Connecting Brain Activity to Language through Mixed-Resolution Visual Representations

Anonymous authors
Paper under double-blind review

Abstract

Brain-to-text decoding aims to describe viewed scenes from fMRI recordings. Pretrained vision-language models provide rich visual and linguistic priors, and their visual representations determine the spatial detail that brain decoders are trained to predict. Inspired by the non-uniform allocation of visual information in human perception, we propose MindMRQ, which uses adaptive visual granularity as a bridge between cortical representations and language models. Its mixed-resolution visual representations retain finer detail in regions sensitive to spatial compression and encode surrounding context more coarsely while maintaining full image coverage. We then connect brain activity to these visual targets through dual-level visual-semantic alignment, combining region-level matching with query-level semantic supervision to support language generation from brain-predicted representations. Extensive experiments demonstrate competitive decoding performance, with controlled comparisons and ablations supporting the effectiveness of the proposed representations and alignment strategy.

1 Introduction

Translating visually evoked brain activity into natural language offers a way to describe human visual experience. Large-scale vision-language models provide rich visual and linguistic priors for learning this mapping from limited brain-signal data. Models such as CLIP (Radford et al., 2021) provide transferable visual representations, while brain encoders map functional magnetic resonance imaging (fMRI) responses into a representation space readable by language models (Xia et al., 2024; Chen et al., 2025a). These visual representations also define the brain decoder's prediction targets, making their spatial organization central to connecting brain activity to language generation.

A straightforward approach uses spatial features covering the entire image as prediction targets. For example, UMBRAE aligns brain representations elementwise with regular spatial features from an intermediate CLIP layer (Xia et al., 2024). In Figure 1(a), uniform tokens represent both the cat and the surrounding bedding with equally sized patches. These uniformly fine-grained targets preserve full-image spatial structure but require the brain encoder to recover regional details that are not equally important for describing the scene.

Subsequent works reorganize alignment supervision through semantic selection. Mind-SA aligns the average representation of adaptively selected image patches with a global brain representation; Yo'Mind uses personalized multimodal soft assignment to align each brain token with a weighted combination of visual and textual semantics (Chen et al., 2025b; 2026). These methods aim to mitigate semantic mismatch between stimulus images and brain signals, focusing supervision on content consistent with brain activity. However, semantic selection does not ensure that useful scene context is retained. Our qualitative comparisons show that Yo'Mind can identify the main objects while still omitting useful scene context (Figure 3). Figure 1(b) schematically illustrates this concern: contextual regions such as the bed receive less emphasis despite their value for scene description.

Inspired by the non-uniform allocation of visual information in human perception, we adapt the spatial granularity of visual prediction targets. A region's need for fine detail is distinct

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

from its value for scene description. In Figure 1, details of the cat's shape help identify the object, while the bed and pillows establish its surroundings without requiring equally fine texture representations.

Based on this principle, we propose MindMRQ, a brain-to-text framework that uses adaptive visual granularity as a bridge between cortical representations and language models. As illustrated in Figure 1(c), we partition the image according to the sensitivity of image features to spatial compression, using smaller regions at detail-sensitive locations and larger regions elsewhere. Resampling these differently sized regions to a common input size retains relatively finer detail in smaller regions, while larger regions encode scene context at a lower sampling density. The resulting mixed-resolution visual representations cover the entire image and form the brain decoder's prediction targets.

Establishing this bridge requires the visual representations to be both predictable from fMRI and readable by a pretrained language interface. We first adapt the mixed-resolution visual encoder and tune its language-interface queries to support language generation. We then train the brain encoder with dual-level visual-semantic alignment. At the visual level, optimal transport (OT) establishes soft correspondence between brain-predicted and image-derived region tokens based on visual features and regional geometry, accommodating image-dependent partitions. At the semantic level, shared queries read the predicted and target representations separately, and the corresponding query outputs are aligned. Our main contributions are summarized as follows:

1. We propose MindMRQ, a brain-to-text framework that uses content-adaptive mixed-resolution visual representations to preserve fine detail and scene context through non-uniform spatial granularity.
2. We introduce dual-level visual-semantic alignment, combining OT-based matching of mixed-resolution tokens with query-output alignment to connect brain activity with language generation.
3. Extensive experiments demonstrate competitive decoding performance, while controlled comparisons and ablations support the effectiveness of the mixed-resolution representations and alignment strategy.

2 Related work

Brain-to-image reconstruction. Brain-to-image methods typically map fMRI into the latent space of a pretrained visual or generative model and then use generative priors to reconstruct visual stimuli. Takagi & Nishimoto (2023) guide a latent diffusion model with multiple decoded inputs; MindEye combines contrastive learning with a diffusion prior for image retrieval and reconstruction (Scotti et al., 2023). MindEye2 subsequently improves adaptation with limited data through a shared subject space and multi-subject pretraining (Scotti et al., 2024). More recently, BrainJanus quantizes brain signals into discrete tokens and jointly learns bidirectional brain-image and brain-text transformations through next-token prediction in a unified autoregressive model (Wu et al., 2026). These works demonstrate the role of pretrained representations and generative priors in visual brain decoding and have also encouraged research on expressing scene information in language.

