot)$ denotes a robust loss function (implemented as the Fair loss). This term mitigates the influence of outliers and stabilizes token-level alignment.

Progressive Dual-Objective Optimization. The two complementary objectives are jointly optimized under a progressive weighting schedule:

$$\mathcal{L}_t = \lambda_t \mathcal{L}_{\mathrm{OT}} + (1-\lambda_t)\mathcal{L}_{\mathrm{RL}}, \quad \lambda_t = \lambda_0 + (\lambda_1 - \lambda_0)\left(\frac{t}{T}\right)^p, \quad t \in [0, T]. \tag{8}$$

where λ_t is a progressive, time-dependent weighting coefficient, $\lambda_0, \lambda_1 \in [0, 1]$ are the initial and final weights for the OT loss, T denotes the total number of training steps, t is the current step, and p controls the curvature of the progression. Early training emphasizes structural distribution alignment, while later stages gradually increase the influence of robust feature refinement.

Only parameters of the brain encoder f and routing module ϕ are updated during training, while vision experts $\{\mathcal{E}_i\}_{i=1}^{k}$ and language decoder $\mathcal{D}$ remain frozen. During inference, fMRI responses are mapped into decoder-compatible brain tokens b and fed into $\mathcal{D}$ with a text prompt to generate captions.

MERPA extends prior multi-expert and OT-based brain alignment methods in two aspects. Compared with MEVOX [30], which relies on static expert aggregation, MERPA uses brain-conditioned dynamic routing to adapt expert contributions according to ROI-specific neural responses. Compared with BrainOT [32], MERPA does not apply OT to a single visual representation space; instead, it performs progressive alignment toward routed hierarchical expert tokens, combining global distribution matching with robust feature-level refinement.

3 Experiments

Dataset and Evaluation Protocol. We conduct experiments on NSD [1], which contains high-resolution 7T fMRI recordings paired with natural images from COCO [11]. The dataset was collected and distributed under approved ethical protocols with informed consent from participants. We use the anonymized fMRI data for scientific evaluation of brain captioning methods. Following prior works [25, 22, 28], we use the standard data split and report results on the held-out test set. Two protocols are considered: (1) single-subject setting, where all models are trained and evaluated on subject S1. (2) multi-subject setting, where models are jointly trained on S1, S2, S5, and S7. We follow BrainHub [28] evaluation, with standard captioning metrics, including BLEU-k [17], METEOR [4], ROUGE-L [10], CIDEr [26], SPICE [2], CLIP-Score [20], and RefCLIP-Score [9].

Implementation Details. The brain encoder is implemented as a transformer-based architecture [28], which maps brain responses into token space as in visual

Table 1. Brain captioning results on BrainHub under both single-subject and multi-subject settings. ZS denotes zero-shot methods, meaning that no external caption data were used during training. The best-performing results are highlighted in **bold**.

	Method	ZS	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S	RefCLIP-S
Single-Sub	BrainSD [25]	✗	36.21	17.11	7.72	3.43	10.03	25.13	13.83	5.02	61.07	66.36
	OneLLM [8]	✗	47.04	26.97	15.49	9.51	13.55	35.05	22.99	6.26	54.80	61.28
	BrainCap [7]	✗	55.96	36.21	22.70	14.51	16.68	40.69	41.30	9.06	64.31	69.90
	MindEye2 [22]	✗	54.82	38.60	26.49	18.16	17.54	43.77	55.70	10.97	67.54	73.73
	NeuroVLA [23]	✗	57.19	37.17	23.78	15.85	18.60	36.67	49.51	12.39	65.49	-
	UMBRAE [28]	✓	57.84	38.43	25.41	17.17	18.70	42.14	53.87	12.27	67.27	73.04
	VINDEX [29]	✓	59.19	40.20	27.43	19.15	18.93	43.36	57.38	12.25	66.26	72.24
	BrainOT [32]	✓	62.73	43.86	30.15	20.66	21.41	46.74	68.43	14.01	70.24	75.60
	MERPA	✓	**62.91**	**43.96**	**30.29**	**20.94**	**21.67**	**47.05**	**71.91**	**14.55**	**70.50**	**75.84**
Multi-Sub	UMBRAE [28]	✓	59.44	40.48	27.66	19.03	19.45	43.71	61.06	12.79	67.78	73.54
	MEVOX [30]	✓	58.56	40.36	28.09	20.11	19.20	44.47	54.37	11.05	64.23	70.36
	VINDEX [29]	✓	60.16	41.05	28.04	19.31	19.70	44.03	61.01	13.03	67.47	73.45
	BrainOT [32]	✓	63.80	45.17	31.84	22.59	22.06	47.71	75.56	15.17	70.90	76.32
	MERPA	✓	**63.89**	**46.01**	**32.62**	**23.13**	**22.52**	**48.39**	**77.46**	**15.65**	**70.93**	**76.40**

experts. The multi-expert encoders are from Prismer [13] and produce expert-derived visual tokens $\tilde{c} \in \mathbb{R}^{B\times N\times L}$, where $N = 1220$ and $L = 1024$. RoBERTa [15] serves as the frozen decoder. We use AdamW [16] as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01, and a one-cycle learning rate schedule [24] starting at 3×10^{-4}. For partial OT, we set the default mass ratio as 1.0. The weighting hyperparameters λ_0, λ_1, and p are set as 0.9, 0.6 and 2, respectively.

