Modality-specific Representations in Multimodal Large Language Models Align with Distinct Neural Systems in the Brain

Jixing Li[1], Qihong Lu[1], Jian-Qiao Zhu[2]
[1]City University of Hong Kong, Hong Kong SAR, China
[2]The University of Hong Kong, Hong Kong SAR, China

The platonic representation hypothesis (Huh et al., 2024, ICML) proposes that as AI models increase in scale and capacity, their internal representations converge toward a shared embedding space of the world. Here, we test whether such convergence extends to alignment between multimodal large language models (LLMs) and the human brain. We collected fMRI and MEG data from two independent cohorts (30 participants each) while participants watched a reality TV show for approximately 30 minutes. The same video segments were input into a multimodal LLM (Qwen-2.5-Omni-7B), from which we extracted text, audio, and visual embeddings using the corresponding modality-specific encoders. We aligned these representations with voxelwise fMRI responses and time-resolved MEG activity across the cortex. Despite being embedded within a unified model, modality-specific representations selectively mapped onto language, auditory, and visual cortices at distinct temporal scales. These results suggest that multimodal LLMs preserve modality-specific representational structure that aligns with dissociable neural systems in the human brain.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*