Brain-to-text decoding. Unlike pixel-level reconstruction, brain-to-text decoding focuses on expressing the semantic content of visual stimuli in natural language. Brain Captioning and MindGPT leverage pretrained models to generate descriptions of visual stimuli from brain activity (Ferrante et al., 2023; Chen et al., 2025a). UMBRAE, Neuro2Language, and MindLLM extend decoding capabilities through multi-subject training, brain-representation modeling, and instruction tuning (Xia et al., 2024; Shen et al., 2024; Qiu et al., 2025). To improve semantic correspondence between brain and visual representations, Mind-SA combines pre-alignment with global image representations and supervision from the average representation of selected image patches; Yo'Mind connects brain tokens with visual and textual semantic sets through optimal transport with partial assignment (Chen et al., 2025b; 2026). These approaches primarily adjust which semantic features are used for alignment

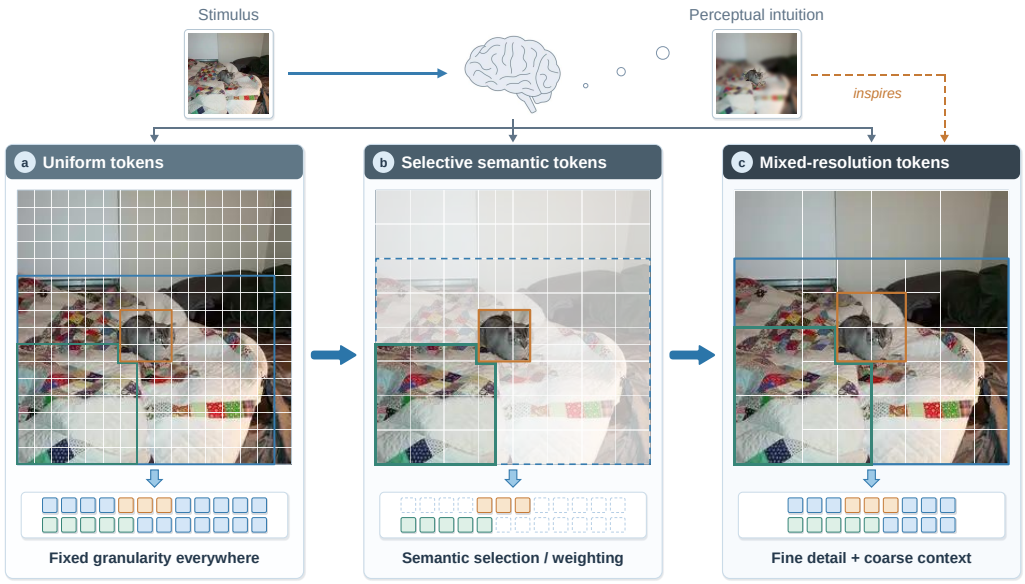

Figure 1: Conceptual comparison of visual targets for brain-to-text alignment. (a) Uniform tokens represent the entire image at a fixed spatial granularity. (b) Selective semantic tokens emphasize selected components through semantic selection or weighting. Faded regions and dashed token outlines indicate reduced emphasis on some scene context. (c) Mixed-resolution tokens retain finer representations in smaller regions while representing surrounding context more coarsely. Orange, teal, and blue boxes highlight the cat, pillow, and bedding context, respectively.

and how they correspond to brain representations. MindMRQ instead reorganizes the spatial granularity of the visual prediction targets themselves, allocating finer representations to compression-sensitive regions and coarser representations to their surroundings while retaining full image coverage.

3 Method

MindMRQ first constructs mixed-resolution visual representations from training images, then learns to predict them from fMRI and generate descriptions. Figure 2 summarizes the training procedure into two parts: visual learning and brain alignment. The former consists of visual representation learning followed by query tuning, while the latter trains the brain encoder against targets from the fixed mixed-resolution visual encoder.

3.1 Content-adaptive mixed-resolution visual representations

We construct content-adaptive image regions through quadtree partitioning with feature-based region scoring (Ronen et al., 2023). Starting from four equally sized coarse regions, we repeatedly divide the highest-scoring eligible region into four equal parts until K leaf nodes are obtained. A frozen feature extractor measures each region's sensitivity to spatial compression by comparing corresponding features in the original image with those in the corresponding downsampled-then-restored image. Regions more sensitive to spatial compression are subdivided first, while the remaining regions retain coarser representations. The scoring and scale settings are detailed in Appendix A.1.

We resample the image patch corresponding to each region to 32×32, denoted by $\widetilde{R}_k$, apply the CLIP ViT-B/32 patch projection, and add position and scale information:

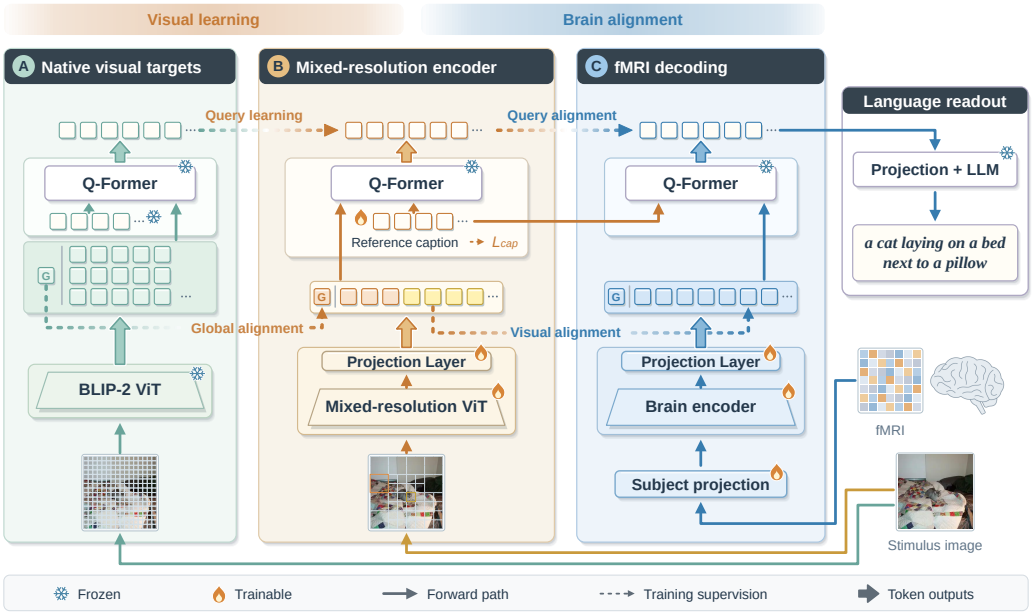

Figure 2: Overview of MindMRQ. During visual learning, the mixed-resolution visual encoder (B) learns from the frozen BLIP-2 branch (A) through global-feature alignment and query-output supervision, followed by tuning only the query inputs with reference captions. During brain alignment, the fMRI branch (C) learns to predict mixed-resolution visual representations through OT-based token matching and query-output alignment, while B remains frozen. At inference, the fixed Q-Former reads the predicted representations, and its projected outputs guide language generation.

