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.

NeuroAlign: A Unified Plug-and-Play Enhancer for Visual and Linguistic Brain Decoding

Jinke Li[1], Yanyan Huang[2], Le Xue[3], Yichi Zhang[3], Yuchen Liu[3], Weihao Zheng[1], Lequan Yu[2], Zhijun Yao[1], and Yu Fu[1(✉)]

[1] Lanzhou University, Lanzhou, China
fuyu@lzu.edu.cn
[2] The University of Hong Kong, Hong Kong, China
[3] Fudan University, Shanghai, China

Abstract. Decoding high-fidelity visual and linguistic content from non-invasive brain recordings is fundamentally impeded by the severe modality misalignment between noisy neural signals and the structured latent spaces of generative foundation models. This gap frequently results in geometric distortions in visual reconstruction and semantic drift in textual decoding. To bridge this divide, we present **NeuroAlign**, a universal, plug-and-play enhancer designed for seamless integration into diverse neural decoding architectures. NeuroAlign orchestrates a dual-pathway alignment strategy: it employs Hierarchical Dynamic Transport to stabilize fine-grained spatial topology for visual fidelity, synergized with Concept Memory Injection to explicitly anchor high-level global semantics. A cross-attention fusion mechanism further adaptively harmonizes these multi-scale representations, resolving feature conflicts before mapping them into the generative prior. Extensive evaluations across state-of-the-art baselines demonstrate the exceptional universality of our framework. NeuroAlign consistently drives simultaneous gains in visual reconstruction and textual consensus, effectively mitigating generative hallucinations. By aligning neural activity with the ground-truth perceptual manifold, our work establishes a robust new standard for unified, multi-modal brain decoding. The code is available at NeuroAlign-Code.

Keywords: Brain Decoding · Multimodal Alignment · Plug-and-Play · Representation Learning

1 Introduction

Deciphering the neural code underpinning visual perception is a foundational pursuit in neuroscience, holding profound implications for the advancement of brain-computer interfaces (BCIs) and the mechanistic understanding of human cognition [4,7,6]. Functional Magnetic Resonance Imaging (fMRI) serves as the primary non-invasive proxy for this endeavor, offering a high-dimensional window into the cortical substrates of visual experience [22,13]. Recently, the integration of deep generative models has catalyzed a paradigm shift, enabling two frontier capabilities: the reconstruction of high-fidelity visual stimuli and the decoding of

semantic captions directly from brain activity [20,24,16,25]. Collectively, these fMRI-to-image and fMRI-to-text tasks constitute complementary modalities for interrogating the visual mind, promising to bridge the chasm between biological intelligence and artificial representation.

Despite these generative strides, neural visual decoding confronts a fundamental bottleneck at the encoding stage. Current fMRI encoders struggle to capture the intrinsic relational structures required to cross the modality gap, resulting in three critical limitations. **First, Geometric Instability**: Pointwise objectives neglect the neural manifold's topological geometry, inducing "manifold fractures" that map adjacent neural states to disjoint latent regions [11]. **Second, Hierarchical Collapse**: Existing architectures conflate multi-scale information through ad-hoc pooling, allowing coarse semantic priors to overwrite fine-grained spatial details [9,14]. **Third, Semantic Indeterminacy**: Without explicit guidance, regression-based projections are ill-posed, leading to semantic ambiguity and drift that severely degrade performance during text-free inference [20,24].

To bridge this divide, we introduce **NeuroAlign**, a universal plug-and-play framework designed to align arbitrary fMRI encoders with pre-trained generative decoders. To effectively address the encoder-decoder misalignment, our architecture isolates the inductive biases required for geometric and semantic preservation into three synergistic modules. First, the Hierarchical Dynamic Transport (HDT) module enforces strict geometric consistency by reformulating feature refinement as a Sinkhorn-regularized optimal transport problem. To complement this, the Concept Memory Injection (CMI) module leverages a retrieval-based mechanism, guided by a Concept Distribution Alignment (CDA) loss, to anchor the decoding process with semantic prototypes, guaranteeing robustness in text-free scenarios. Finally, the Cross-Attention Fusion (CAF) module harmonizes these fine-grained spatial signals with coarse semantic priors via asymmetric gated attention, preventing hierarchical collapse. Ultimately, **NeuroAlign** resolves intrinsic representation bottlenecks, offering a robust enhancement paradigm that elevates decoding performance across both visual and linguistic modalities without modality-specific retraining.

