A Large-Scale Study of Social Scene Judgments Reveals Alignment with Deep Neural Networks and Social-Affective Features

Kathy Garcia (kgarci18@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
Baltimore, MD 21218, U.S.A.

Leyla Isik (lisik@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
Department of Biomedical Engineering, Johns Hopkins University
Baltimore, MD 21218, U.S.A.

Abstract

Similarity judgments offer a window into the mental representations people use to make sense of objects and events—yet most prior work has focused on static images, leaving dynamic scene understanding underexplored. We introduce a novel large-scale dataset of 56,000+ odd-one-out similarity judgments for 250 3-second videos of everyday naturalistic social actions. Using representational similarity analysis (RSA), we compare human similarity judgments to behavioral ratings of social-visual features, fMRI responses, and embeddings from pretrained deep neural networks (DNNs). Language model embeddings from human video annotations explained the most unique variance in behavior, followed by social-affective features and visual DNNs. Neural activity in high-level social perception regions (EBA, LOC, STS, FFA) mirrored the behavioral similarity structure, whereas early visual and scene-selective areas did not. Variance partitioning showed that behavioral and model-derived features captured both overlapping and complementary structure, and their combination reached the level of split-half agreement in the data. This highlights how current AI models, especially language models encoding semantic information, perform well in approximating human judgments. Together, these findings and dataset reveal the nature of social event representations and offer a framework for evaluating model–brain–behavior alignment in dynamic social perception.

Keywords: social perception; RSA; dynamic visual scenes; similarity judgments; DNNs; human-model alignment

Introduction

Humans effortlessly understand different types of social interactions and use this information to make social judgments from early in development (Thomas et al., 2022). Measuring perceived similarity between interactions provides a window into the structure of the underlying mental representations. While prior work has explored these representations using static objects (Hebart et al., 2020) or actions (Dima et al., 2022), it remains unclear which features drive similarity judgments in naturalistic, dynamic social interactions—and how these judgments align with neural and model representations

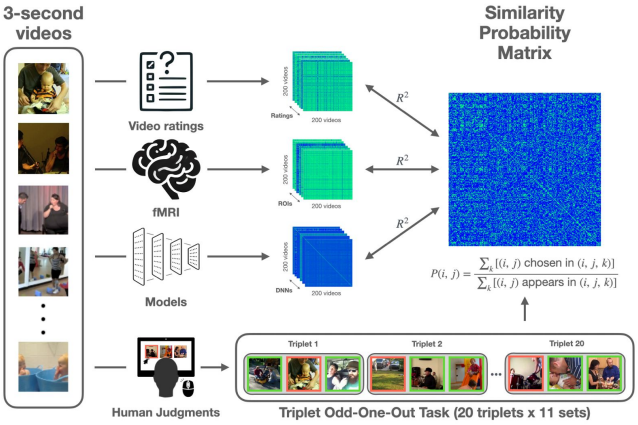

Figure 1: Overview of RSA methodology used to calculate the representational similarity between various features and human similarity judgments.

(Peterson et al., 2016; Vo et al., 2019). To extend this work to rich, dynamic social interactions, we investigate how people assess similarity between naturalistic videos depicting two people engaged in everyday actions. Using a triplet odd-one-out design, we collected targeted comparisons that emphasize salient social-visual distinctions. We then applied RSA to relate these human judgments to: (1) behavioral ratings of social and perceptual attributes, (2) fMRI responses from visual and social brain regions, and (3) embeddings from image-, video-, and language-based DNNs. This framework allows us to evaluate the extent to which each feature space explains human judgments.

Methods

Stimuli and Similarity Judgments We collected 56,421 triplet judgments on an existing dataset of 250 short (3s) videos of everyday social interactions (McMahon et al., 2023) curated from the Moments in Time dataset (Monfort et al., 2020). For this analysis we focused on the 25,623 trials from the pre-defined 200-video training set in (Garcia et al., 2025; McMahon et al., 2023). Participants (N=215) completed odd-one-out similarity judgment tasks (e.g., Hebart et al., 2020) on Meadows Research (`https://meadows-research`

