CorText-AMA: brain-language fusion as a new tool for probing visually evoked brain responses

Victoria Bosch
victoria.bosch@uni-osnabrueck.de
Institute of Cognitive Science, University of Osnabrück
Osnabrück, NI 49090, Germany

Daniel Anthes
Institute of Cognitive Science, Osnabrück University
Osnabrück, NI 49090, Germany

Adrien Doerig
Department of Psychology and Education, Freie Universität Berlin
Berlin, Germany
Institute of Cognitive Science, University of Osnabrück
Osnabrück, NI 49090, Germany

Sushrut Thorat
Institute of Cognitive Science, Osnabrück University
Osnabrück, NI 49090, Germany

Peter König
Institute of Cognitive Science, Osnabrück University
Osnabrück, NI 49090, Germany

Tim C Kietzmann
tim.kietzmann@uni-osnabrueck.de
Institute of Cognitive Science, University of Osnabrück
Osnabrück, NI 49090, Germany

Abstract

Cognitive computational neuroscience embraces machine learning techniques to gain insight into how the brain represents and transforms visual information. Recent advances have allowed the field to move from classic category inventory approaches to more contextualized, semantic aspects, e.g. by mapping visual responses to natural scenes to corresponding language embeddings of scene captions. While the latter is powerful, single embedding vectors or captions may not fully capture the distributed cortical feature selectivity or complex spatial and semantic interactions in natural scenes. To go beyond passive representation analysis and develop interactive approaches to neural data interpretation, we extend large language models to combine a natural language interface with brain data. The resulting framework, CorText-AMA, provides an interactive chat interface that enables researchers to interrogate neural representations of natural scenes. This approach preserves semantic context while simultaneously allowing us to isolate and examine specific dimensions of brain representations. To make this possible, we combine a transformer-based multimodal model and functional brain alignment with a large instruction-finetuning dataset of question-answer pairs defined on natural scenes. The current model enables flexible probing of decodable information in visual cortex and outperforms control models. Future work will further investigate the usage of CorText-AMA as an interactive diagnostic readout that allows contrasting which questions can be answered based on neural responses in specific brain regions, and which cannot.

Keywords: neural decoding; question-answering; vision; transformer; scene perception; alignment; semantics

Results

Here, we present *CorText-AMA*, a novel end-to-end trainable multimodal model, combining fMRI data and language for flexible neural decoding and diagnostic probing of visually evoked neural responses to complex scenes. By implementing an intermediate fusion approach that combines brain and language embeddings, we enable conditional generation where neural "context" directly guides answer generation. We first parcellate the neural responses of visual cortex to complex scenes, and embed the responses from the resulting 82 regions with region-specific linear encoders. Together with the question word embeddings, these are fed as a multimodal input token sequence to the language decoder (Llama 2), which autoregressively continues the sequence by answering the question. The model is trained using the image captions and instruction question-answering pairs available for the stimuli (X. Chen et al., 2015; Liu, Li, Wu, & Lee, 2023) (Fig. 1A; see Methods). *Cortext-AMA* is trained using the Natural Scenes Dataset (NSD), a 7T fMRI dataset of neural responses to images of complex natural scenes (Allen et al., 2022). To address the data-hungry nature of transformers, we make use of Shared Response Modeling (SRM), a functional alignment technique that maps individual subjects' neural data into a shared, lower-dimensional representational space that generalizes to unpaired data (P. C. Chen et al., 2015). This approach effectively expands our available training data by aligning unique data from all 8 subjects (Fig. 1B). As control models, we train one model per subject, which results in decoding performance (CLIPScore) of captions and answers above the performance of a shuffled control (8 models; $\mu_{cap} = 0.51$, $\mu_{q\&a} = 0.72$ Fig. C), and a model on unaligned aggregated data from all 8 subjects ($\mu_{cap} = 0.49$, $\mu_{q\&a} = 0.70$). The multisubject (SRM) model, with eight-fold multiplication of training data, yields significant improvements in caption decoding performance and question-answer capacity compared to the control models ($\mu_{cap} = 0.54$, individual vs aligned: $p < 0.01$, unaligned vs aligned: $p < 0.001$; $\mu_{q\&a} = 0.83$, individual vs aligned: $p < 0.001$, unaligned vs aligned: $p < 0.001$; independent permutation tests, n=10000; Fig. 1C, see 1D for examples). This demonstrates that by using functional alignment, neural decoding performance might be further improved without having to collect more data for individual subjects.

