Linguistic and visual similarity judgements predict EEG representational dynamics in visual perception and sentence reading

Katerina Marie Simkova[a], Jasper J. F. van den Bosch[b], Clayton Hickey[a], Ian Charest[c]

CHBH, School of Psychology, University of Birmingham[a], School of Psychology, University of Leeds[b], cerebrUM, Département de Psychologie, Université de Montréal[c]
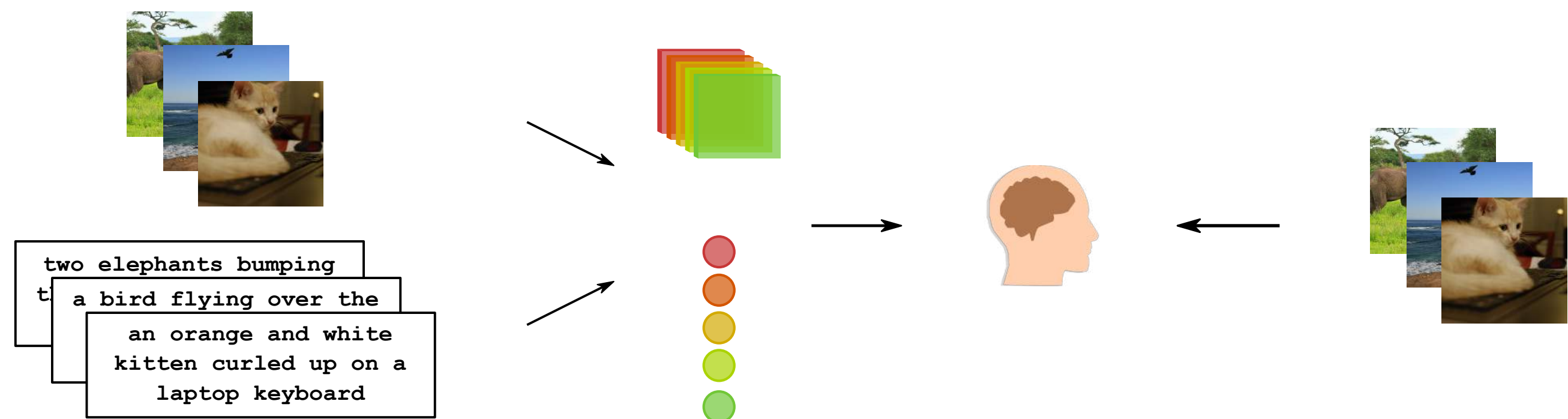
contact: km.simkova@outlook.com

Background

Computational models jointly trained on text and images outperform unimodal models at predicting visual response patterns in the human brain[1,2,3,4].


Evidence from behavioural[5] and neuroimaging[6] studies demonstrates a close relationship between visual and linguistic representations.

Response times in a cross-modal priming experiment predicted by prime-target dissimilarity
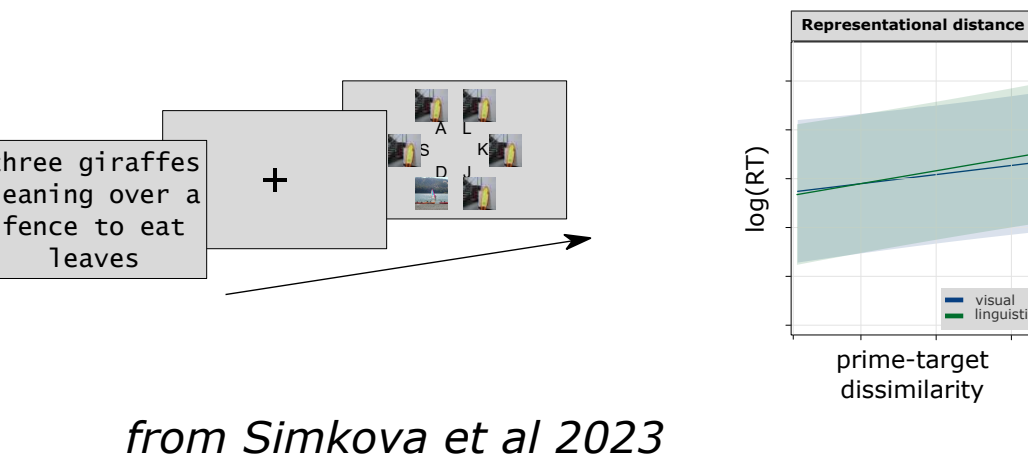
from Simkova et al 2023

NSD fMRI[7] predicted by a) visual similarity judgements*, b) linguistic similarity judgements** in multiple arrangement tasks
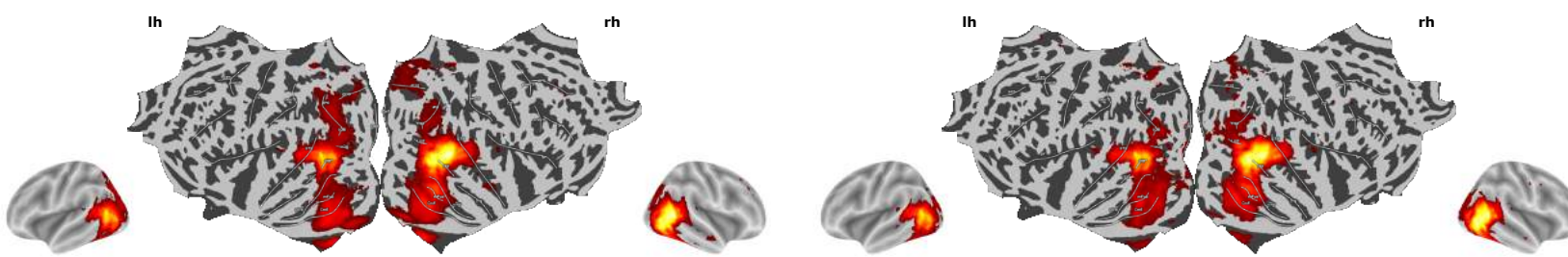
from Simkova, Hickey, & Charest (in prep)