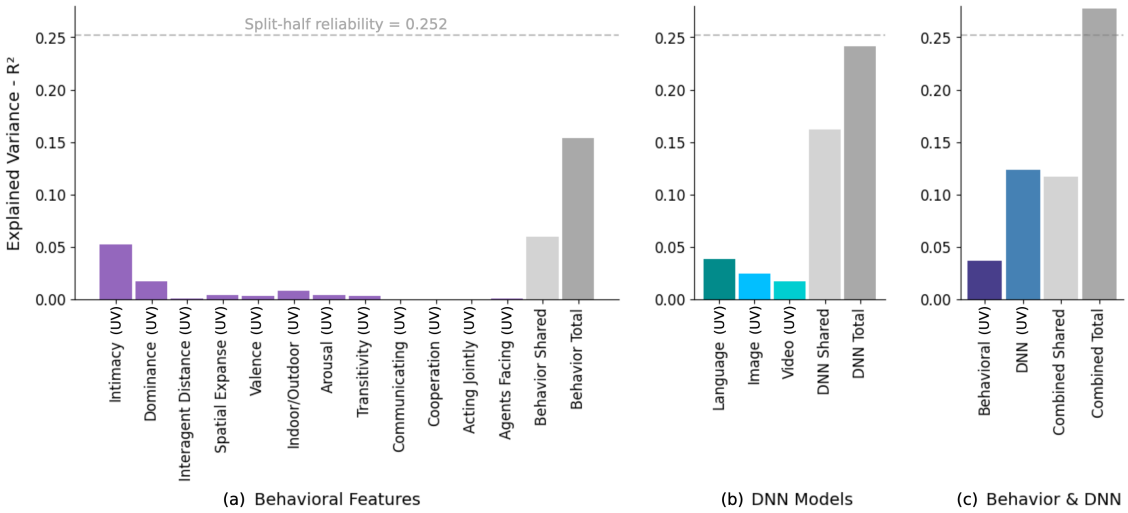

Figure 2: Unique and shared explained variance (R^2) for behavioral features and DNN modalities obtained via multiple linear regression. The variance is decomposed into (a) unique contributions from individual features (behavior), (b) best modality RSM (DNNs), and (c) shared and total variance in predicting human similarity judgments. Behavior & DNN is decomposed into unique (UV), shared, and total (all possible features combined) variance. The dashed line is the split-half reliability ($R^2 = 0.252$)

.com), selecting the video least similar to the other two in each trial. Previously, an independent sample (*N*=150; McMahon et al., 2023) rated the video set across 13 social-affective and perceptual attributes, which were used to compute representational similarity matrices (RSMs) based on pairwise euclidean distances between videos in each feature space.

Unique and Shared Explained Variance Partitioning We computed a 200×200 similarity probability matrix from human odd-one-out judgments (Figure 1) and compared it to representational similarity matrices (RSMs) derived from (1) behavioral ratings, (2) embeddings from pretrained DNNs (language: e.g., MPNet (Song et al., 2020) embeddings evaluated on sentence captions of the videos, image: e.g., CLIP (Radford et al., 2021); video: e.g., X3D (Feichtenhofer, 2020)), and (3) fMRI responses from regions implicated in social and visual processing (EBA, STS, LOC; Pitcher & Ungerleider, 2021; Wurm & Caramazza, 2022), face-selective (FFA; Kanwisher et al., 1997), early visual (EVC), and scene-selective (PPA; Epstein & Kanwisher, 1998) regions.

Next, using cross-validated multiple linear regression RSA, we quantified the unique and shared variance in human similarity judgments explained by each feature space individually, as well as by the combined behavioral-DNN feature.

Results

Similarity structure corresponds to high-level social ratings and DNN features We first conducted RSA between each feature space and behavior similarity. Figures are not shown due to space constraints. For behavior features, intimacy, dominance, agent distance, and expanse each significantly predicted similarity judgments (ps $< .05$), though no single feature approached the split half reliability of similarity judgments. Intimacy emerged as the strongest single behavior predictor ($R^2 = 0.1$), though follow-up analyses suggest this largely reflects the presence of children in videos.

RSA between behavior similarity and fMRI revealed strong alignment with social perception regions along the lateral and ventral streams (EBA, LOC, FFA, STS). In contrast, the early visual cortex (EVC) and scene-selective regions (PPA) showed minimal correspondence (not shown for space).

Among pretrained DNNs, average embeddings from language models best predicted human judgments, outperforming vision models. Video models uniquely captured far less variance, showing limitations in current model architectures.

Multiple regression RSA and variance partitioning Next, we conducted multiple regression RSA and variance partitioning using all behavioral features and the top DNN in each model modality (image: CLIP_R50; video: X3D_M; language: Multilingual MPNet). Results (Figure 2) showed that behavioral and DNN features each uniquely explained human similarity judgments, with DNNs accounting for the majority.