$$\mathbf{e}_k = \mathrm{PatchEmbed}(\widetilde{R}_k) + \mathbf{e}_{\mathrm{pos}}(\mathbf{p}_k) + \mathbf{e}_{\mathrm{scale}}(\ell_k). \tag{1}$$

Here $\mathbf{p}_k$ denotes the region center and ℓ_k is the scale index. All patches yield embeddings of the same dimensionality but cover different extents: larger patches retain context at a lower sampling density, while smaller patches retain finer spatial detail. The patch embeddings and a class token are then fed into the mixed-resolution visual Transformer, followed by visual projection layers to form representations $M = E_\phi(I) = [\mathbf{g}; \mathbf{v}_1; \ldots; \mathbf{v}_K]$. Here I is the stimulus image and E_ϕ is the mixed-resolution visual encoder with parameters ϕ. The global summary token $\mathbf{g}$ complements the mixed-resolution tokens $\{\mathbf{v}_k\}_{k=1}^{K}$, each associated with a partition region.

3.2 Adapting the mixed-resolution visual encoder

To connect mixed-resolution visual representations to language generation, we adopt the pretrained language interface of BLIP-2 (Li et al., 2023). The Q-Former reads the visual representations, and a language projection layer maps the resulting query outputs into the input space of the frozen language model. We denote the query inputs by Q and the Q-Former mapping by $\mathcal{Q}_\psi$, whose parameters ψ remain frozen. Visual learning consists of the following two steps.

Visual representation learning. A frozen native BLIP-2 branch provides global-feature and query-output targets for the same training images. The mixed-resolution branch learns these targets jointly through query alignment, image-level contrastive learning, and global-feature alignment. During training, we update the mixed-resolution visual encoder's last two Transformer blocks, scale embeddings, and visual projection layers.

Query tuning. We then fix the mixed-resolution visual encoder and fine-tune only the 32 query inputs Q using reference-caption negative log-likelihood. The tuned $Q^\star$ will be fixed during subsequent brain training, with the mixed-resolution visual encoder supplying representations and corresponding query-output targets. Complete objectives and settings are given in Appendices A.3-A.4.

3.3 Brain-to-text decoding

We use a shared brain encoder with subject-specific input mappings (Xia et al., 2024) to predict mixed-resolution visual representations $\widehat{M} = F_\theta(\mathbf{x}^{(s)}, s)$ from fMRI. The predicted representations contain K mixed-resolution tokens and one global token, for a total length of $K+1$. An auxiliary head predicts a geometry vector $\widehat{\mathbf{c}}_k$ for each region token, encoding the corresponding region's center coordinates, width, and height for set matching during training. Architectural details are provided in Appendix A.2.

Dual-level visual-semantic alignment. Content-adaptive partitioning changes the spatial extents associated with region tokens across images, so token indices alone do not define consistent spatial correspondence. We therefore use optimal transport (OT) to match brain-predicted and image-derived region-token sets based on visual-feature and regional-geometry similarity. For sample i, $\Pi_{i,ab}$ is the mass assigned from predicted token a to target token b. The matching cost and token-level visual alignment loss are

$$\begin{aligned} D_{i,ab} &= 1 - \cos(\widehat{\mathbf{v}}_{i,a}, \mathbf{v}_{i,b}) + \frac{\lambda_{\text{geo}}}{4}\|\widehat{\mathbf{c}}_{i,a} - \mathbf{c}_{i,b}\|_1, \\ \mathcal{L}_{\text{visual}} &= \frac{1}{B}\sum_{i=1}^{B}\langle \Pi_i, D_i \rangle. \end{aligned} \tag{2}$$

Here B is the batch size, and $\widehat{\mathbf{c}}_{i,a}$ and $\mathbf{c}_{i,b}$ encode predicted and target region center coordinates, widths, and heights. We compute Π_i with entropy-regularized Sinkhorn iterations (Cuturi, 2013), assigning each of the K region tokens mass $1/K$. The OT formulation and implementation are given in Appendix A.3.

At the semantic level, the fixed Q-Former reads the visual-encoder and brain-predicted representations using the same queries $Q^\star$, producing output representations Z_i and $\widehat{Z}_i$, respectively:

$$Z_i = \mathcal{Q}_\psi(Q^\star; M_i), \qquad \widehat{Z}_i = \mathcal{Q}_\psi(Q^\star; \widehat{M}_i). \tag{3}$$

The query alignment loss $\mathcal{L}_{\text{query}}$ aligns the normalized query outputs through cosine distance and Smooth L1; the image-level contrastive loss $\mathcal{L}_{\text{contrastive}}$ applies symmetric InfoNCE (Oord et al., 2018) to the means of the output representations. Together, these losses form the query-level supervision. The overall brain-training objective is

$$\mathcal{L}_{\text{brain}} = \mathcal{L}_{\text{visual}} + \lambda_{\text{q}}\mathcal{L}_{\text{query}} + \lambda_{\text{c}}\mathcal{L}_{\text{contrastive}}. \tag{4}$$

Here λ_{q} and λ_{c} weight the query alignment and image-level contrastive losses, respectively. Brain alignment updates only the brain encoder and prediction heads. The complete losses are given in Appendix A.3. At inference, the fixed Q-Former reads the predicted representations using $Q^\star$. The language projection layer maps the resulting query outputs to a continuous prefix for OPT-2.7B, which generates the description.