2 Method

2.1 Overview of NeuroAlign

As illustrated in Fig. 1, the **NeuroAlign** pipeline processes fMRI tokens through three sequential stages: geometric regularization, semantic calibration, and cross-scale fusion, as follows:

$$\mathbf{X} \xrightarrow{\text{HDT}} \mathbf{Z}_{\text{fine}} \xrightarrow{\text{Pool}} \mathbf{Z}_{\text{coarse}} \xrightarrow{\text{CMI}} \tilde{\mathbf{Z}}_{\text{coarse}} \xrightarrow{\text{CAF}} \tilde{\mathbf{Z}}. \tag{1}$$

Specifically, the **Hierarchical Dynamic Transport (HDT)** module first stabilizes the geometric structure of tokens at fine resolution (e.g., T_f=256). These features are then compressed to coarse resolution (e.g., T_c=128), where **Concept**

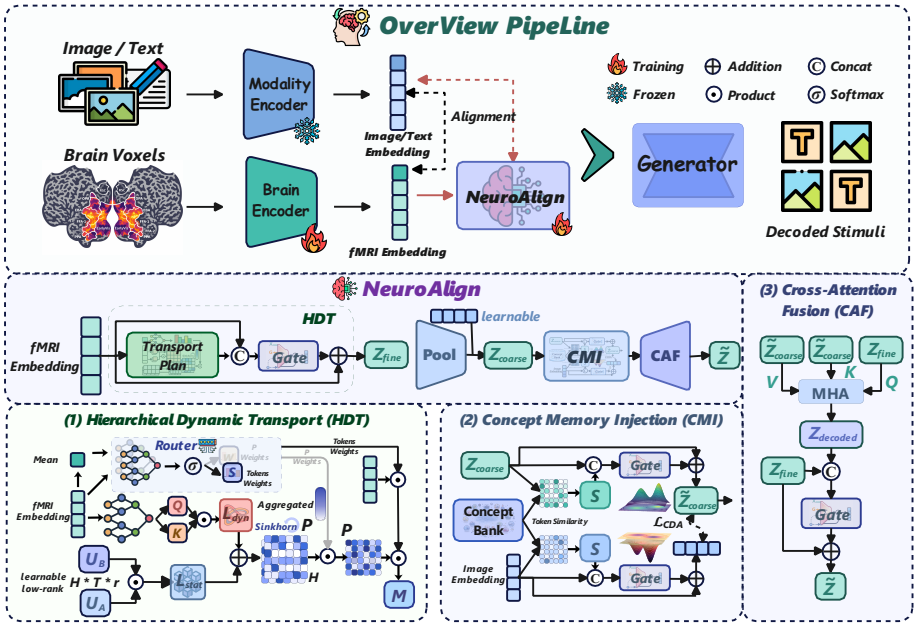

Fig. 1. Overview of the plug-and-play **NeuroAlign** framework. (a) **HDT** aligns fine-grained token geometry via regularized optimal transport (Z_{fine}); (b) **CMI** rectifies semantic drift by injecting retrieved prototypes from a concept bank into pooled coarse tokens ($\tilde{Z}_{\text{coarse}}$); (c) **CAF** integrates calibrated semantics back into the fine resolution via asymmetric cross-attention, yielding the robust condition $\tilde{Z}$ for decoding.

Memory Injection (CMI) rectifies semantic drift using a retrieved concept bank. Finally, **Cross-Attention Fusion (CAF)** propagates the calibrated semantics back to the fine-grained level via gated fusion.

2.2 Hierarchical Dynamic Transport (HDT)

Standard attention mechanisms inherently rely on the assumption of token exchangeability, which neglects the spatial consistency required for fMRI signal decoding. Here, we introduce HDT to formulate feature refinement as an entropy-regularized optimal transport problem, aiming to transport probability mass between voxel-derived and latent token distributions from the noisy voxel space to the structured latent manifold while minimizing geometric distortion costs.