Discussion

Semantic decoding of neural content has significantly advanced in the last few years, with new machine learning techniques fueling its development. Notably, large language models offer a promising new tool that can be used to decode semantic embeddings and image captions from neural data (Doerig et al., 2022; Zhang, Han, Worth, & Liu, 2020; Ferrante, Ozcelik, Boccato, VanRullen, & Toschi, 2023; Luo, Henderson, Tarr, & Wehbe, 2023; Matsuyama, Nishimoto, & Takagi, 2025; Scotti et al., 2024; Bosch et al., 2024). While few studies on brain-language fusion for question-answering have been presented recently (Huang, Ma, Xie, & Wang, 2025; J. Chen, Qi, Wang, & Pan, 2023; Qiu et al., 2025), our work makes several contributions to this emerging topic. First, we provide an end-to-end training pipeline that neither includes access to the underlying images during training, nor pre-trained models such as CLIP (Radford et al., 2021) that have seen the underlying stimulus materials. This reduces the risk of the model memorizing a simple look-up table. Second, by conceptualizing the brain as a non-linear "image embedder" and implementing fully linear brain data encoders, we enable encoding reversal back into vertex space and allow for explicit testing of linear decodability. Lastly, we also show that aligning neural data of multiple participants increases performance, possibly indicating a data deficiency that needs to be overcome for training large-scale neural decoding models (Banville, Benchetrit, d'Ascoli, Rapin, & King, 2025). Together, these results reveal the potential of *CorText-AMA* as an interactive diagnostic tool that enables targeted probing of neural data.

Methods

Dataset The Natural Scenes Dataset (NSD) contains 7T fMRI measurements of 8 participants who have each viewed

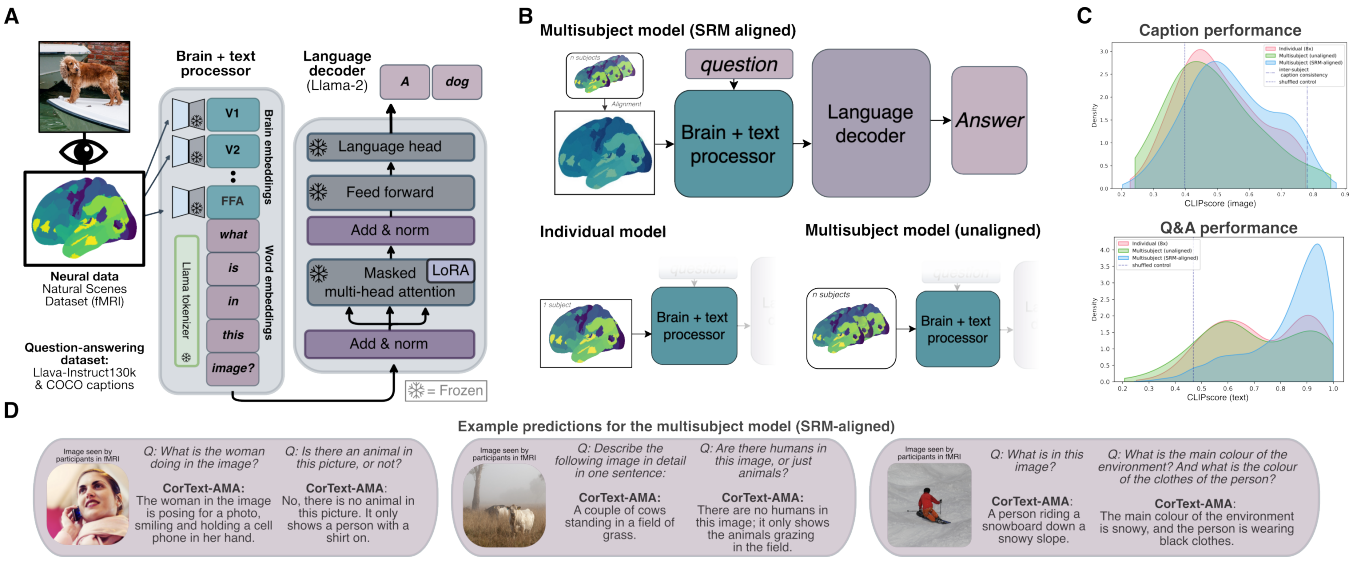

Figure 1: **A - The CorText-AMA architecture:** *The fMRI data of visual cortex is parcellated and linearly embedded into 4096-dimensional embeddings to match the input word embedding dimensionality. Together with the question word embeddings, these are fed as a sequence to the language decoder, which autoregressively generates an answer.* **B - Model input data:** *We compare three approaches to neural data input: individual models (n=8) trained on data from single subjects, a model trained on 'naively' aggregated unaligned data from all subjects, and one model trained on functionally aligned data from all subjects using Shared Response Modeling (SRM).* **C - Performance evaluation:** *Caption generation and question-answering performance across model types demonstrates that the SRM-aligned model outperforms controls.* **D - Example predictions:** *Answers generated by the multisubject SRM-aligned model in response to human questions about image-evoked neural activity.*

9000 unique images and up to 1000 shared images from the MS COCO dataset (Lin et al., 2014). We use the beta values of the 1.8-mm volume preparation in fsaverage space. We parcellate the visual cortex into 82 regions of interest (ROI) per the HCP-MMP1 atlas (Glasser et al., 2016). Each image in NSD has five human captions from MS COCO, which we use to construct question-answers pairs asking e.g. "*Describe the following image*". In addition, each image has up to 5 question-answer pairs from the Llava-Instruct-150k visual question-answering dataset for multimodal instruction-following (Liu et al., 2023). We use the unique trials as training data, and the remaining shared trials as test dataset.