4 Experiments

We evaluate MindMRQ's brain-to-text decoding performance on NSD and analyze the roles of mixed-resolution visual representations and the accompanying decoding design through partitioning comparisons, spatial granularity analysis, and component ablations.

Dataset. We use data from Subjects 1, 2, 5, and 7 of the Natural Scenes Dataset (NSD) (Allen et al., 2022). For each subject, we use preprocessed fMRI responses from the nsdgeneral region of interest (ROI), which comprises visually responsive voxels in the posterior

cortex. Each subject has 8,859 training stimulus images, and the four subjects share 982 test stimulus images. The mixed-resolution visual encoder and queries are adapted using training stimulus images, and the brain encoder is jointly trained on responses from different subjects. At inference, repeated responses to the same image are averaged.

Implementation details. We use 40 mixed-resolution tokens and one global token. The visual encoder and queries are adapted for 15 and 5 epochs, respectively, and the brain encoder is trained to convergence using AdamW for 9,000 steps. Optimization and generation settings are given in Appendix A.4.

Evaluation metrics. Following prior brain-to-text studies (Shen et al., 2024; Chen et al., 2025b; 2026), we evaluate generated descriptions using BLEU-1/2/3/4 (Papineni et al., 2002), METEOR (Denkowski & Lavie, 2014), ROUGE-L (Lin & Hovy, 2003), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). We use COCO reference captions and average the scores across the four subjects. In tables, B@1–B@4 abbreviate BLEU-1–BLEU-4.

4.1 Comparison with existing methods

Qualitative comparison. We first compare MindMRQ with UMBRAE and Yo'Mind on the same visual stimuli. UMBRAE provides a representative baseline based on uniform visual representations, while Yo'Mind represents recent approaches that adaptively select semantic components to reduce the mismatch between brain and visual representations. As shown in Figure 3, MindMRQ identifies the main objects while retaining specific surroundings and relationships between objects; some outputs from the other methods contain object misidentifications or omissions.

MindMRQ appears to recover specific object and action details beyond a general scene description. For instance, it identifies doughnuts in a display case, whereas Yo'Mind describes food on a table and UMBRAE refers to boxed items in a store (Figure 3(d)). In the skateboarding scene, MindMRQ describes the rider as "in the air." Yo'Mind also recognizes skateboarding but describes movement down a ramp, while UMBRAE describes a person standing indoors (Figure 3(e)). MindMRQ can also combine object descriptions with their surroundings. In the train-station scene, it mentions people waiting for a train, whereas both other methods recognize the people but place them on a sidewalk or city street (Figure 3(b)). Similarly, UMBRAE identifies the flowers but omits the vase and table; MindMRQ includes all three, while Yo'Mind describes bananas on a table (Figure 3(h)). These observations suggest the potential of mixed-resolution visual representations to retain fine-grained content alongside scene context.

Quantitative comparison. We compare MindMRQ with the reported brain-to-text decoding results of seven existing methods: Brain Captioning, Takagi et al., MindGPT, Neuro2Language, UMBRAE, Mind-SA, and Yo'Mind. Table 1 summarizes all eight metrics on NSD Subjects 1, 2, 5, and 7. MindMRQ obtains the highest reported scores on seven metrics: BLEU-1/2/3/4, ROUGE-L, CIDEr, and SPICE.

Compared with UMBRAE, which predicts uniform spatial visual representations, MindMRQ obtains relative improvements of 48.2%, 46.7%, and 24.5% in BLEU-4, CIDEr, and SPICE, respectively. MindMRQ also achieves higher scores than Mind-SA and Yo'Mind on the seven metrics listed above. In particular, its BLEU-4 and CIDEr scores exceed those of Yo'Mind by 30.6% and 51.1%, respectively. Overall, these results suggest that MindMRQ provides an effective framework for decoding scene descriptions from brain activity, complementing the qualitative observations of recovered object details and scene context.

4.2 Analyzing mixed-resolution visual representations

We compare partitioning strategies at a fixed region-token count, then vary K to examine the effect of spatial granularity on caption quality.

Content-adaptive spatial partitioning. We fix $K = 40$ and compare our content-adaptive strategy (labeled Content-driven), random quadtree partitioning, and a spatially shifted

