UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Relational Information Predicts Human Behavior and Neural Responses to Complex Social Scenes

Permalink
https://escholarship.org/uc/item/4680v4ws

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Qin, Wenshuo
Malik, Manasi
Isik, Leyla

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Relational Information Predicts Human Behavior and Neural Responses to Complex Social Scenes

Wenshuo Qin (wqin6@jh.edu)
Department of Cognitive Science, 3400 N Charles St,
Baltimore, MD 21218 USA

Manasi Malik (mmlik16@jhu.edu)
Department of Cognitive Science, 3400 N Charles St,
Baltimore, MD 21218 USA

Leyla Isik (lisik@jhu.edu)
Department of Cognitive Science, 3400 N Charles St,
Baltimore, MD 21218 USA

Abstract

Understanding social scenes depends on tracking relational visual information, which is prioritized behaviorally and represented in the superior temporal sulcus (STS), a region involved in processing social scenes. Despite its importance, relational information has been underutilized in computational models of social vision. In this study, we evaluate two neural network models—SocialGNN and RNN Edge—that explicitly incorporate relational cues, and compare their performance to state-of-the-art (SOTA) AI vision models. SocialGNN utilizes a graph neural network to organize input information about each video frame into a graph structure with nodes representing faces and key objects, and edges encoding relational information such as gaze direction and physical contact. RNN Edge is an even simpler model that processes only relational information without node features or graph-based structures. These models were tested on behavioral and neural data from 3-second natural videos of two people engaged in everyday activities, as well as on the PHASE dataset, a collection of 2D animations depicting agent-object interactions inspired by Heider and Simmel. Across both datasets, SocialGNN and RNN Edge achieved strong performance in predicting human behavioral ratings of social interactions and were comparable to SOTA AI models in behavioral encoding tasks, despite being trained on significantly less data and with simpler architectures. Notably, the success of RNN Edge suggests that additional visual features and the graph-based framework of SocialGNN do not significantly enhance performance, underscoring the primacy of gaze and physical contact as essential relational cues. These findings emphasize the importance of integrating relational information into computational models to develop better models of social perception and human-aligned AI.

Keywords: Artificial Intelligence; Cognitive Neuroscience; Social Cognition; Neural Networks; fMRI

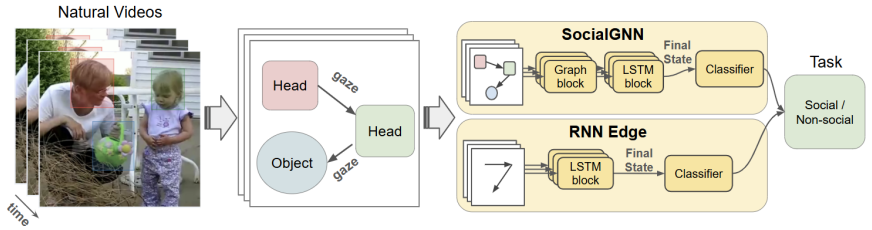
Introduction

Recognizing social interactions between others is a core human ability that humans use to gather information about the social world. The superior temporal sulcus (STS) plays a central role in processing these interactions, particularly when individuals observe face-to-face exchanges, highlighting its involvement in interpreting social dynamics (Allison, Puce, & McCarthy, 2000; Deen, Koldewyn, Kanwisher, & Saxe, 2015; Deen, Saxe, & Kanwisher, 2020; Isik, Koldewyn, Beeler, & Kanwisher, 2017; Kreifelts, Ethofer, Shiozawa, Grodd, & Wildgruber, 2009; McMahon, Bonner, & Isik, 2023). Recent studies emphasize that this process often relies on bottom-up visual cues, which are integral to social understanding and relationship inference (Fox, 2005; McMahon & Isik, 2023; Salatiello, Hovaidi-Ardestani, & Giese, 2021). Relational information, such as gaze direction, proximity, and physical contact, are bottom-up visual cues that are considered to play a crucial role in social interaction perception (Hafri & Firestone, 2021; McMahon et al., 2023; Papeo, 2020). However, current research often overlooks the integration of these cues into computational models. Perhaps as a result, even state-of-the-art (SOTA) AI models do a poor job of recognizing and understanding social scenes (Bolotta & Dumas, 2022; Garcia, McMahon, Conwell, Bonner, & Isik, 2024; Shu et al., 2021). Addressing this gap could significantly enhance our understanding of social cognition by providing a more comprehensive framework for analyzing social scenes and lead to better human-aligned AI.