Shared Response Modeling The neural data for 8 subjects is aligned using SRM, which is fit using the first 515 shared test images of each subject. The number of features for the SRM fit is determined by performing PCA over the train data per ROI and determining how many components are required to capture 99% of variance in that region. The learnt fit is then used to transform all remaining data into the shared subject space.

Architecture Cortext-AMA is a multimodal decoder-only transformer-based architecture, using Llama 2 (7B-instruct), an autoregressive causal language model, as backbone (Touvron et al., 2023). To enable multimodal fusion between neural data and language, we use 82 linear encoders, one for each ROI in visual cortex. To reduce the encoder size, we use a low-rank linear projection to embed neural data of each ROI. We identify the number of principal components required to capture 95% of variance in the original 4096-dimensional embeddings of all captions in the train set of subject 1. We project the data of each ROI in a linear layer of that size (921 components). We retain the PCA projection and use its transpose as a frozen intermediate linear layer to upscale the brain embedding to the word embedding space dimensionality. The neural data for each trial is embedded by the ROI encoders and concatenated with the Llama tokenizer-embedded question for that trial, resulting in a multimodal input sequence of brain and text embeddings.

Training Model training consists of two phases (20 epochs each): during pre-training, only the brain encoders and layer normalization throughout the Llama decoder are trainable, effectively training a brain-tokenizer compatible with the language decoder. For finetuning, we employ Quantized Low-Rank Adaptation (QLoRA; (Dettmers, Pagnoni, Holtzman, & Zettlemoyer, 2023)) to adapt the Q and V projection matrices of the decoder, while keeping the original decoder frozen and quantized for memory efficiency. We use QLoRA with a rank of 16, α = 16, and dropout of 5e-2. The model minimises cross-entropy between generated and true answers for each trial. Models are pre-trained with AdamW-8bit and a batch size of 25, a learning rate of 1e-3, a cosine lr scheduler, and L2 encoder weight regularization (2e-1). Finetuning continues with a reduced learning rate (2e-5) and L2 (5e-4).

Metrics To evaluate the quality of the answers generated from the neural data test set, we assess two tasks. We measure the ability of the model to generate scene captions (responses to questions such as "describe the content of this image") using CLIPScore (Hessel, Holtzman, Forbes, Bras, & Choi, 2021) to evaluate the correspondence of the generated caption with the stimulus image. The ceiling is set by the average CLIPScore of all MS COCO human captions. To evaluate the answers generated with the Llava-Instruct130k data, we use RefCLIPscore to capture semantic correspondences with ground truth answers, dealing with answers that differ syntactically but have similar meaning. Lower bounds reflect the correspondence between shuffled generated and true answers.

Acknowledgments

The authors acknowledge support by the ERC stg grant 101039524 TIME (Bosch, Doerig, Kietzmann), SNF grant n.203018 (Doerig), RTG GRK2340 DFG (Anthes). Compute resources for this project are in part funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 456666331.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Banville, H., Benchetrit, Y., d'Ascoli, S., Rapin, J., & King, J.-R. (2025). Scaling laws for decoding images from brain activity. *arXiv preprint arXiv:2501.15322*.

Bosch, V., Gutlin, D., Doerig, A., Anthes, D., Thorat, S., Konig, P., & Kietzmann, T. C. (2024). Cortext: large language models for cross-modal transformations from visually evoked brain responses to text captions. *Cognitive Computational Neuroscience*.

Chen, J., Qi, Y., Wang, Y., & Pan, G. (2023). Mindgpt: Interpreting what you see with non-invasive brain recordings. *arXiv preprint arXiv:2309.15729*.

Chen, P. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A reduced-dimension fmri shared response model. *Advances in neural information processing systems*, *28*.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, *36*, 10088–10115.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*.

Ferrante, M., Ozcelik, F., Boccato, T., VanRullen, R., & Toschi, N. (2023). Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., ... others (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., & Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Huang, W., Ma, K., Xie, T., & Wang, H. (2025). Brainchat: Interactive semantic information decoding from fmri using large-scale vision-language pretrained models. In *Icassp 2025-2025 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, *36*, 34892–34916.

Luo, A. F., Henderson, M. M., Tarr, M. J., & Wehbe, L. (2023). Brainscuba: Fine-grained natural language captions of visual cortex selectivity. *arXiv preprint arXiv:2310.04420*.

Matsuyama, T., Nishimoto, S., & Takagi, Y. (2025). Lavca: Llm-assisted visual cortex captioning. *arXiv preprint arXiv:2502.13606*.

Qiu, W., Huang, Z., Hu, H., Feng, A., Yan, Y., & Ying, R. (2025). Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding. *arXiv preprint arXiv:2502.15786*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Scotti, P. S., Tripathy, M., Villanueva, C. K. T., Kneeland, R., Chen, T., Narang, A., ... others (2024). Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhang, Y., Han, K., Worth, R., & Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, *11*(1), 1877.