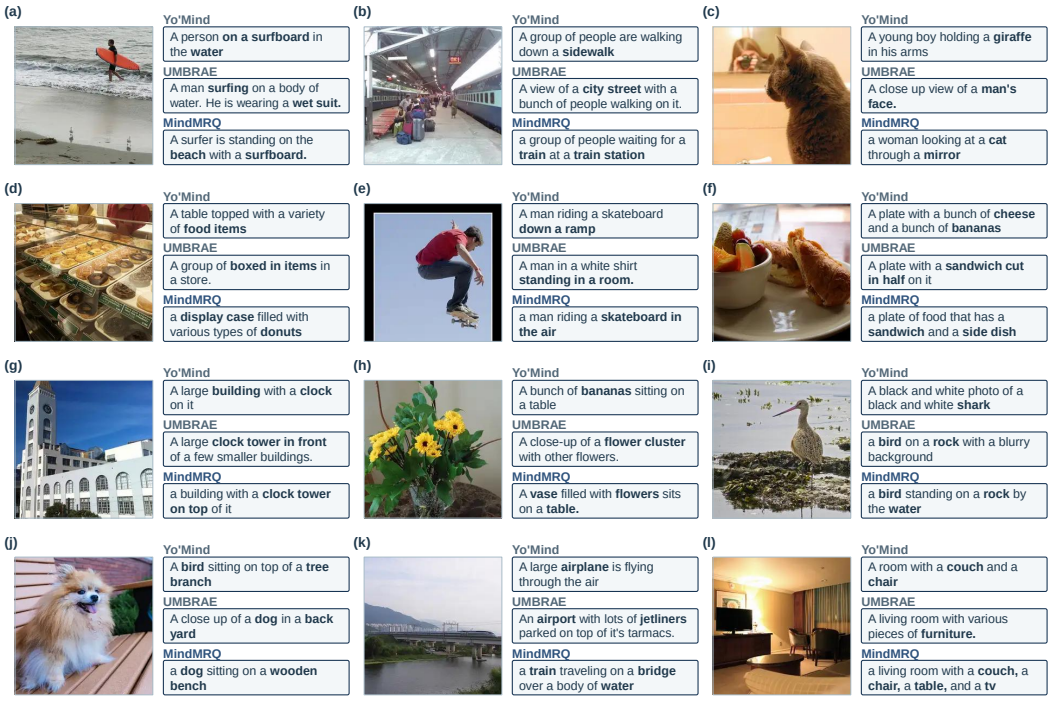

Figure 3: Qualitative brain-to-text comparisons for NSD Subject 1. Each example shows the stimulus on the left and the original outputs of Yo'Mind, UMBRAE, and MindMRQ from top to bottom on the right. Bold text highlights key phrases in the comparison.

Table 1: Brain-to-text decoding results on NSD Subjects 1, 2, 5, and 7. Bold and underlined values indicate the best and second-best scores, respectively. See Appendix B.2 for score sources and comparison protocols.

Method	B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr	SPICE
Brain Captioning	54.82	35.14	21.96	14.04	16.24	40.39	38.86	8.84
Takagi et al.	35.22	16.37	7.22	3.30	9.77	24.61	13.52	4.88
MindGPT	56.91	36.95	23.41	15.87	20.04	38.41	43.56	10.97
Neuro2Language	57.19	37.17	23.78	15.85	18.60	36.67	49.51	12.39
UMBRAE	59.09	40.13	27.05	18.40	19.24	<u>43.64</u>	<u>57.76</u>	<u>12.42</u>
Mind-SA	59.76	39.81	27.05	18.86	<u>38.08</u>	42.72	54.03	12.25
Yo'Mind	<u>61.81</u>	<u>42.75</u>	<u>29.76</u>	<u>20.88</u>	38.40	43.63	56.07	<u>12.42</u>
MindMRQ	67.53	50.71	37.24	27.27	22.42	50.16	84.73	15.46

partition. All three settings use 40 region tokens and one global token, as well as the same model architecture, adaptation procedure, and brain-training objective. For the shifted variant, we circularly shift the original content-driven partition by half the image width and height, thereby disrupting the correspondence between the partition layout and the underlying image content.

As shown in Table 2, content-driven partitioning scores higher on all five metrics, improving BLEU-4 by 1.62 and 0.94 and CIDEr by 3.06 and 2.91 over the random and shifted strategies, respectively. Consistent with these gains, Figure 4 shows finer content-driven regions around the surfer, bench, and smaller elephant, with captions that better describe the beach and roadside context and distinguish the baby from the adult elephant. Together,

Table 2: Spatial partitioning strategies. Bold indicates the best result in each column.

Partitioning strategy	B@4	METEOR	ROUGE-L	CIDEr	SPICE
Random quadtree	25.65	22.05	49.48	81.67	15.03
Shifted partition	26.33	22.08	49.80	81.82	15.03
Content-driven (MindMRQ)	27.27	22.42	50.16	84.73	15.46

these results suggest that content-adaptive spatial allocation can improve decoding while preserving object details alongside scene context.

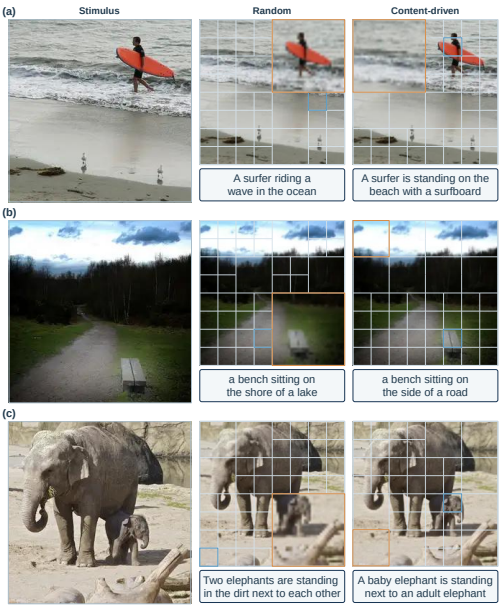

Figure 4: Spatial partitioning and information retention. From left to right: the stimulus and visualizations of random and content-driven partitions. Blue and orange borders highlight selected fine- and coarse-grained regions, respectively. The corresponding brain-decoded captions are shown below.

Spatial granularity. We vary the region-token count K while retaining one global token in every configuration (Table 3). All eight metrics improve as K increases from 10 to 40. The default setting, $K = 40$, achieves the highest scores among the evaluated settings, while $K = 55$ and $K = 61$ yield slightly lower scores. These results suggest that refining coarse spatial partitions improves caption quality, but further refinement does not necessarily yield additional gains.

4.3 Component ablations

We examine representation composition and brain-stage supervision separately. Query-level supervision in Table 5 comprises $\mathcal{L}_{\text{query}}$ and $\mathcal{L}_{\text{contrastive}}$ in Eq. 4. The visual-only variant uses magnitude and global-direction constraints to stabilize decoding (Appendix A.5); per-subject scores for the full model are in Appendix B.1.

Composition of mixed-resolution visual representations. We compare each token-removal variant separately with the full model containing both token types. As shown in Table 4, each variant scores lower on all five reported metrics under its respective target settings in Appendix A.2. These results support retaining both token types in MindMRQ. A possible

Table 3: Effect of spatial granularity on caption quality across region-token counts K. The default setting ($K = 40$) is shown in bold.