Static-Dynamic Logit Decomposition. To capture both the invariant functional topography and instance-specific activation patterns, we decompose the transport cost for each head h. The static component $\mathbf{L}_{\text{stat}}^{(h)}$ utilizes learnable low-rank factors $\mathbf{U}_A, \mathbf{U}_B \in \mathbb{R}^{T\times r}$ to encode spatial priors, while the dynamic

component $\mathbf{L}_{\text{dyn}}^{(h)}$ is derived from content-aware query-key similarities:

$$\mathbf{L}^{(h)} = (1-\eta)\underbrace{\frac{\mathbf{U}_A^{(h)}\mathbf{U}_B^{(h)\top}}{\sqrt{r}}}_{\text{Static Prior}} + \eta\underbrace{\frac{(\mathbf{X}\mathbf{W}_Q^{(h)})(\mathbf{X}\mathbf{W}_K^{(h)})^\top}{\sqrt{d_k}}}_{\text{Dynamic Content}}. \tag{2}$$

Sinkhorn Transport and Regularization. We obtain the transport plan $\mathbf{P}^{(h)}$ by applying the Sinkhorn-Knopp algorithm to $\exp(\mathbf{L}^{(h)}/\tau(t))$, where $\tau(t)$ is a temperature (entropic regularization strength) scheduled over training. To encourage near doubly-stochastic transport and mass conservation, we add $\mathcal{L}_{\text{DS}}$ to penalize row- and column-sum deviations:

$$\mathcal{L}_{\text{DS}} = \frac{1}{B}\sum_{i=1}^{B}\left(\|\mathbf{P}_i\mathbf{1}-\mathbf{1}\|_2^2 + \|\mathbf{P}_i^\top\mathbf{1}-\mathbf{1}\|_2^2\right). \tag{3}$$

where $\mathbf{1}$ is a vector of ones. This auxiliary objective accelerates the convergence of the Sinkhorn iterations.

Conservative Geometric Update. Using the aggregated plan $\mathbf{P}$, we transport the input features via a gated residual update with ReZero initialization:

$$\mathbf{Z}_{\text{fine}} = \mathbf{X} + \alpha_{\text{geo}} \cdot \sigma(\mathbf{g}_{\text{geo}}) \odot (\mathbf{P}(\mathbf{X}\odot\mathbf{s}) - \mathbf{X}). \tag{4}$$

where $\mathbf{s}$ denotes token importance scores. This step aligns tokens geometrically before semantic processing.

2.3 Semantic Calibration via CMI and CDA

While HDT improves structural consistency, fMRI features often suffer from semantic ambiguity. We address this in the coarse-grained space to reduce computational cost and enhance semantic robustness, as follows:

Hierarchical Pooling. First, we compress the fine features $\mathbf{Z}_{\text{fine}}$ into a coarse representation $\mathbf{Z}_{\text{coarse}} \in \mathbb{R}^{B\times T_c\times D}$ using attention pooling with $T_c = 128$ learnable queries. This abstraction filters out high-frequency noise, creating a stable basis for semantic injection.

Concept Memory Injection (CMI). We maintain a frozen concept bank $\mathbf{M} \in \mathbb{R}^{K\times D}$ containing $K{\approx}1500$ prototype embeddings (e.g., from COCO-Stuff [2], ImageNet [5] and Places [26]). For each coarse token, we retrieve relevant concepts via sparse retrieval:

$$\mathbf{A} = \text{softmax}\left(\text{Mask}_{\text{top-}k}\left(\frac{\mathbf{Z}_{\text{coarse}}\mathbf{W}_q(\mathbf{M}\mathbf{W}_k)^\top}{\tau_c}\right)\right) \tag{5}$$

The retrieved semantic context $\mathbf{S} = \mathbf{A}\mathbf{M}$ is then injected to rectify the coarse features:

$$\tilde{\mathbf{Z}}_{\text{coarse}} = \mathbf{Z}_{\text{coarse}} + \alpha_{\text{sem}} \cdot \sigma(\text{MLP}_g([\mathbf{Z}_{\text{coarse}};\mathbf{S}])) \odot \text{MLP}_\Delta([\mathbf{Z}_{\text{coarse}};\mathbf{S}]). \tag{6}$$

This allows the model to "consult" clear semantic prototypes when the fMRI signal is ambiguous.