How do linguistic representations that are traditionally considered to be symbolic generalise to depictive visual representations?
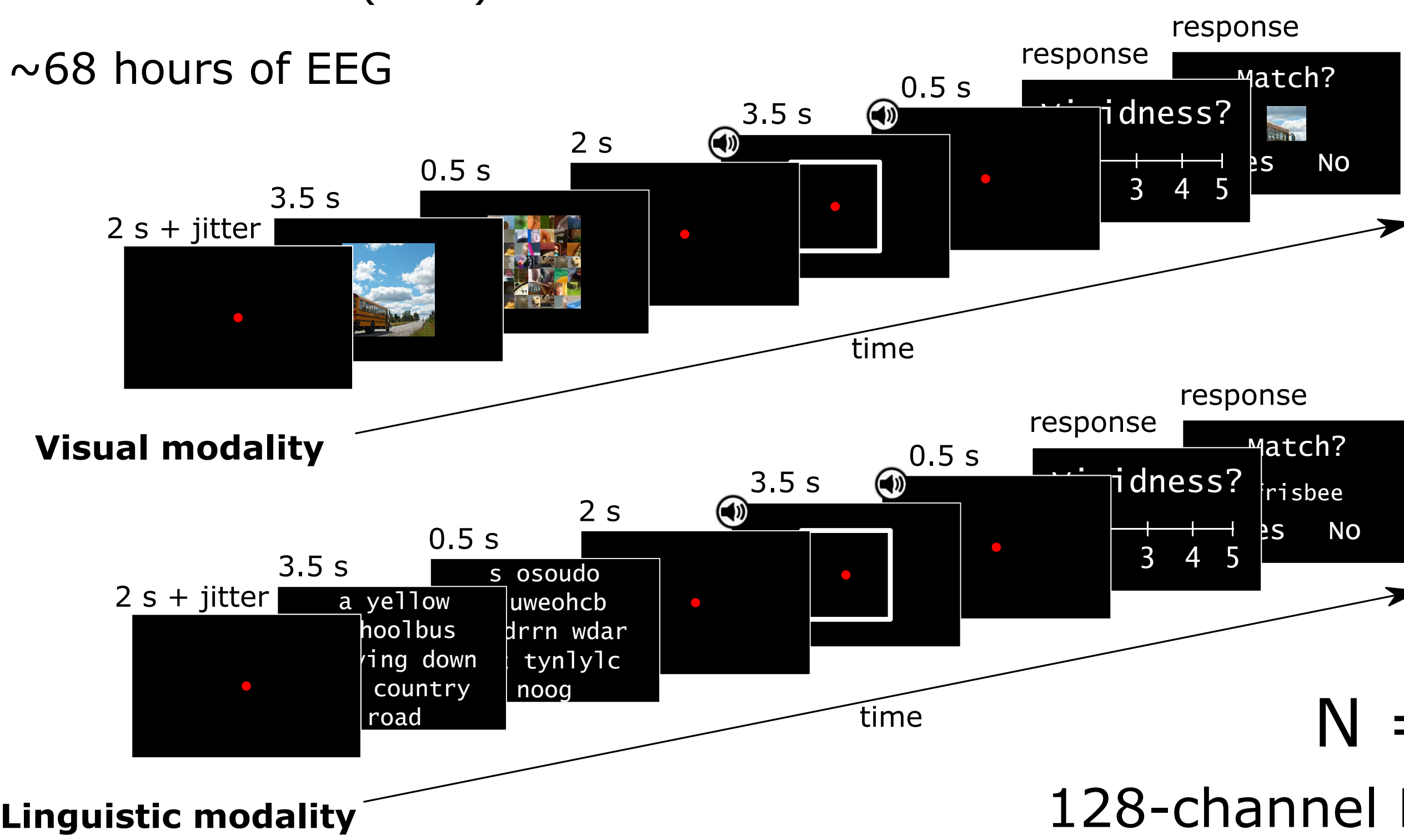
What is the precise representational format of shared representations in vision and language?

Methods

Multimodal mental imagery task to explore natural mental representations associated with vision and language

