Mapping intracranial EEG to semantic embedding spaces with a deep transformer decoder

William L. Gross, MD, PhD[1,2,3,4] Alex J. Skitowski[2] Sunil Mathew, PhD[5] Dewan I. Ahmed[3] Hernan G. Rey, PhD[3,5] Kunal Gupta, MD, PhD[5] Andrew J. Anderson, PhD[3,4,5]

Departments of [1]Anesthesiology, [2]Biophysics, [4]Neurology, and [5]Neurosurgery, Medical College of Wisconsin

[3]Joint Department of Biomedical Engineering, Medical College of Wisconsin and Marquette University; Milwaukee, Wisconsin, USA

Why transformers?

Semantic decoding predominantly has used linear [1,2] and convolutional [3] semantic decoding models. Transformers are able to efficiently integrate information across time and electrodes. Does combining sparse responses across **time and electrodes** improve concept decoding of viewed images?

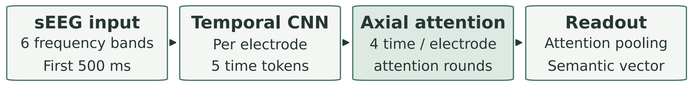
Task images



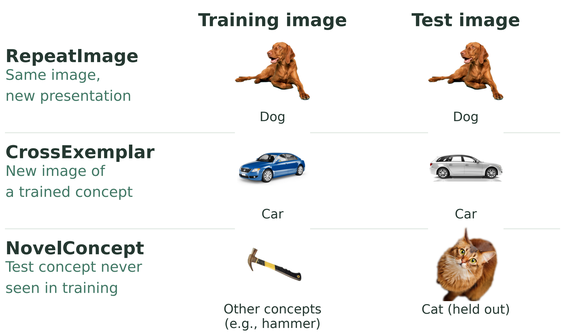
Two visual exemplars each of truck, hammer, and dog. Images: iStockPhoto.com, adapted for the task.

Single-trial sEEG to semantic vectors

Five participants; 20 concrete concepts; multiple image exemplars. Images shown every 500ms in rapid succession with a category judgment task. A shared temporal CNN and axial attention map the first 500 ms of band activity to a 3,072-dimensional embedding (*text-embedding-3-large*).

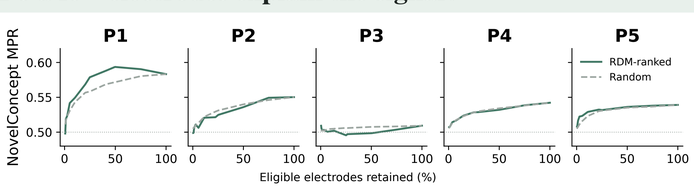


Three levels of generalization



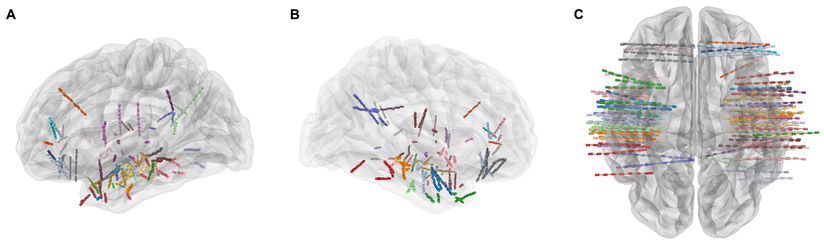
Illustrative task images; assignments vary by split. Images: iStockPhoto.com, adapted for the task. Mean percentile rank (MPR): higher is better; random ranking = 0.5. Matched controls refit models with shuffled training targets.

Do a few electrodes explain the signal?



Separate ridge / NovelConcept benchmark. Solid: training-only RDM ranking; dashed: matched random subsets. Sparse subsets generally underperform larger subsets; ranking has no universal advantage.

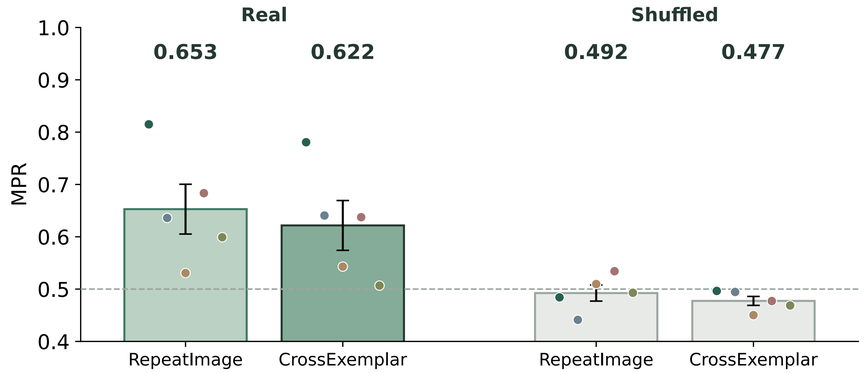
Recording coverage



sEEG electrode locations in MNI space: **A**, left lateral; **B**, right lateral; **C**, inferior view.