K	B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr	SPICE
10	65.83	48.05	34.33	24.42	21.48	48.51	75.65	14.21
16	65.97	48.97	35.39	25.45	21.58	49.12	79.24	14.63
25	66.68	49.43	35.89	25.95	21.90	49.28	81.42	14.95
28	67.03	50.18	36.61	26.62	22.11	49.87	83.31	15.11
40	67.53	50.71	37.24	27.27	22.42	50.16	84.73	15.46
55	67.51	50.64	37.06	26.93	22.26	50.09	84.27	15.36
61	67.30	50.35	36.69	26.56	22.08	49.91	83.93	15.28

Table 4: Representation ablations of the full model. Full-model scores are bold.

Mixed-resolution tokens	Global token	B@4	METEOR	ROUGE-L	CIDEr	SPICE
✓	×	23.77	20.33	46.37	75.62	13.90
×	✓	25.93	21.32	49.24	79.05	14.48
✓	✓	27.27	22.42	50.16	84.73	15.46

explanation is that mixed-resolution tokens retain regional detail and context, while the global token provides a complementary image-level summary.

Table 5: Ablation of dual-level visual-semantic alignment. Bold indicates the best result in each column.

Visual alignment $\mathcal{L}_{\text{visual}}$	Query-level supervision	B@4	METEOR	ROUGE-L	CIDEr	SPICE
✓	×	26.40	21.93	49.68	82.48	15.03
×	✓	26.63	22.19	49.65	82.76	15.27
✓	✓	27.27	22.42	50.16	84.73	15.46

Dual-level visual-semantic alignment. We examine the benefit of aligning both visual representations and query outputs. As shown in Table 5, the full objective achieves higher scores on all five reported metrics than the stabilized visual-only and query-only variants. These results suggest that the two levels provide complementary supervision: token-level visual alignment constrains the predicted mixed-resolution token sets, while query-level supervision constrains the Q-Former outputs used for language generation.

5 Discussion and conclusion

We propose MindMRQ, a brain-to-text decoding framework that connects brain activity to language through mixed-resolution visual representations. By allocating spatial precision according to image content, the framework retains fine detail alongside scene context. Visual encoder adaptation and query tuning make these representations accessible to a pretrained language interface, while dual-level visual-semantic alignment supports their prediction from fMRI. Experiments on NSD demonstrate competitive decoding performance and support content-adaptive partitioning at a moderate spatial granularity. Component ablations further support the combination of mixed-resolution tokens and a global summary token, as well as the accompanying alignment design.

These findings highlight the potential of adapting visual prediction targets to the demands of brain-to-text decoding. However, the partitioning scores reflect image-feature sensitivity to spatial compression rather than individual neural responses. Future work could inves-

tigate how brain signals can inform spatial allocation and whether mixed-resolution visual representations generalize to new subjects and broader decoding settings.

AI use statement

In this work, we used generative AI tools for assisting with translation, improving English expression, and revising manuscript text. We have not used generative AI tools for developing theoretical models or conceptual frameworks, proposing or refining hypotheses, implementing methods, supporting qualitative and thematic data analysis, interpreting results, or designing and providing feedback on research methodology or experiments. In addition, generating synthetic datasets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and cleaning and reformatting datasets are not applicable to this work. We have reviewed all AI-assisted work. In particular, we carefully reviewed the LLM-generated text. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

This study uses the publicly available Natural Scenes Dataset (NSD) and does not involve the collection of new neural recordings. Our analyses focus on describing viewed scenes from the released fMRI responses, rather than identifying participants or inferring sensitive personal attributes. We recognize the privacy implications of brain decoding and emphasize that its applications should comply with dataset usage policies and protect the confidentiality of neural data. The proposed framework is intended for research on visual brain-to-text decoding and should not be used to infer information from individuals' brain activity without their authorization.

References

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience, 25(1):116–126, 2022.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In European conference on computer vision, pp. 382–398. Springer, 2016.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mindgpt: Interpreting what you see with non-invasive brain recordings. IEEE Transactions on Image Processing, 34:3281–3293, 2025a.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Bridging the gap between brain and machine in interpreting visual semantics: Towards self-adaptive brain-to-text decoding. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21938–21948. IEEE, 2025b.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Reducing semantic mismatch in brain-to-text decoding through personalized multimodal masking. In International Conference on Learning Representations, volume 2026, pp. 68777–68793, 2026.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems, 26, 2013.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation, pp. 376–380, 2014.

Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. arXiv preprint arXiv:2305.11560, 2023.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pp. 19730–19742. PmLR, 2023.

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics, pp. 150–157, 2003.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Rex Ying. Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding. arXiv preprint arXiv:2502.15786, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PmLR, 2021.

Tomer Ronen, Omer Levy, and Avram Golbert. Vision transformers with mixed-resolution tokenization. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4613–4622. IEEE, 2023.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. Advances in Neural Information Processing Systems, 36:24705–24728, 2023.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. arXiv preprint arXiv:2403.11207, 2024.

Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. Advances in Neural Information Processing Systems, 37:98083–98110, 2024.

Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. arXiv preprint arXiv:2306.11536, 2023.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4566–4575, 2015.

Haitao Wu, Qirui Zhang, Zhouheng Yao, Shangquan Sun, Qihao Zheng, Mianxin Liu, Chi Zhang, Wanli Ouyang, Chunfeng Song, Changqing Zhang, et al. Brainjanus: A unified model for understanding and generation across brain, vision, and language. arXiv preprint arXiv:2606.30319, 2026.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In European Conference on Computer Vision, pp. 242–259. Springer, 2024.

A Implementation and Training Details

A.1 Partitioning and Region Encoding

The CLIP processor resizes the shorter image side to 224, center-crops to 224×224, and normalizes the input. Quadtree side lengths are 112, 56, and 28, giving $K = 4 + 3t$ after t subdivisions.