100 natural scene images and sentence captions from Natural Scenes Dataset (NSD)[7]
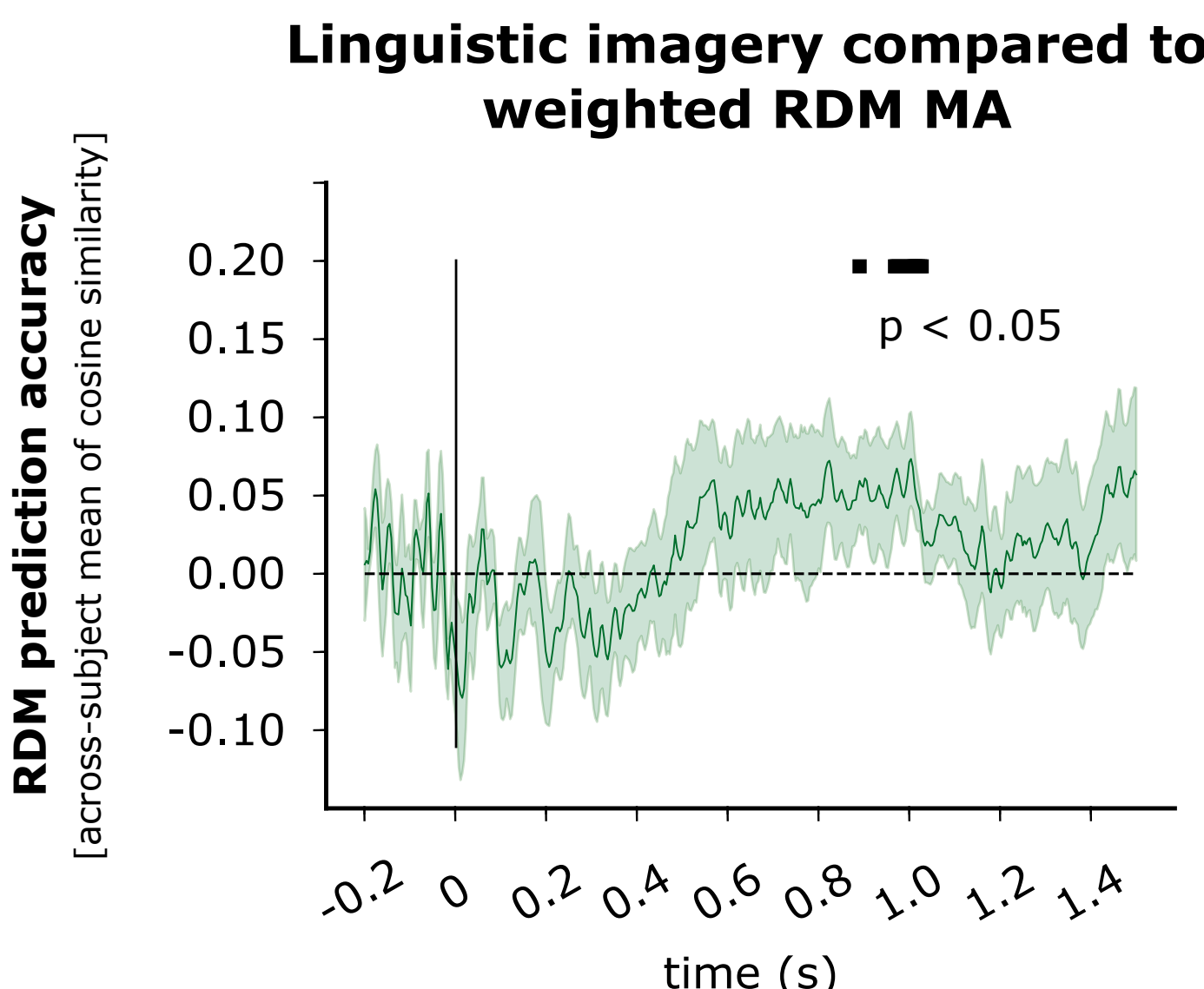
~68 hours of EEG