Concept Distribution Alignment (CDA). To supervise CMI, we introduce a training-only objective. We compute the probability distribution over the concept bank for both the image (teacher) and fMRI (student) embeddings. The CDA loss minimizes the symmetrized KL divergence between them:

$$\mathcal{L}_{\text{CDA}} = \frac{1}{2}\left(D_{\text{KL}}(p_{\text{img}} \| p_{\text{fMRI}}) + D_{\text{KL}}(p_{\text{fMRI}} \| p_{\text{img}})\right). \tag{7}$$

Here, p_{img} acts as a semantic anchor, guiding the fMRI representation to match the true concept distribution of the visual stimulus.

2.4 Cross-Attention Fusion (CAF)

The final stage uses a residual, gated fine-to-coarse attention block to propagate calibrated semantics without overwriting fine-grained geometry.

Asymmetric Unpooling. Specifically, fine-resolution tokens $\mathbf{Z}_{\text{fine}}$ serve as queries, while CMI-rectified coarse tokens $\tilde{\mathbf{Z}}_{\text{coarse}}$ provide keys and values:

$$\mathbf{Z}_{\text{decoded}} = \text{MultiHead}(\mathbf{Q} = \mathbf{Z}_{\text{fine}}, \mathbf{K} = \tilde{\mathbf{Z}}_{\text{coarse}}, \mathbf{V} = \tilde{\mathbf{Z}}_{\text{coarse}}). \tag{8}$$

Gated Fusion. An element-wise sigmoid gate then injects the decoded semantic signal through a residual path, making CAF conflict-aware rather than a plain cross-attention overwrite:

$$\tilde{\mathbf{Z}} = \mathbf{Z}_{\text{fine}} + \sigma(\text{MLP}([\mathbf{Z}_{\text{fine}}; \mathbf{Z}_{\text{decoded}}])) \odot \mathbf{Z}_{\text{decoded}}. \tag{9}$$

The final output $\tilde{\mathbf{Z}}$ is then fed into the generative decoder.

2.5 Optimization Objectives

The framework is trained end-to-end. The total objective combines the reconstruction loss with the module-specific regularizations defined in Sec. 2.2 and Sec. 2.3:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{CDA}}\mathcal{L}_{\text{CDA}} + \lambda_{\text{DS}}\mathcal{L}_{\text{DS}}. \tag{10}$$

where $\mathcal{L}_{\text{recon}}$ is the standard reconstruction loss (e.g., MSE or contrastive loss depending on the decoder), and $\lambda_{\text{CDA}}, \lambda_{\text{DS}}$ are hyperparameters balancing the constraints, we fix $\lambda_{\text{CDA}} = 0.2$ and $\lambda_{\text{DS}} = 1$ for all experiments.

3 Experiments

3.1 Datasets and Implementation

Datasets. We evaluate our framework on the 7T high-resolution Natural Scenes Dataset (NSD) [1], utilizing the four subjects (sub-01, 02, 05, 07) who completed the full protocol. Neural responses were extracted from the "nsdgeneral" ROI in native space (1.8mm resolution, 13k–16k voxels) to preserve subject-specific spatial topography. Single-trial β-weights, estimated via GLMsingle [15], were voxel-wise z-scored using training statistics to prevent data leakage. Following

Table 1. NeuroAlign (NA) plug-in results on fMRI→image (I) and fMRI→text (T) pipelines. Task abbreviations combine modality (I/T) and setting (S: Single-subject, C: Cross-subject). For frameworks, the suffix "-H" indicates the use of only the high-level branch, whereas its absence denotes a dual-branch (low- and high-level) configuration. The ↑ / ↓ indicate higher/lower is better.