Following Ronen et al. (2023), a frozen ShuffleNet-V2-x0.5 scores sensitivity to spatial compression. For the three region sizes, the full image is resized to 64, 128, or 224 pixels per side and restored to 224 using bilinear interpolation. Squared feature differences are averaged across channels, upsampled, and averaged within each candidate region. The random policy assigns reproducible priorities by image and region. The shifted policy circularly shifts the content-driven partition by 112 pixels on both axes, corresponding to half the image width and height. This is implemented by applying the same shift to the score maps before subdivision, which yields an equivalent shifted partition under the present quadtree configuration.

CLIP provides a learned positional embedding at each location of its 7×7 patch grid. For each region, we obtain its positional embedding by bilinearly interpolating these embeddings at the region center. We also use a learnable scale embedding indexed by $\ell = \log_2(112/s)$, where s is the region's side length before resampling. Thus, regions with side lengths of 112, 56, and 28 pixels use scale indices 0, 1, and 2, respectively. All scale embeddings are initialized to zero and updated during visual adaptation.

A.2 Visual Representations and Brain Encoder

Separate projection layers map the mixed-resolution ViT's region-token and class-token states from 768 to 1408 dimensions, matching the visual input dimension of the frozen BLIP-2 Q-Former. The native BLIP-2 target has 256 patch tokens and one class token from its final visual layer. Its OPT-2.7B interface uses 32 queries and produces 768-dimensional query outputs, which a frozen language projection maps to a 32×2560 continuous prefix for OPT-2.7B.

The randomly initialized brain encoder follows UMBRAE's design (Xia et al., 2024). Subject-specific affine mappings, voxel projections, and LayerNorm produce a 256×1024 input sequence. Six shared layers use $K + 1$ learned latents of width 1024, 16 attention heads of width 96, and feedforward expansion four with GELU. Latents attend to the concatenation of input and latent features. Separate prediction heads output the global token and the region tokens in 1408 dimensions.

For each brain-predicted region token, we initialize a reference box $\mathbf{c}_k^0 = (x_k^0, y_k^0, w_k^0, h_k^0)$, containing its normalized center coordinates, width, and height. The geometry head learns to refine this initial box from the corresponding brain-encoder feature $\mathbf{h}_k$. Geometry is predicted as $\widehat{\mathbf{c}}_k = \sigma(\mathrm{logit}(\mathbf{c}_k^0) + W_c\mathbf{h}_k + \mathbf{b}_c)$, with a zero-initialized residual head. Initial centers are the first K entries of a $\lceil\sqrt{K}\rceil \times \lceil\sqrt{K}\rceil$ grid.

Representation ablations. For Table 4, the variant without the global token retains only the mixed-resolution tokens in both branches and recomputes the query-output targets from the visual encoder. The global-only brain branch predicts only the global token while retaining query-output targets from the complete mixed-resolution visual representations.

A.3 Objectives and Matching

Query alignment. For two sets of Q-Former output representations $U, V \in \mathbb{R}^{B\times m\times d}$, bars denote ℓ_2 normalization along the feature dimension, and ρ is Smooth L1 with threshold

one. Define

$$D_\beta(\mathbf{u}, \mathbf{v}) = 1 - \bar{\mathbf{u}}^\top \bar{\mathbf{v}} + \frac{\beta}{d} \sum_c \rho(\bar{u}_c - \bar{v}_c),$$
$$\mathcal{A}_\beta(U, V) = \frac{1}{Bm} \sum_{i,j} D_\beta(\mathbf{u}_{ij}, \mathbf{v}_{ij}). \tag{A1}$$

During brain training, $U = \widehat{Z}$ and $V = Z$ are the Q-Former outputs obtained from brain-predicted and image-derived mixed-resolution visual representations, respectively, using the same queries $Q^\star$. We use $\mathcal{L}_{\text{query}} = \mathcal{A}_{0.25}(\widehat{Z}, Z)$ with $m = 32$ and $d = 768$. The image-derived outputs Z serve as fixed targets. Gradients propagate through $\widehat{Z}$ to update the brain encoder, while the Q-Former parameters remain frozen.

Image-level contrast. Mean-pool and normalize each query set to obtain $\mathbf{z}_i(U)$. Define $S_{ik} = \mathbf{z}_i(U)^\top \mathbf{z}_k(V)/\tau$ with temperature $\tau = 0.07$. Symmetric InfoNCE is

$$\mathcal{N}(U, V) = -\frac{1}{2B} \sum_i \left[\log \frac{e^{S_{ii}}}{\sum_k e^{S_{ik}}} + \log \frac{e^{S_{ii}}}{\sum_k e^{S_{ki}}} \right]. \tag{A2}$$

Thus $\mathcal{L}_{\text{contrastive}} = \mathcal{N}(\widehat{Z}, Z)$. Negatives come from the forward batches in Table A1; gradient accumulation does not enlarge this pool.

Visual representation learning. Using BLIP-2's original queries Q_0, let $Z^0 = \mathcal{Q}_\psi(Q_0; M)$ and $Z^{\text{native}} = \mathcal{Q}_\psi(Q_0; M^{\text{native}})$. Visual representation learning minimizes

$$\mathcal{L}_{\text{adapt}} = 2\mathcal{A}_{0.25}(Z^0, Z^{\text{native}}) + 0.1\mathcal{N}(Z^0, Z^{\text{native}}) + \frac{0.25}{B} \sum_i \left[1 - \cos(\mathbf{g}_i, \mathbf{g}_i^{\text{native}})\right], \tag{A3}$$

where $\mathbf{g}^{\text{native}}$ is the global [CLS] token from the final layer of the frozen native BLIP-2 visual encoder.