Visual modality
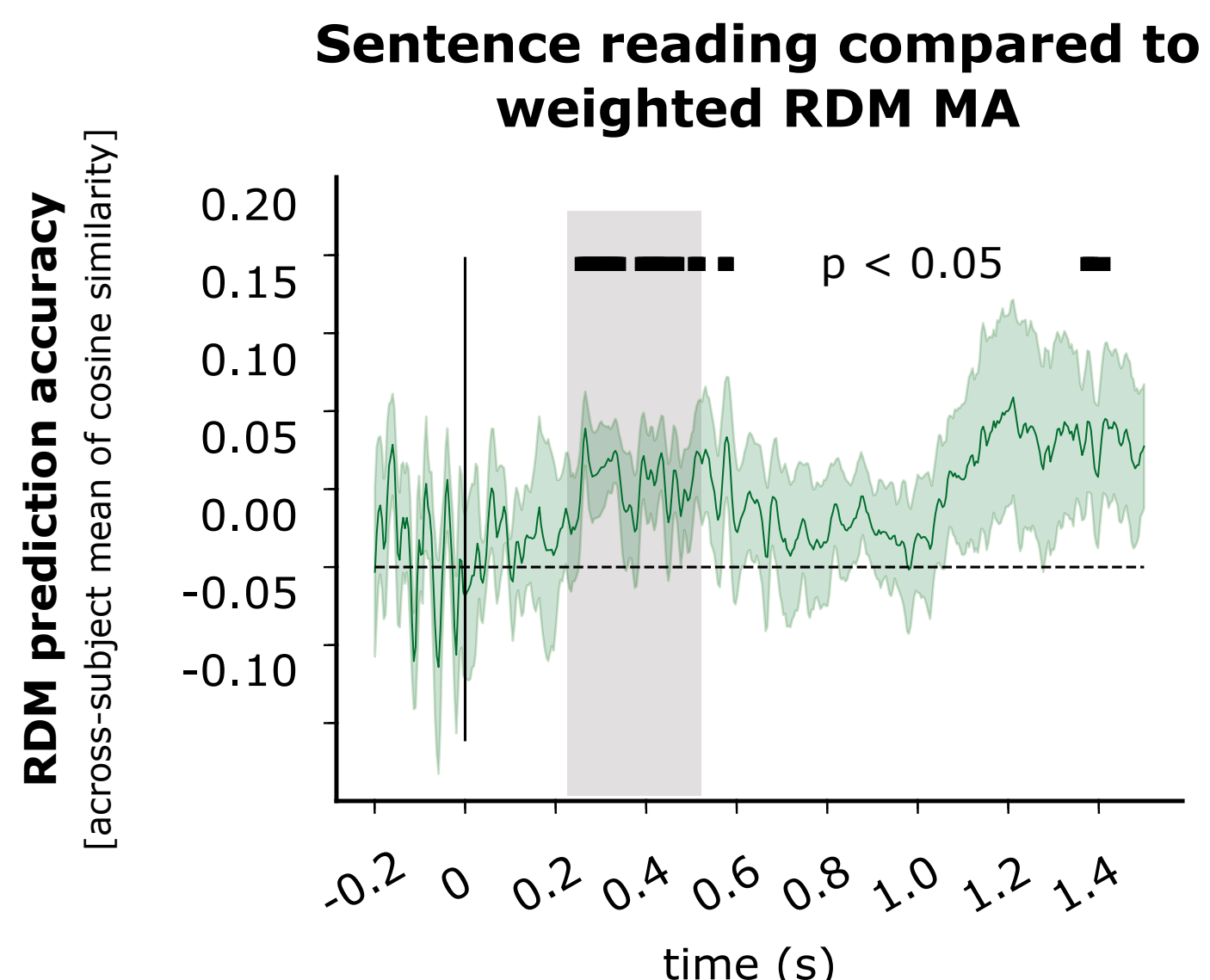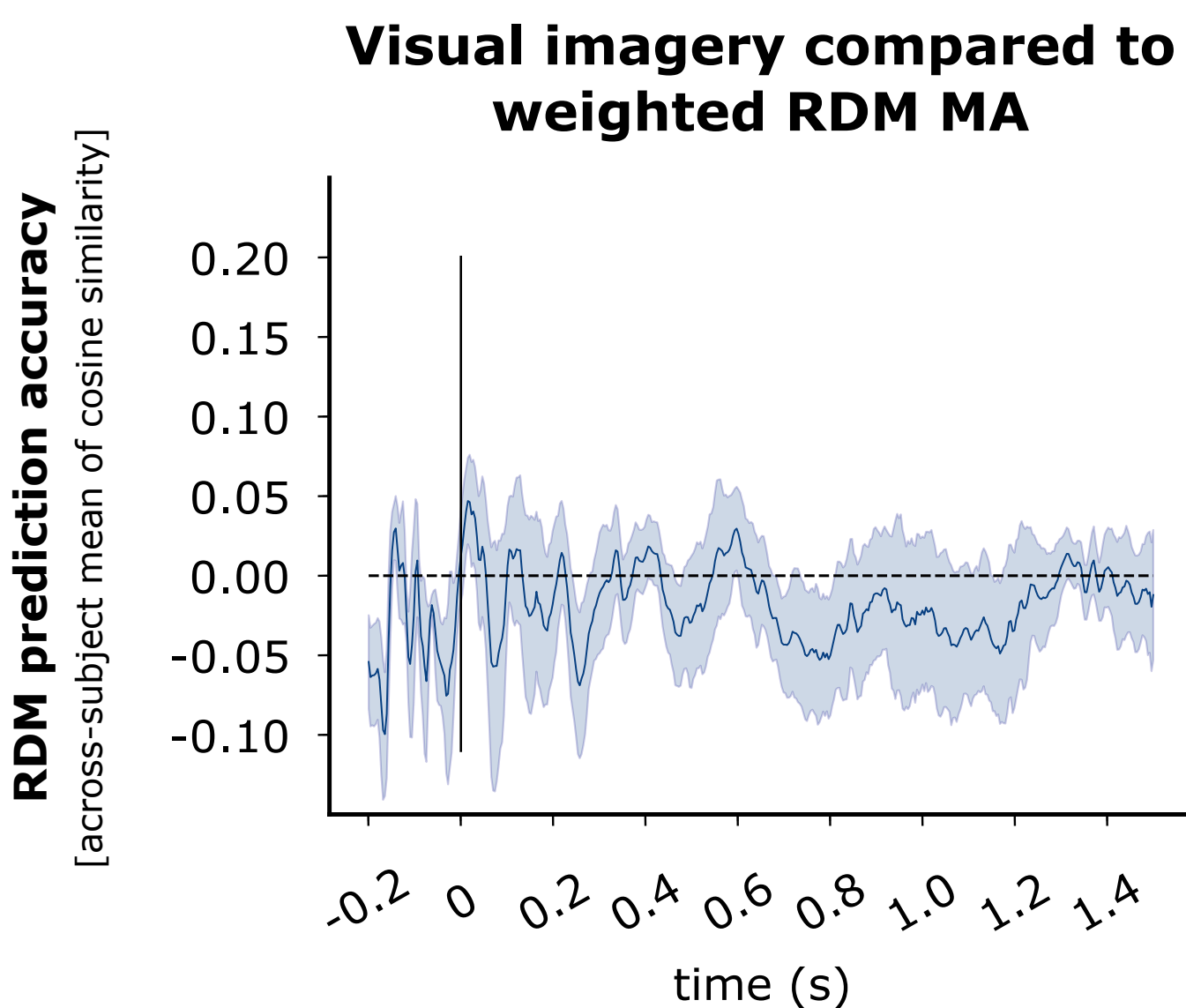