Discussion

We introduced a novel large-scale dataset of over 56,000 human similarity judgments to systematically characterize how people represent dynamic social interactions. Our results reveal that humans rely on combined social-semantic dimensions, as well as visual cues. Neural responses in specialized social-perception regions—particularly STS and EBA—robustly mirrored this representational structure, while early visual and scene-selective areas showed minimal correspondence. Our modeling analyses showed that modern DNNs can closely align with the structure of human similarity judgments on dynamic social scenes, especially language

models evaluated on sentence captions ($R^2 = 0.170$). Pure vision models alone ($R^2_{image} = 0.134,\ R^2_{video} = 0.127$) had lower performance (consistent with Garcia et al., 2025). Further, DNNs combined with human behavioral ratings, showed higher total explained variance, and reaching the split half reliability. We use behavioral ratings as a perceptually grounded reference to evaluate how well model-derived representations align with human social perception. Together, our dataset and results offer a robust framework for model–brain–behavior alignment and a road map for developing integrative models that capture the complexity of human social perception.

Acknowledgments

This work was funded in part by NSF GRFP DGE-2139757 awarded to K.G., and NIMH R01MH132826 awarded to L.I.

References

Dima, D. C., Tomita, T. M., Honey, C. J., & Isik, L. (2022). Social-affective features drive human representations of observed actions. *eLife*, *11*. doi: 10.7554/eLife.75027

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. Retrieved 2024-02-07, from https://www.nature.com/articles/33402 doi: 10.1038/33402

Feichtenhofer, C. (2020). *X3d: Expanding architectures for efficient video recognition.* arXiv. Retrieved from https://arxiv.org/abs/2004.04730 doi: 10.48550/arxiv.2004.04730

Garcia, K., McMahon, E., Conwell, C., Bonner, M. F., & Isik, L. (2025). Modeling dynamic social vision reveals gaps between deep learning and the human brain. In *Proceedings of the thirteenth international conference on learning representations.* Retrieved from https://openreview.net/pdf?id=wAXsx2MYgV

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*, 1173–1185. Retrieved from https://doi.org/10.1038/s41562-020-00951-3 doi: 10.1038/s41562-020-00951-3

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, *17*(11), 4302–4311. Retrieved 2024-02-07, from https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.17-11-04302.1997 doi: 10.1523/JNEUROSCI.17-11-04302.1997

McMahon, E., Bonner, M. F., & Isik, L. (2023). Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, *33*(23), 5035–5047.e8. Retrieved from http://dx.doi.org/10.1016/j.cub.2023.10.015 doi: 10.1016/j.cub.2023.10.015

Monfort, M., Vondrick, C., Oliva, A., Andonian, A., Zhou, B., Ramakrishnan, K., ... Gutfreund, D. (2020). Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 502–508. Retrieved from http://dx.doi.org/10.1109/TPAMI.2019.2901464 doi: 10.1109/tpami.2019.2901464

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2367–2372).

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, *25*(2), 100–110. Retrieved 2024-02-07, from https://linkinghub.elsevier.com/retrieve/pii/S1364661320302783 doi: 10.1016/j.tics.2020.11.006

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021, February). *Learning Transferable Visual Models From Natural Language Supervision.* arXiv. Retrieved 2024-04-25, from http://arxiv.org/abs/2103.00020 (arXiv:2103.00020 [cs]) doi: 10.48550/arXiv.2103.00020

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). *Mpnet: Masked and permuted pre-training for language understanding.* arXiv. Retrieved from https://arxiv.org/abs/2004.09297 doi: 10.48550/arxiv.2004.09297

Thomas, A. J., Saxe, R., & Spelke, E. S. (2022, August). Infants infer potential social partners by observing the interactions of their parent with unknown others. *Proceedings of the National Academy of Sciences*, *119*(32), e2121390119. Retrieved 2023-02-03, from https://www.pnas.org/doi/10.1073/pnas.2121390119 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2121390119

Vo, V.-A., Wang, L., & Voss, M. W. (2019). Similarity judgment for natural images reflects perceptual similarity more than conceptual similarity. *Journal of Experimental Psychology: General*, *148*(6), 994–1005. doi: 10.1037/xge0000612

Wurm, M. F., & Caramazza, A. (2022). Two 'what' pathways for action and object recognition. *Trends in Cognitive Sciences*, *26*(2), 103–116. Retrieved 2024-02-07, from https://linkinghub.elsevier.com/retrieve/pii/S1364661321002588 doi: 10.1016/j.tics.2021.10.003