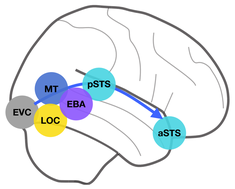
Modeling Dynamic Social Vision Highlights Gaps Between Deep Learning and Humans

Kathy Garcia[1], Emalie McMahon[1], Colin Conwell[1], Michael F. Bonner[1], Leyla Isik[1,2]

[1]Department of Cognitive Science [2]Department of Biomedical Engineering | Johns Hopkins University | Baltimore, MD

Introduction



Human visual system processes dynamic social scenes in distinct regions along the lateral surface of the brain, forming the recently proposed "lateral stream" **[1] [2].**

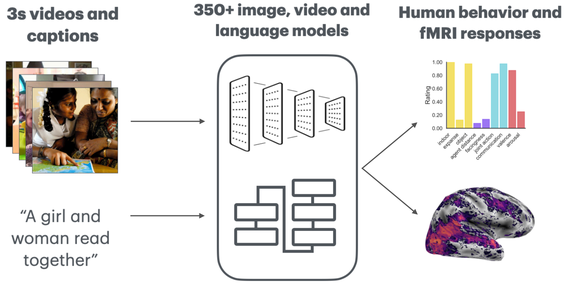
DNNs excel with static images **[3] [4]**, but their match to the brain in terms of perceiving and interpreting dynamic social scenes is still unknown **[5] [6] [7].**

Evaluating the match between human and DNN responses to dynamic social scenes can reveal gaps and guide improvements in AI models' understanding of complex, real-world stimuli.

Objectives

- Benchmark DNNs on dynamic, social visual responses in the lateral visual stream
- Identify computational factors improving lateral stream predictivity that improve match to human behavior and brain responses
- Explore whether there is a hierarchical correspondence between DNNs and the lateral stream, similar to the observed hierarchy in the ventral stream.

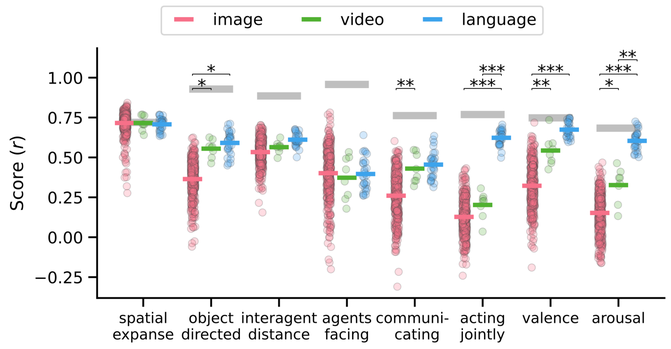
Methods Overview



Results

Behavioral Ratings

Language models best predict the human behavioral ratings

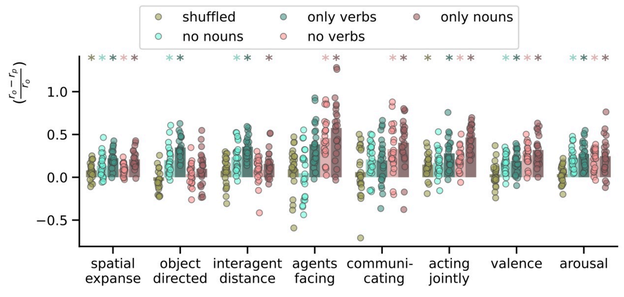


Alignment does not rely on linguistic composition, instead relies on both noun AND verb content

A

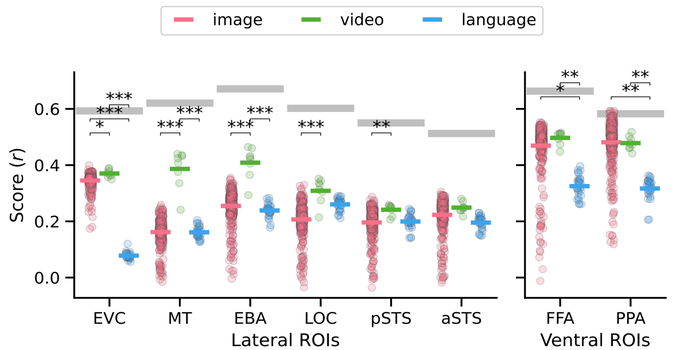
Category	Linguistic perturbation	Example sentence
Original	—	two men are outside and talking
Shuffled	compositional structure	two and outside talking men are
No nouns	nouns	two [MASK] are outside and talking
Only verbs	nouns + context	[MASK] [MASK] are [MASK] [MASK] talking
No verbs	verbs	two men [MASK] outside and [MASK]
Only nouns	verbs + context	[MASK] men [MASK] [MASK] [MASK] [MASK]

B

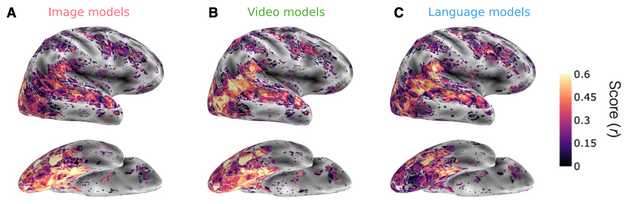


Neural Responses

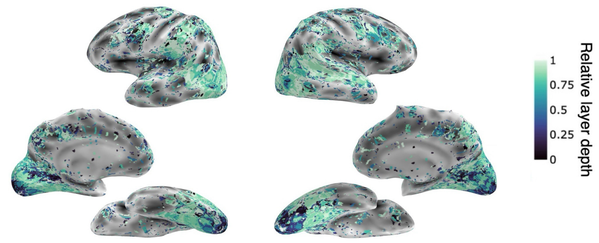
Video models excel in predicting mid-level lateral stream responses (e.g. MT, EBA, LOC)



Whole brain voxel-wise encoding confirms that video models are the most predictive of lateral regions



Whole brain analysis shows evidence of hierarchical correspondence in the ventral BUT NOT lateral stream



Conclusions

Language models align with human social ratings, relying on nouns and verbs, **but** show poor performance in matching neural responses.

Video models show improvement in mid-level lateral regions and brain responses in more anterior regions **but** fall short in predicting social behavior.

DNNs provide the proposed hierarchical correspondence to the ventral stream, but do not reproduce the known hierarchy **[7]** in the lateral stream.

No model class could match **both** human behavior and brain responses!

The success of language & video models in matching behavior and brain responses, respectively, highlights the need for models with dynamic, relational processing.

References

(1) Pitcher. & Ungerleider, Trends in CogSci, 2021.
(2) Wurm & Caramazza, Trends in CogSci, 2022.
(3) Schrimpf, M., et al., Neuroscience, 2018.
(4) Conwell, et al., Journal of Vision, 2023.
(5) Bolotta & Dumas. Frontiers in CS, 2022.
(6) McMahon & Isik, Trends in CogSci, 2023.
(7) McMahon, Bonner, & Isik, CurrentBio, 2023.

GitHub Repo | Twitter | Preprint | Website