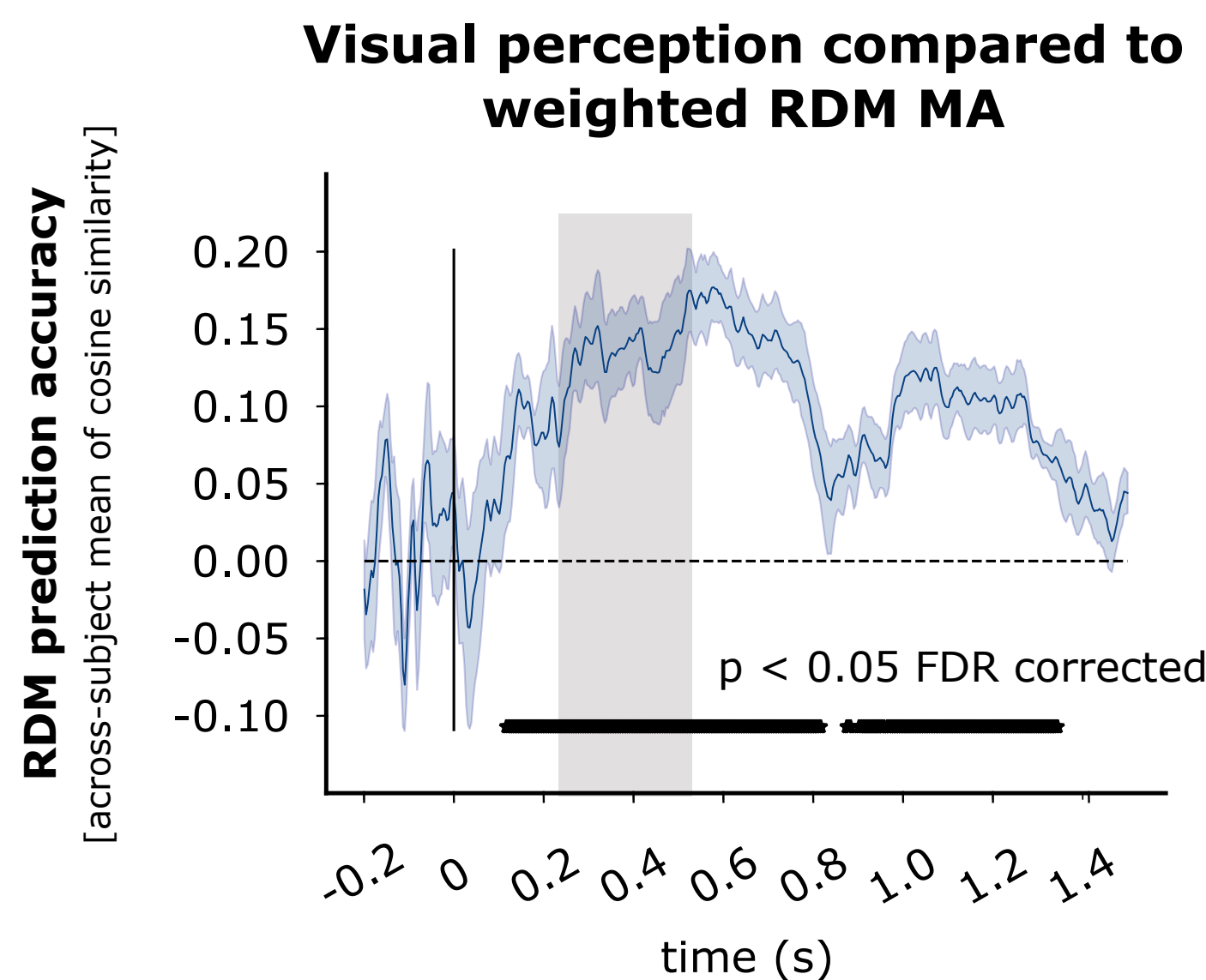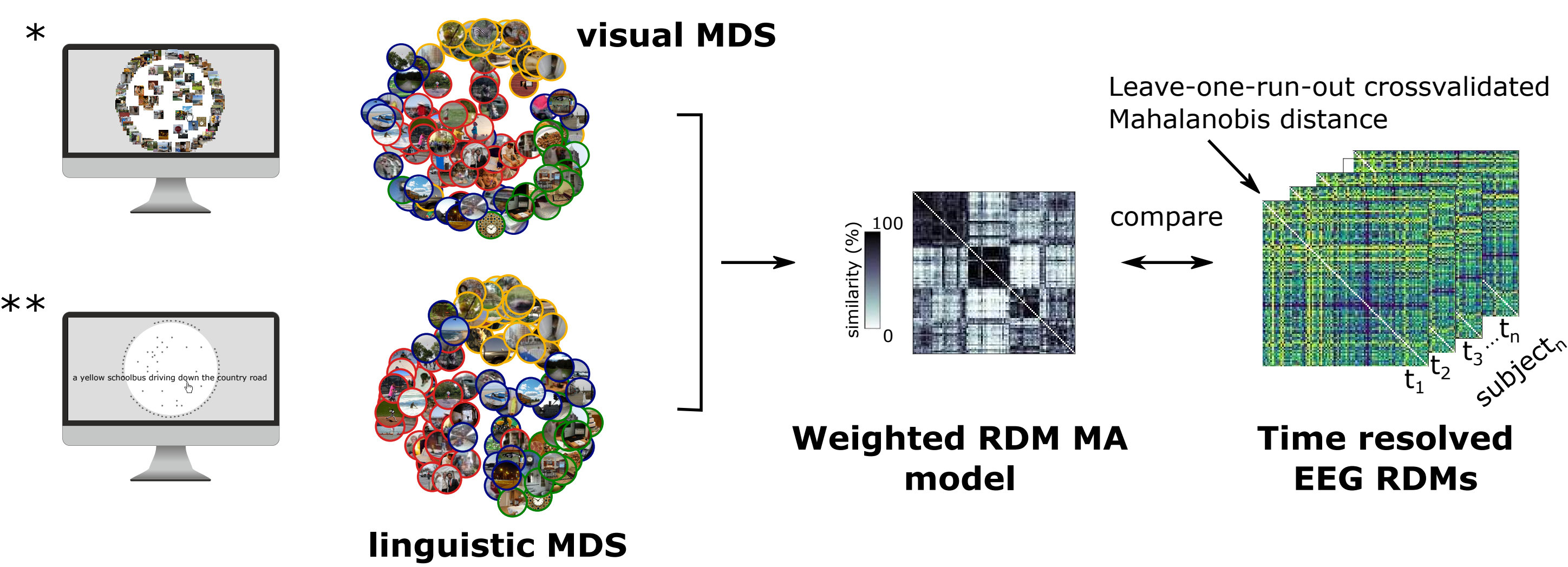
Linguistic modality