Task	Framework	NA	PixC↑	SSIM↑	A2↑	A5↑	Inc↑	CLIP↑	Eff↓	SwAV↓
I-S	MindEye	0	.3016	.3256	92.88	96.58	94.07	93.04	.6542	.3774
I-S	MindEye	**1**	**.3169**	**.3334**	**93.21**	**96.76**	**94.59**	**93.38**	**.6367**	**.3731**
I-S	MindBridge(S-H)	0	.1425	.2511	85.01	94.08	92.02	94.21	.7277	.4237
I-S	MindBridge(S-H)	**1**	**.1691**	**.2653**	**87.60**	**96.49**	**95.65**	**96.32**	**.6709**	**.3935**
I-C	MindBridge(C-H)	0	.1596	.2654	86.27	95.01	92.44	94.35	.7182	.4141
I-C	MindBridge(C-H)	**1**	**.1764**	**.2684**	**88.52**	**96.66**	**95.92**	**96.88**	**.6505**	**.3827**
I-C	MindBridge(C)	0	.3066	.3367	93.55	96.77	93.17	95.79	.6857	.3998
I-C	MindBridge(C)	**1**	**.3197**	**.3334**	**94.18**	**97.58**	**95.84**	**96.68**	**.6557**	**.3727**
Task	**Framework**	**NA**	**B1↑**	**B2↑**	**B3↑**	**B4↑**	**CIDEr↑**	**SP↑**	**C-S↑**	**RC-S↑**
T-S	BrainCap	0	54.71	35.01	21.95	14.14	38.35	8.77	64.34	69.68
T-S	BrainCap	**1**	**56.28**	**38.87**	**26.47**	**18.29**	**50.30**	**10.74**	**67.00**	**72.07**
T-S	Umbrae(S)	0	57.57	38.36	25.43	17.23	52.24	11.88	66.41	72.21
T-S	Umbrae(S)	**1**	**58.19**	**38.98**	**25.99**	**17.72**	**54.29**	**12.42**	**66.95**	**72.75**
T-C	Umbrae(C)	0	58.80	39.65	26.65	18.19	55.73	12.46	67.08	72.87
T-C	Umbrae(C)	**1**	**60.92**	**42.87**	**28.48**	**20.58**	**62.12**	**14.42**	**70.29**	**75.78**

standard protocols, we used 24,980 unique training trials per subject and 982 shared test images (averaged across 3 repetitions), paired with MS-COCO [12] captions for multimodal tasks.

Implementation Details. Implemented in PyTorch on an NVIDIA RTX 5880 GPU, NeuroAlign was evaluated as a plug-and-play module across four state-of-the-art baselines: MindBridge [23], MindEye [18], UMBRAE [25], and Brain-Captioning [8]. Specifically, NeuroAlign was inserted between the fMRI encoder and the generative decoder, with the entire architecture jointly optimized end-to-end. For rigorous and fair comparison, all enhanced models and reproduced baselines were trained under strictly identical hyperparameter, preprocessing, and evaluation protocols.

Evaluation Metrics. We comprehensively evaluate multimodal decoding performance across visual and linguistic dimensions. Visual reconstruction is evaluated using pixel-wise Pearson correlation (PixC), Structural Similarity Index (SSIM), and AlexNet [10] features (A2, A5) for structural fidelity, alongside InceptionV3 (Inc) [19], CLIP-Vision (CLIP) [17], EfficientNet-B1 (Eff) [21], and SwAV-ResNet50 (SwAV) [3] for semantic alignment. Linguistic decoding is assessed via BLEU (B1–B4), CIDEr, and SPICE (SP), supplemented by cross-modal CLIP-Score (C-S) and RefCLIP-Score (RC-S).

3.2 Result Analysis

Quantitative Results. To rigorously benchmark the universality of NeuroAlign, we integrated our framework into state-of-the-art architectures across both visual

Fig. 2. Qualitative decoding results on Subject 1. We compare MindBridge and UMBRAE baselines against their NeuroAlign-enhanced versions. Original baselines frequently exhibit semantic drift by missing key entities like surfboards or misidentifying birds as fish. Conversely, integrating NeuroAlign effectively anchors global semantics to ensure superior geometric fidelity in images and precise entity capture in text captions.