Query tuning. We initialize the query inputs Q from Q_0 and optimize only Q using the reference-caption negative log-likelihood:

$$\mathbf{P}_i(Q) = P_\omega[\mathcal{Q}_\psi(Q; M_i)],$$
$$\mathcal{L}_{\text{cap}}(Q) = -\frac{1}{\sum_{i=1}^B T_i} \sum_{i=1}^B \sum_{t=1}^{T_i} \log p_\eta(y_{i,t} \mid \mathbf{P}_i(Q), y_{\text{BOS}}, y_{i,<t}). \tag{A4}$$

Here, M_i denotes the mixed-resolution visual representations of image i, P_ω is the frozen language projection, and p_η is the frozen OPT-2.7B language model. The continuous prefix $\mathbf{P}_i(Q)$ conditions caption generation, and $y_{i,<t}$ denotes the preceding reference tokens. Captions are truncated to 25 tokens before appending the termination token; T_i counts the resulting target tokens.

Mixed-resolution token matching. For Equation 2, the transport plan $\Pi_i \in \mathbb{R}_+^{K \times K}$ assigns mass from predicted to target region tokens, excluding the global token. The inner product $\langle \Pi_i, D_i \rangle = \sum_{a,b} \Pi_{i,ab} D_{i,ab}$ aggregates the pairwise costs according to this plan. We obtain Π_i by approximately solving the entropy-regularized OT problem

$$\min_{\Pi \in \mathcal{U}_K} [\langle \Pi, D_i \rangle - \varepsilon H(\Pi)], \qquad H(\Pi) = -\sum_{a,b} \Pi_{ab} \log \Pi_{ab},$$

where $\mathcal{U}_K$ is the set of nonnegative matrices whose rows and columns each sum to $1/K$, and $\varepsilon > 0$ controls the strength of entropy regularization. The Sinkhorn algorithm (Cuturi, 2013) alternately rescales the rows and columns of the elementwise exponential kernel $\exp(-D_i/\varepsilon)$ to match these prescribed marginals. We set $\lambda_{\text{geo}} = 0.25$ and $\varepsilon = 0.08$, and approximate the transport plan using 30 log-domain Sinkhorn iterations. Gradients propagate through both the cost matrix D_i and the transport plan Π_i via the unrolled Sinkhorn iterations.

Table A1: Training settings.

Setting	Visual representation learning	Query tuning	Brain training
Learning rate	2×10^{-5}	10^{-4}	2×10^{-4}
Weight decay	0.01	0	5×10^{-4}
Forward / effective batch	16 / 16	16 / 64	32 / 128
Epochs / evaluated step	15 epochs	5 epochs	EMA step 9,000
Schedule	Constant	5% warmup, then cosine	2,000-step warmup, then constant

A.4 Optimization and Generation

All stages use AdamW and gradient-norm clipping at 1.0 (Table A1). Brain loss weights are $\lambda_{\mathrm{q}} = 1.2$ and $\lambda_{\mathrm{c}} = 0.1$, with a unit weight for $\mathcal{L}_{\mathrm{visual}}$.

Brain evaluation uses exponential moving average (EMA) parameters with decay 0.999. Generation uses an empty prompt, three beams, length penalty 0.7, no repeated trigrams, and at most 25 new tokens, with sampling disabled.

A.5 Alignment Ablation Settings

Query-level supervision combines $\mathcal{L}_{\mathrm{query}}$ and $\mathcal{L}_{\mathrm{contrastive}}$; these terms are enabled or removed jointly in Table 5. The visual-only variant sets $\lambda_{\mathrm{q}} = \lambda_{\mathrm{c}} = 0$; the fixed queries $Q^{\star}$ are still used for language readout.

With geometry fixed, the cosine feature cost in Eq. 2 is invariant to positive rescaling of predicted region tokens, leaving their magnitudes unconstrained. Moreover, the global token is excluded from OT matching. In our experiments, removing query-level supervision without further constraints led to decoding collapse. We therefore use a stabilized visual-only variant (the first row of Table 5) with magnitude and global-direction constraints.

For $X \in \mathbb{R}^{n \times 1408}$, define the whole-matrix RMS $r(X) = \max(\|X\|_F / \sqrt{1408n}, 10^{-6})$ and $\mathcal{S}(X, Y) = \rho(\log r(X) - \log r(Y))$. With V denoting the region-token representations, global vectors treated as one-row matrices, and D_β from Equation A1 evaluated at $d = 1408$, the objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{scale}} &= \frac{1}{2B} \sum_i \left[\mathcal{S}(\widehat{V}_i, V_i) + \mathcal{S}(\widehat{\mathbf{g}}_i, \mathbf{g}_i) \right], \\ \mathcal{L} &= \mathcal{L}_{\mathrm{visual}} + \mathcal{L}_{\mathrm{scale}} + \frac{0.1}{B} \sum_i D_{0.1}(\widehat{\mathbf{g}}_i, \mathbf{g}_i). \end{aligned} \tag{A5}$$

B Supplementary Results

B.1 Per-Subject Results

Table B1 gives the subject scores underlying the main-model macro-average.

Table B1: MindMRQ results by subject.

Subject	B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr	SPICE
1	67.74	51.10	37.43	27.25	22.76	50.47	86.18	15.82
2	66.29	49.57	36.11	26.42	21.84	49.37	81.35	14.86
5	69.33	52.68	39.09	28.65	23.40	51.42	91.45	16.61
7	66.77	49.48	36.32	26.79	21.68	49.39	79.92	14.55

B.2 Published Results and Comparison Protocols

Table 1 retains the published COCO-reference values collated by Chen et al. (2026).

Table B2 reports BrainJanus's original GT=COCO results (Wu et al., 2026). Its test-image set has not been harmonized with our four-subject protocol, so it is kept separate from the ranking in Table 1.

Table B2: Published BrainJanus results under its original GT=COCO protocol; the original ROUGE label (R) is retained.

Method	B@1	B@2	B@3	B@4	METEOR	R	CIDEr	SPICE
BrainJanus	63.20	44.10	31.25	22.45	20.21	46.91	62.37	13.10