N = 6
128-channel EEG

Results

EEG representational dynamics predicted by semantic similarity structure

Visually evoked activity is predictable by semantic similarity structure. How does semantic similarity structure relate to language?

visual MDS

linguistic MDS

Weighted RDM MA model

Leave-one-run-out crossvalidated Mahalanobis distance

compare

Time resolved EEG RDMs

Visual perception compared to weighted RDM MA

p < 0.05 FDR corrected

Visual imagery compared to weighted RDM MA

Sentence reading compared to weighted RDM MA

p < 0.05

Linguistic imagery compared to weighted RDM MA
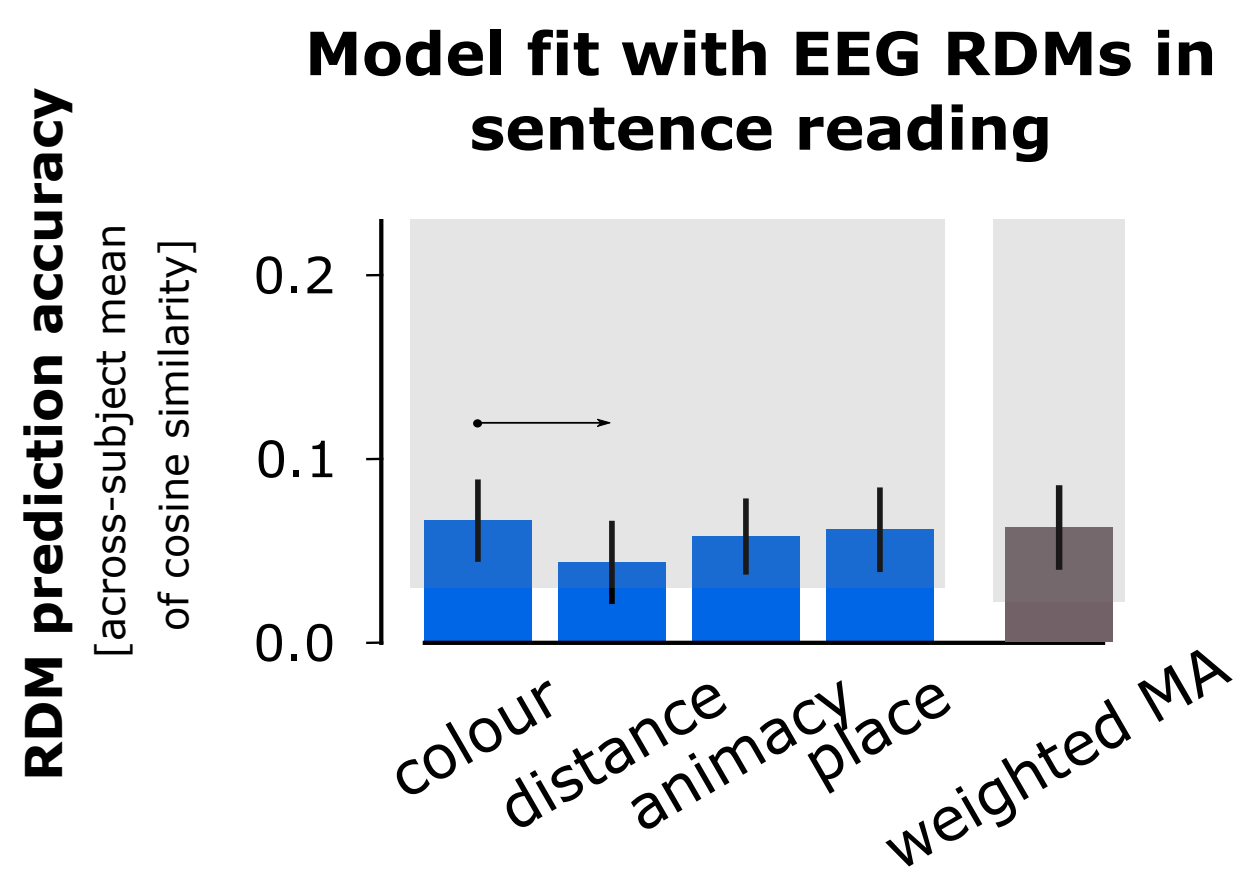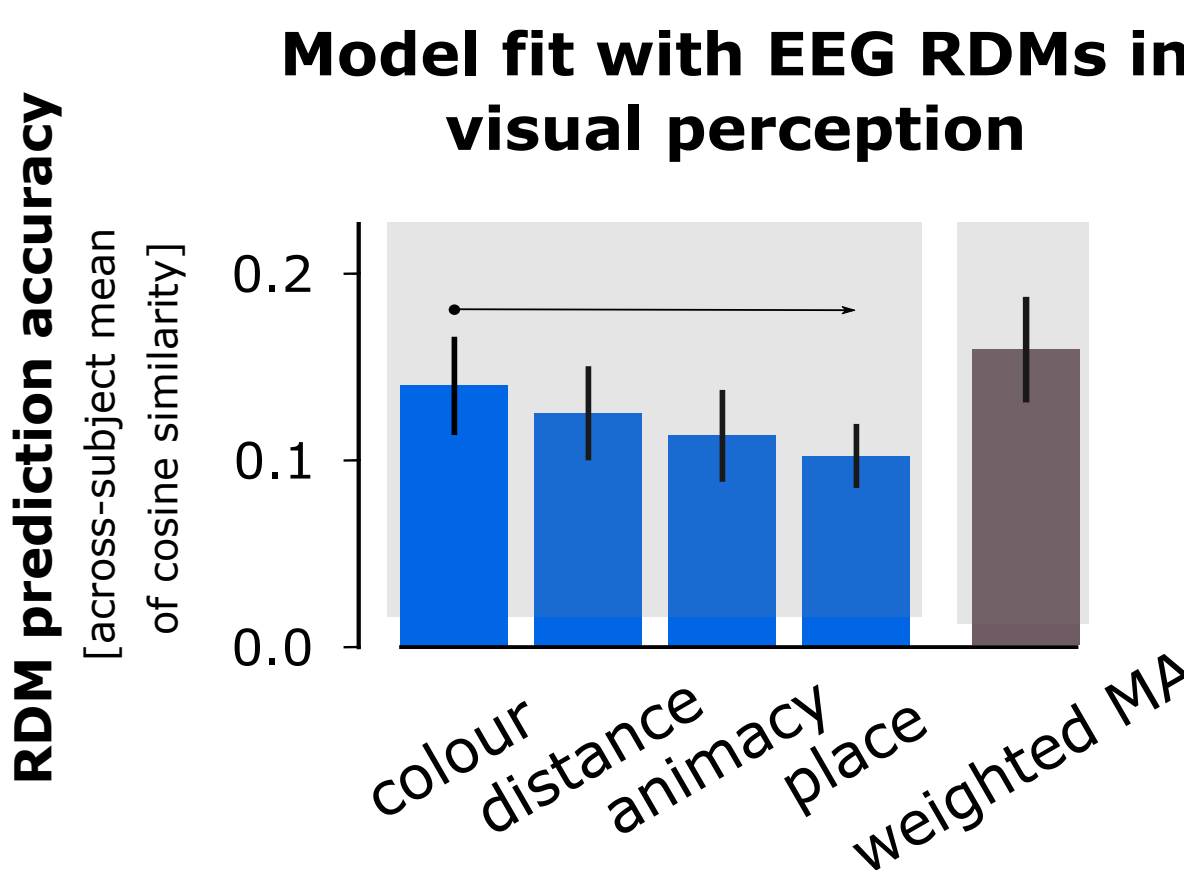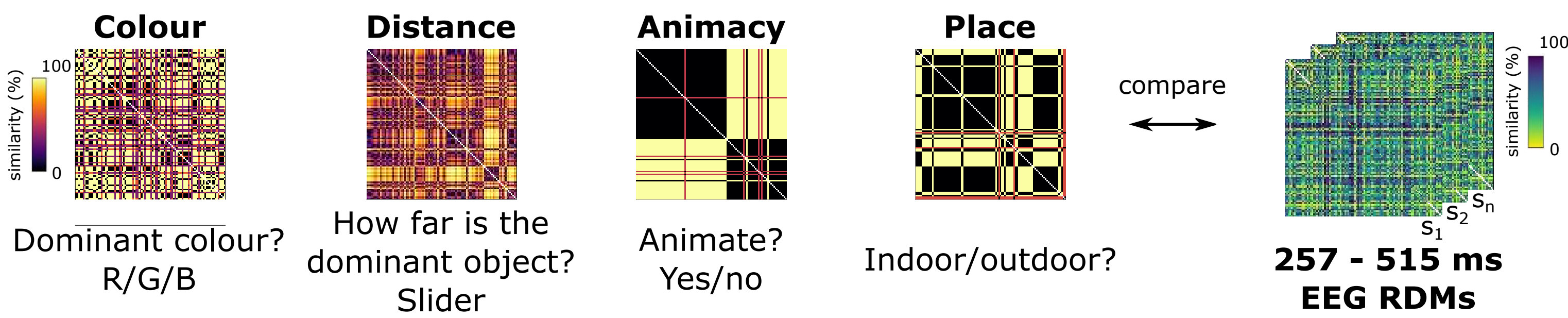
p < 0.05