Results Analysis. We compare our method MERPA against recent state-of-the-art brain captioning approaches, including text-aligned and visual-aligned methods. Text-aligned methods use externally annotated COCO captions [11] as supervision, whereas visual-aligned methods rely solely on stimulus images during training and can be regarded as zero-shot (ZS) with respect to textual supervision. Table 1 presents a comparative evaluation. Our method consistently outperforms prior models in both single-subject and multi-subject settings. These quantitative results underscore the effectiveness of the proposed MERPA method. Notably, performance gains are more pronounced in the multi-subject setting, demonstrating improved robustness to inter-subject variability.

To further analyze the contribution of individual components, we conduct ablation experiments. These experimental results in Table 2 confirm that each enhancement contributes meaningfully to model performance.

Effect of Routing Strategies. Replacing dynamic routing with *static* expert token aggregation, as in MEVOX [30], consistently degrades captioning performance, demonstrating that adaptive expert modulation more effectively captures neural context variability. These results suggest the importance of context-aware routing for modeling heterogeneous cortical contributions. Dynamic modulation captures context-dependent neural dynamics and enables more precise alignment of visual representations with stimulus- and region-specific neural responses.

Effect of Alignment Objectives. We further analyze the impact of different alignment strategies. Replacing the proposed alignment with mean squared error

Table 2. Ablation Studies.

	Strategy	BLEU1	METEOR	ROUGE	CIDEr	SPICE	CLIP-S	RefCLIP-S
Routing	Static	56.84	18.52	43.17	49.89	10.33	63.81	69.60
	Dynamic	62.91	21.67	47.05	71.91	14.55	70.50	75.84
Alignment	MSE	57.74	18.28	42.22	52.53	12.08	66.59	72.41
	L1	60.73	20.07	44.49	61.75	13.48	68.54	74.50
	Fair	60.90	20.27	44.83	63.58	13.67	68.54	74.41
	OT	62.73	21.41	46.74	68.43	14.01	70.24	75.60
	Fixed	62.53	21.36	46.53	71.86	14.29	70.32	75.72
	Progressive	62.91	21.67	47.05	71.91	14.55	70.50	75.84
Expert	L: D/N/E	59.84	19.52	44.38	58.76	12.89	67.21	72.95
	H: O/S/T	61.42	20.74	45.91	66.38	13.96	69.32	74.88
	All	62.91	21.67	47.05	71.91	14.55	70.50	75.84

(MSE), as in UMBRAE [28], yields the lowest performance across all metrics. MSE enforces rigid element-wise matching between brain and visual tokens, which may be overly restrictive given the distributed and non-deterministic nature of neural representations [29]. Such point-wise supervision is sensitive to noise and may fail to capture higher-order structural correspondences across modalities. Substituting MSE with alternative robust loss functions (e.g., L1 or Fair loss) leads to consistent improvements. These robust objectives reduce the influence of noisy residuals and alleviate potential over-penalization of mismatched tokens, resulting in more effective alignment between visual stimuli and brain responses. Unlike regression-based losses, OT aligns token distributions rather than enforcing fixed positional correspondence, allowing flexible matching between neural and visual embeddings. This form of structural alignment is better suited to the distributed coding properties of cortical representations. We further examine the combination of OT and Fair loss. A fixed weighting scheme of 0.7 OT and 0.3 Fair improves semantic-sensitive metrics, including CIDEr, SPICE, CLIP-S, and RefCLIP-S compared with OT alone, indicating enhanced robustness. However, slight degradation is observed on surface-level n-gram metrics such as BLEU-1, METEOR, ROUGE-L, indicating a trade-off between lexical precision and semantic consistency. Finally, the proposed progressive dual-objective alignment strategy achieves the best overall performance across metrics. Through this dual-objective design, the model first establishes reliable structural alignment and then gradually enhances robust refinement, enabling a balanced optimization between global distribution matching and local noise suppression.

Effect of Expert Modeling. We analyze the impact of different expert groups by evaluating low-level (L), high-level (H), and combined (L+H) configurations. The low-level experts include depth (D), surface normals (N), and edges (E), which primarily encode geometric and structural cues. The high-level experts consist of object detection (O), semantic segmentation (S), and OCR labels (T), capturing semantic and categorical information. Using only low-level experts (D/N/E) results in weaker semantic metrics, as geometric cues alone are

insufficient to recover high-level scene descriptions. Conversely, relying solely on high-level experts (O/S/T) improves semantic fidelity but lacks fine-grained structural detail. The combined configuration (L+H) consistently achieves the best performance across metrics, indicating that geometric and semantic supervision provide complementary information for brain alignment. These results suggest that hierarchical expert modeling is crucial for capturing the multi-level organization of cortical representations, where both low-level perceptual signals and high-level semantic abstractions contribute to natural language decoding.