reconstruction and linguistic decoding tasks. As summarized in Table 3.1, NeuroAlign consistently outperforms baselines across almost all metrics, validating its ability to enhance both geometric fidelity and semantic alignment. Table 3.1 validates the universality of NeuroAlign across diverse decoding architectures. **(1) Visual Reconstruction**: NeuroAlign consistently enhances geometric fidelity and semantic alignment. In the MindEye single-subject setting, it elevates Pixel Correlation to 0.3169 and SSIM to 0.3334, while boosting CLIP similarity to 93.38%. This robustness extends to MindBridge, where the challenging cross-subject dual-branch integration achieves a Pixel Correlation of 0.3197 and a CLIP similarity of 96.68%, significantly surpassing baseline capacities in recovering spatial topology. **(2) Linguistic Decoding**: Improvements are even more pronounced in captioning tasks. By equipping BrainCap with our concept injection mechanism, we observe a massive +11.95 point surge in CIDEr (reaching 50.30). Similarly, under the UMBRAE cross-subject setting, our model attains a CIDEr score of 62.12 (vs. baseline 55.73). These substantial margins confirms that NeuroAlign effectively anchors global semantics to mitigate hallucinations and generate precise entity descriptions.

Table 2. Ablation study of NeuroAlign components. We incrementally evaluate the impact of Hierarchical Dynamic Transport (HDT), Concept Memory Injection (CMI), and Cross-Attention Fusion (CAF) using the Mindbridge subject-1 baseline. ↑ / ↓ indicate higher/lower is better.

Variant	HDT	CMI	CAF	PixC↑	SSIM↑	A2↑	A5↑	Inc↑	CLIP↑	Eff↓	SwAV↓
(a) Baseline				.1616	.2870	89.08	96.07	93.19	95.45	.6909	.4024
(b) + HDT	✓			.1675	.3013	89.57	97.12	96.96	96.79	.6766	.3960
(c) + CMI	✓	✓		.1712	.3068	89.78	97.80	96.57	97.41	.6775	.3874
(d) Full	✓	✓	✓	**.1776**	**.3074**	**90.31**	**98.21**	**97.28**	**98.11**	**.6684**	**.3652**

Qualitative Analysis. Fig. 2 visualizes the qualitative superiority of our method. **(1) Baseline Failures**: Vanilla baselines are plagued by severe semantic drift and structural incoherence. Specifically, UMBRAE suffers from generative hallucinations—misinterpreting a surfboard as "glistening sunlight", misclassifying a bird as a "fish" and describing a hot dog consumption scene as merely a "generic smile". Similarly, MindBridge fails to synthesize coherent visual entities, leaving core structures absent. **(2) NeuroAlign Rectification**: Conversely, integrating Concept Memory Injection anchors global semantics, enabling the upgraded UMBRAE to correctly retrieve precise prototypes (e.g., surfboard, bird, hot dog). This semantic grounding synergizes with Dynamic Transport to elevate visual fidelity in MindBridge, successfully reconstructing intricate details such as surfboard and giraffe. By harmonizing local textures with global semantics, NeuroAlign ensures decoded outputs faithfully reflect the original stimuli.

3.3 Ablation Studies

Table 2 reports an incremental, pipeline-order ablation on the MindBridge baseline rather than a full factorial isolation. (1) Geometric Foundation: adding **HDT** improves Pixel Correlation from 0.1616 to **0.1675** and SSIM from 0.2870 to **0.3013**. (2) Semantic Anchoring: adding **CMI** further improves CLIP similarity from 96.79 to **97.41** and A5 from 97.12 to **97.80**. (3) Unified Harmony: adding **CAF** integrates calibrated coarse semantics with fine-grained features, yielding the best Pixel Correlation (**0.1776**) and CLIP similarity (**98.11**). These results support the designed module order, while standalone modules, two-module variants, and simpler alignment controls remain future controlled comparisons.

4 Conclusion

In this work, we presented **NeuroAlign**, a universal plug-and-play framework designed to bridge the fundamental modality gap in neural decoding. Through comprehensive evaluations across diverse reconstruction architectures, we demonstrated that NeuroAlign successfully reconciles the inherent tension between spatial geometric fidelity and global semantic alignment. By synergizing Hierarchical Dynamic Transport with Concept Memory Injection, our approach explicitly

anchors ambiguous neural signals to precise semantic prototypes, thereby effectively rectifying generative hallucinations and semantic drift. Furthermore, ablation studies confirm that the Cross-Attention Fusion mechanism is essential for adaptively harmonizing these multi-scale representations. Ultimately, NeuroAlign establishes a robust and highly adaptable new standard for unified multi-modal brain decoding.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (Grant No. 62577037) and the Central Government Guides Local Science and Technology Development Fund Projects (Grant No. 25ZYJA017), in part by the Youth Science and Technology Talent Innovation Project of Lanzhou City (Grant No. 2025-QN-037).