Conclusion

We have promising results showing similar, behaviourally relevant representations elicited in the brain when perceiving natural scenes and reading scene captions depicting the same natural scenes.

Some of these seem to also be elicited, albeit to a lesser extent, when imagining the natural scenes (either from reading the scene caption, or seeing the scene to be imagined)

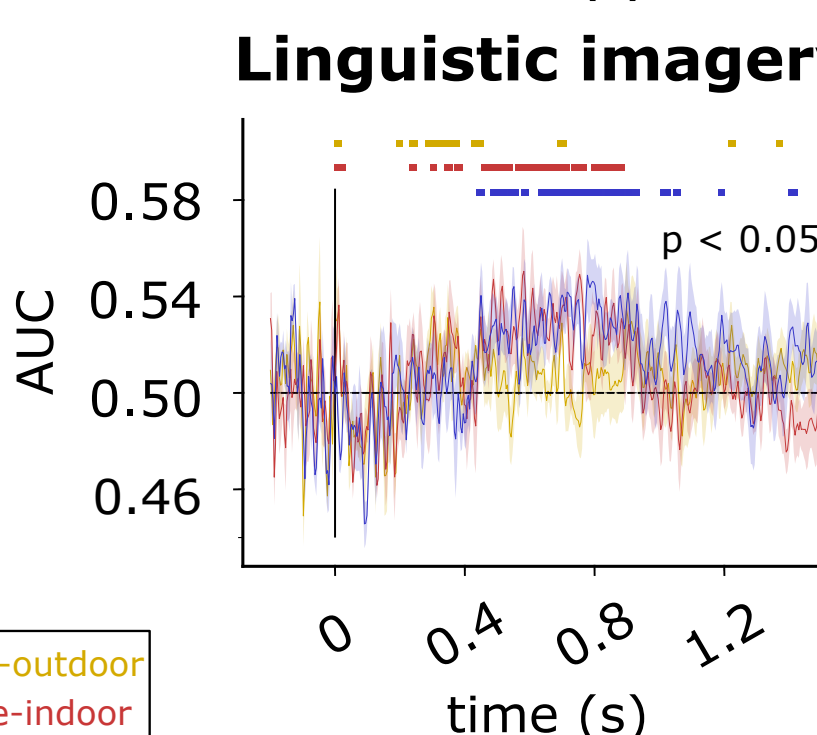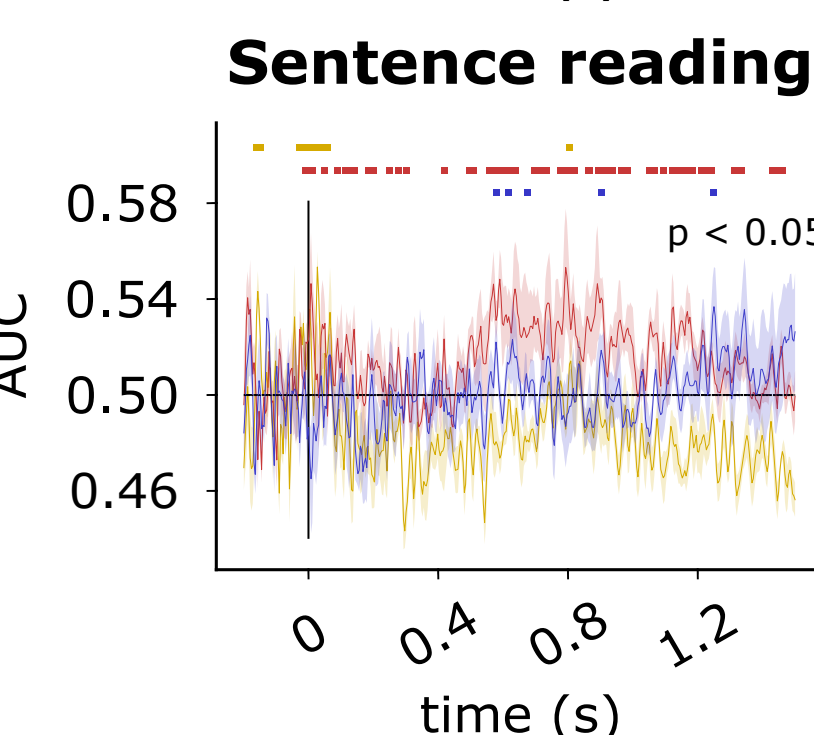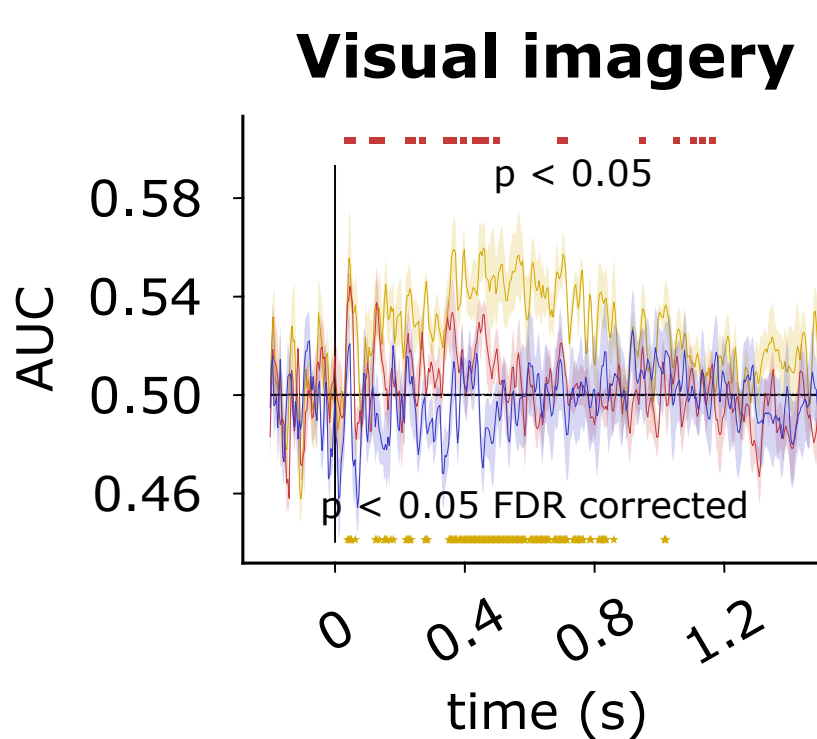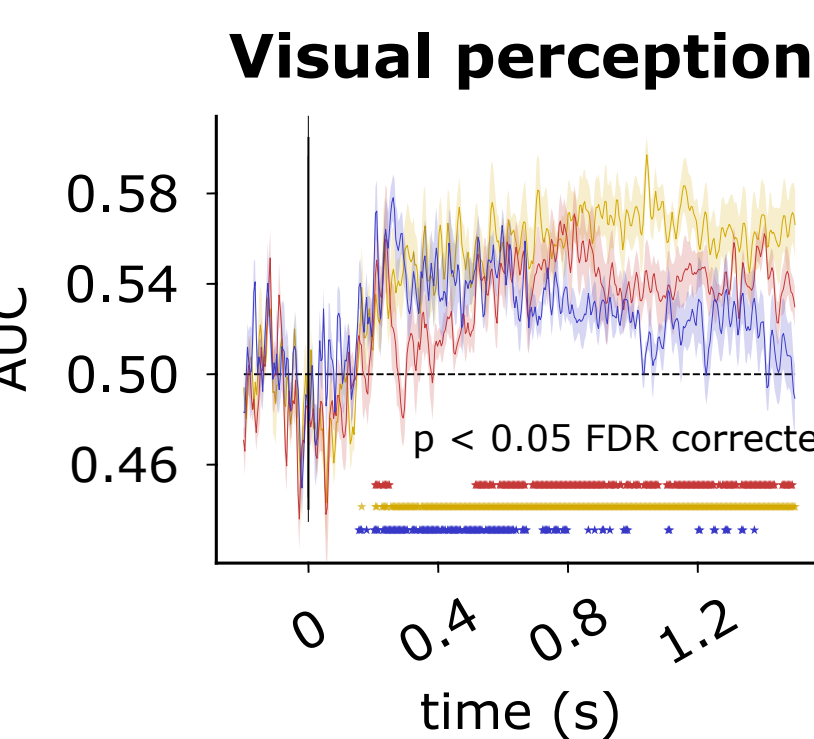### Mid- and high-level features: representations 257 ms - 515 ms post-stimulus

Representational dynamics in both visual perception and sentence reading correlated with semantic similarity structure. What is the relative weighting of mid- and high-level features?

2 independent subjects provided annotations of the NSD images:

Colour
Dominant colour? R/G/B

Distance
How far is the dominant object? Slider

Animacy
Animate? Yes/no

Place
Indoor/outdoor?

compare

257 - 515 ms EEG RDMs

Model fit with EEG RDMs in visual perception

Model fit with EEG RDMs in sentence reading

Categorical representations: within-modality decoding across time

Multidimensional scaling (MDS) of the multimodal similarity judgements revealed four key high-level categories. We explored whether these categories are also distinguishable from multimodal EEG using linear discriminant analysis (LDA).

Animate   **Outdoor**   **Indoor**

Visual perception
p < 0.05 FDR corrected

Visual imagery
p < 0.05
p < 0.05 FDR corrected

Sentence reading
p < 0.05

Linguistic imagery
p < 0.05

animate-outdoor
animate-indoor
outdoor-indoor

References

[1] Doerig et al (2022). Semantic scene descriptions as an objective of human vision. *arXiv*.

[2] Tang et al (2021). Cortical representations of concrete and abstract concepts in language combine visual and linguistic representations. *bioRxiv*.

[3] Tang et al (2023). Brain encoding models based on multimodal transformers can transfer across language and vision. *arXiv*.

[4] Wang et al (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5, 1415-1426.

[5] Simkova et al (2023). The semantic distance between a linguistic prime and a natural scene target predicts reaction times in a visual search experiment. *Journal of Vision August 2023*, 23(5055).

[6] Popham et al (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11), 1628-1636.

[7] Allen et al (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25, 116-126.

[QR placeholder]