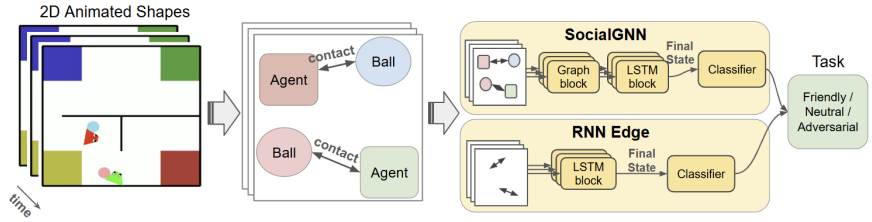
Here, we analyzed two models: SocialGNN (Malik & Isik, 2023) and a new simpler counterpart, the RNN Edge model. SocialGNN combines the strengths of recurrent neural networks (RNNs) and graph neural networks (GNNs) by representing each video frame as a graph, where nodes capture visual information about human faces and objects, and edges encode relational cues such as gaze direction or physical contact. These graphs are processed over a time series, enabling SocialGNN to capture both spatial and temporal aspects of social interactions. RNN Edge is a simpler model that takes only the edge information (gaze direction or physical contact) between entities as input to an RNN.

We conducted three main experiments. In our first experiment, the RNN Edge model outperformed SocialGNN and a matched non-relational model in classifying social interactions within the train-test splits of the natural video datasets on which the models were trained. This suggests that relational information alone can accurately predict human behavior. Next, we extended our analysis to a natural video dataset and found that both SocialGNN and RNN Edge effectively encode human behavioral and neural responses, even when these videos are out of distribution relative to the model's training set. Remarkably, both models performed at the level of SOTA AI vision models in behavioral encoding. Addi-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).



(a) Natural video processing pipeline



(b) PHASE 2d animation videos processing pipeline

Figure 1: **GNN and RNN model overview.** The processing pipelines of SocialGNN and RNN Edge models for two types of datasets: (a) natural videos, and (b) 2D animated PHASE videos. (a) For SocialGNN, each frame from the natural videos is represented as a graph, where nodes correspond to individual faces, and edges represent directed gaze interactions. The graph-based representation is passed to a graph block containing a node and an edge processing unit, both with linear transformations, and outputs a vector. An LSTM block then captures the temporal dependencies across video frames, and a subsequent classifier predicts whether the video depicts a social interaction based on the LSTM's final hidden state. For RNN Edge, only the directed gaze edges of the graph are extracted and processed, without the node-level representation. These edges are similarly passed through an LSTM block (without a graph block), final state, and classifier. (b) Similarly, for SocialGNN, each video frame in the PHASE dataset is converted into a graph where nodes represent velocity, position, size, and entity type; the bidirectional edges capture physical contact between entities. The graph is subsequently processed by a graph block, followed by an LSTM block to model temporal dynamics, and a final classifier that categorizes the interaction between agents as friendly, neutral, or adversarial. For the RNN Edge, only the bidirectional physical contact edges are extracted for processing through an LSTM block, final state, and classifier, excluding node-specific information.

tionally, they encoded more meaningful information in lateral brain regions, particularly the superior temporal sulcus (STS), compared to other brain regions. Finally, we demonstrated the importance of relational information generalized to animated social scenes. SocialGNN and RNN Edge excelled at predicting human judgments on abstract, shape-based interactions, further underscoring the critical role of relational cues in human social scene understanding.