4 Conclusion

In this paper, we present a multi-expert routing framework with progressive alignment optimization for brain captioning. Unlike conventional static and monolithic alignment paradigms, our method decomposes visual supervision into task-specialized experts and dynamically modulates contributions based on neural activity, enabling region-sensitive alignment that better reflects the hierarchical organization of cortical representations. To address neural noise and inter-subject variability, we introduce a progressive weighting scheme that first aligns structural distributions and then refines token-level robustness. Our method, despite zero-shot, outperforms state-of-the-arts, demonstrating the effectiveness of combining specialized vision experts with adaptive routing for multimodal brain decoding.

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

References

1. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al.: A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience **25**(1), 116–126 (2022)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: ECCV. pp. 382–398 (2016)
3. Bae, G., Budvytis, I., Cipolla, R.: Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In: ICCV. pp. 13137–13146 (2021)
4. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: ACL Workshop. pp. 65–72 (2005)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1290–1299 (2022)
6. Desimone, R., Albright, T.D., Gross, C.G., Bruce, C.: Stimulus-selective properties of inferior temporal neurons in the macaque. Journal of Neuroscience **4**(8), 2051–2062 (1984)
7. Ferrante, M., Boccato, T., Ozcelik, F., VanRullen, R., Toschi, N.: Multimodal decoding of human brain activity into images and text. In: NeurIPS Workshop (2024)

8. Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., Yue, X.: Onellm: One framework to align all modalities with language. In: CVPR. pp. 26584–26595 (2024)
9. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: EMNLP. pp. 7514–7528 (2021)
10. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755 (2014)
12. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. pp. 34892–34916 (2023)
13. Liu, S., Fan, L., Johns, E., Yu, Z., Xiao, C., Anandkumar, A.: Prismer: A vision-language model with multi-task experts. TMLR (2024)
14. Liu, W., Chen, C., Wong, K.Y.: Char-net: A character-aware neural network for distorted scene text recognition. In: AAAI. vol. 32 (2018)
15. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002)
18. Poma, X.S., Riba, E., Sappa, A.: Dense extreme inception network: Towards a robust cnn model for edge detection. In: WACV. pp. 1923–1932 (2020)
19. Puce, A., Allison, T., Asgari, M., Gore, J.C., McCarthy, G.: Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. Journal of neuroscience **16**(16), 5205–5215 (1996)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
21. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)
22. Scotti, P.S., Tripathy, M., Villanueva, C.K.T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K.A., Abraham, T.M.: Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In: ICML. pp. 44038–44059 (2024)
23. Shen, G., Zhao, D., He, X., Feng, L., Dong, Y., Wang, J., Zhang, Q., Zeng, Y.: Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. In: NeurIPS. pp. 98083–98110 (2024)
24. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. vol. 11006, pp. 369–386. SPIE (2019)
25. Takagi, Y., Nishimoto, S.: Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. arXiv preprint arXiv:2306.11536 (2023)
26. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR. pp. 4566–4575 (2015)
27. Xia, W., de Charette, R., Oztireli, C., Xue, J.H.: Dream: Visual decoding from reversing human visual system. In: WACV. pp. 8226–8235 (2024)
28. Xia, W., de Charette, R., Oztireli, C., Xue, J.H.: Umbrae: Unified multimodal brain decoding. In: ECCV. pp. 242–259 (2024)

29. Xia, W., Oztireli, C.: Exploring the visual feature space for multimodal neural decoding. In: ICCV. pp. 4370–4379 (2025)
30. Xia, W., Oztireli, C.: Mevox: Multi-task vision experts for brain captioning. In: CVPR Workshop (2025)
31. Xia, W., Oztireli, C.: Multigranular evaluation for brain visual decoding. In: AAAI. pp. 17868–17876 (2026)
32. Xiao, Y., Lu, W., Ji, J., Ye, R., Li, G., Ma, X., Hui, B.: Optimal transport for brain-image alignment: Unveiling redundancy and synergy in neural information processing. In: ICCV. pp. 20445–20455 (2025)
33. Zhang, X., Xia, W., Liu, Y., Yang, B., Bozzon, A., Wang, P.: Single-stage fmri-to-3d reconstruction via viewpoint-aware embedding and hierarchical guidance. In: AAAI. pp. 18037–18045 (2026)
34. Zhou, X., Koltun, V., Krähenbühl, P.: Simple multi-dataset detection. In: CVPR. pp. 7571–7580 (2022)