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., Hutchinson, J.B., Naselaris, T., Kay, K.: A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence **25**(1), 116–126. https://doi.org/10.1038/s41593-021-00962-x
2. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: Thing and Stuff Classes in Context. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1209–1218. https://doi.org/10.1109/CVPR.2018.00132
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc. (2020)
4. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., King, J.R.: Decoding speech perception from non-invasive brain recordings **5**(10), 1097–1107. https://doi.org/10.1038/s42256-023-00714-5
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. IEEE (2009)
6. Doerig, A., Kietzmann, T.C., Allen, E., Wu, Y., Naselaris, T., Kay, K., Charest, I.: High-level visual representations in the human brain are aligned with large language models **7**(8), 1220–1234. https://doi.org/10.1038/s42256-025-01072-0
7. Du, C., Fu, K., Li, J., He, H.: Decoding Visual Neural Representations by Multimodal Learning of Brain-Visual-Linguistic Features. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10760–10777 (Sep 2023). https://doi.org/10.1109/TPAMI.2023.3263181
8. Ferrante, M., Ozcelik, F., Boccato, T., VanRullen, R., Toschi, N.: Brain Captioning: Decoding human brain activity into images and text. https://doi.org/10.48550/arXiv.2305.11560
9. Guo, W., Sun, G., He, J., Shao, T., Wang, S., Chen, Z., Hong, M., Sun, Y., Xiong, H.: A Survey of fMRI to Image Reconstruction. https://doi.org/10.48550/arXiv.2502.16861

10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (May 2017). https://doi.org/10.1145/3065386
11. Lin, B., Kriegeskorte, N.: The topology and geometry of neural representations **121**(42), e2317881121. https://doi.org/10.1073/pnas.2317881121
12. 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: Computer Vision – ECCV 2014. pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
13. Muttenthaler, L., Greff, K., Born, F., Spitzer, B., Kornblith, S., Mozer, M.C., Müller, K.R., Unterthiner, T., Lampinen, A.K.: Aligning machine and human visual representations across abstraction levels **647**(8089), 349–355. https://doi.org/10.1038/s41586-025-09631-6
14. Ozcelik, F., VanRullen, R.: Natural scene reconstruction from fMRI signals using generative latent diffusion **13**(1), 15666. https://doi.org/10.1038/s41598-023-42891-8
15. Prince, J.S., Charest, I., Kurzawski, J.W., Pyles, J.A., Tarr, M.J., Kay, K.N.: Improving the accuracy of single-trial fMRI response estimates using GLMsingle **11**, e77599. https://doi.org/10.7554/eLife.77599
16. Qiu, W., Huang, Z., Hu, H., Feng, A., Yan, Y., Ying, R.: Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=EiAQrilPYP
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning (2021)
18. Scotti, P., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., cohen, e., Dempster, A., Verlinde, N., Yundler, E., Weisberg, D., Norman, K., Abraham, T.: Reconstructing the minds eye: fMRI-to-Image with contrastive learning and diffusion priors. In: Advances in Neural Information Processing Systems. vol. 36, pp. 24705–24728. Curran Associates, Inc. (2023)
19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308
20. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14453–14463 (2023). https://doi.org/10.1109/CVPR52729.2023.01389
21. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (09–15 Jun 2019)
22. Tang, J., LeBel, A., Jain, S., Huth, A.G.: Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience **26**(5), 858–866 (May 2023). https://doi.org/10.1038/s41593-023-01304-9
23. Wang, S., Liu, S., Tan, Z., Wang, X.: Mindbridge: A cross-subject brain decoding framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11333–11342 (June 2024)

24. Xia, W., de Charette, R., Öztireli, C., Xue, J.H.: Dream: Visual decoding from reversing human visual system. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8211–8220. IEEE Computer Society (2024). https://doi.org/10.1109/WACV57701.2024.00804
25. Xia, W., de Charette, R., Oztireli, C., Xue, J.H.: UMBRAE: Unified Multimodal Brain Decoding. In: Computer Vision – ECCV 2024. pp. 242–259. Cham (2025). https://doi.org/10.1007/978-3-031-72667-5_14
26. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)