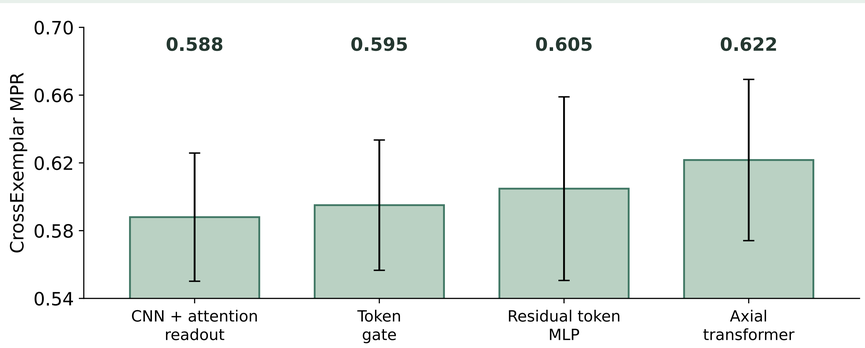
The transformer learns concept-discriminating responses that transfer across pictures. Unrestricted semantic generalization remains the next challenge.

Cross-exemplar decoding survives shuffling controls



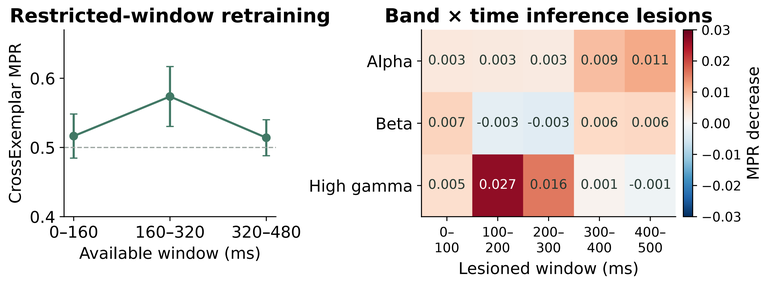
Bars: mean ± SEM; colored dots: five participant means, each averaged across three seeds. Dashed line: chance (0.5). All five participant means for CrossExemplar exceed their shuffle means.

What does each model add?



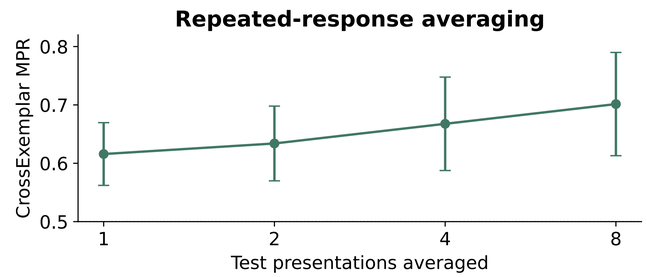
Bars: mean ± SEM across five participants. Displayed MPR range: 0.54–0.70. CNN/readout: local features + pooling. Gate: independent feature gating. MLP: independent token transformations. Axial: repeated time/electrode interactions. All retain readout attention. Axial has the highest CrossExemplar mean: +0.034 over CNN/readout and +0.017 over MLP. The advantage varies across participants.

When is the useful information?



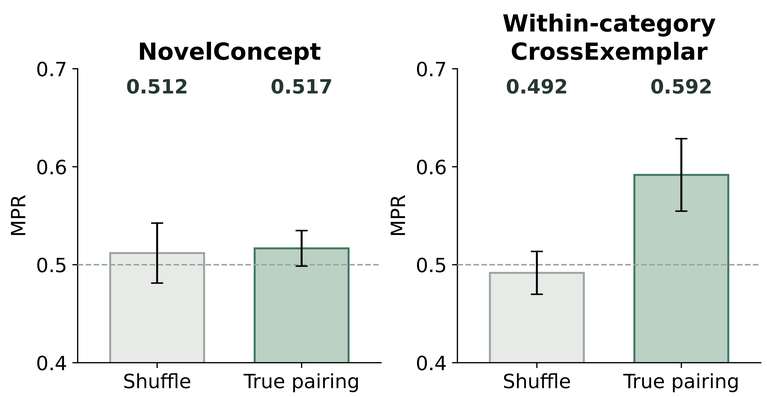
Left: window retraining, mean ± SEM ($n = 5$). Right: lesion-induced MPR loss. 160–320 ms supports the strongest window-only decoding. High gamma at 100–300 ms produces the largest tested lesion effects.

More reliable predictions through averaging



Averaging predictions across repeated presentations of a fixed set of held-out images. Points: mean ± SEM across five participants. Averaging $1 \rightarrow 8$ test presentations increases MPR from 0.616 $\rightarrow$ 0.701; gains vary across participants.

Where does generalization stop?



Bars: mean ± SEM across five participants. Displayed MPR range: 0.40–0.70. Different candidate sets: NovelConcept uses outside-task foils; within-category ranking uses familiar concepts from the same broad category.

Selected references

[1] Pereira et al. *Nat Commun* (2018). doi:10.1038/s41467-018-03068-4.
[2] Tang et al. *Nat Neurosci* (2023). doi:10.1038/s41593-023-01304-9.
[3] Défossez et al. *Nat Mach Intell* (2023). doi:10.1038/s42256-023-00714-5.

Supported by NIH/NIDCD R21DC021013 (William L. Gross).