Methods

SocialGNN

SocialGNN combines an RNN and GNN to process graphical representations of visual information in video frames (Malik & Isik, 2023). Following prior work, we employed two versions of SocialGNN: one trained on natural videos from the VACATION (Video gAze CommunicATION) dataset (Fan, Wang, Huang, Tang, & Zhu, 2019) and another trained on animated shape videos from the PHASE (PHysically-grounded Abstract Social Events) dataset (Netanyahu, Shu, Katz, Barbu, & Tenenbaum, 2021) (Figure 1). Graphical information was processed through a linear layer before being input into the LSTM block. In the prior study by Malik and Isik (2023), SocialGNN demonstrated superior performance in social versus non-social classification tasks compared to baseline RNNs with the same node information but no graph structure.

In this study, to improve SocialGNN's performance and mitigate overfitting, we adjusted the L2 regularization term from 0.05 to 0.2 for both versions. For the natural video task, while the original study reduced VGG19 node features to 20 principal components (PCs), explaining 50% of the variance, we extended this to 90 PCs to capture 75% of the variance, providing a richer set of visual features. The model was retrained on the same binary social interaction discrimination task using the original 20 bootstrapped train/test splits, each of which contained around 740 training clips and 215 testing clips. For the PHASE dataset, we retrained the model on the same friendly/neutral/adversarial classification task using the original split of 400 training and 100 testing videos.

RNN Edge

To further assess the role of gaze direction in social interaction understanding, we developed an even simpler model called the RNN Edge model. This model was trained using only gaze direction inputs, represented as a simple 20-dimensional vector (Figure 1a). Each entry in this vector was binary (1 or 0), indicating whether an agent was gazing at a

corresponding object. We also made a PHASE version of the RNN Edge, where the inputs are the 12-dimensional vectors, representing bidirectional contact-based physical interactions (Figure 1b). This approach allowed us to isolate the contribution of relational information to the model's performance.

Other Models

We included the VisualRNN model described in Malik and Isik (2023) for comparison. VisualRNN is an RNN model trained exclusively on node information without any relational representations. This model was preserved to specifically evaluate the contributions of node features alone in our analysis. As with SocialGNN, we also retrained the model with an L2 regularization term of 0.2 instead of 0.05 to avoid overfitting. In the original study, the authors utilized both node and context information for the PHASE analysis. However, in this study, we restricted the input to node information only, renaming this model "RNN Node" to distinguish it from the "RNN Edge" described above.

In the PHASE experiment, we included a generative inverse planning model. The generative inverse planning model is a Bayesian framework that generates hypotheses about an agent's internal goals and recognizes interactions by comparing predictions based on those hypotheses to observed interactions to infer the social interaction type. SIMPLE, an implementation of this approach, has demonstrated SOTA performance on the PHASE dataset, making it a benchmark for evaluating goal-directed behavior modeling (Netanyahu et al., 2021).

Natural Video Encoding Dataset

We utilized a natural social interaction dataset comprising behavior and functional magnetic resonance imaging (fMRI) responses to 250 three-second videos of people engaged in everyday activities (McMahon et al., 2023). Online human raters provided judgments on a range of visual and social scene features in the videos, and separate subjects viewed the videos while their neural activity was recorded using fMRI (McMahon et al., 2023). Prior neuroAI benchmarking work has shown that even modern SOTA models struggle to match human behavior and neural responses to these videos (Garcia et al., 2024).

In this dataset, each video included human behavioral ratings across six dimensions: spatial expanse (small versus large scenes), inter-agent distance, the extent to which agents are facing, the presence of object-directed actions, joint physical interactions between agents, and communicative interactions. Additionally, the dataset contained fMRI neural responses from four participants, covering regions of interest (ROIs) in the early visual cortex (EVC), as well as areas along the lateral and ventral streams. The lateral stream includes motion-selective middle temporal area (MT), extrastriate body area (EBA), and posterior and anterior social-interaction selective regions in the superior temporal sulcus (pSTS and aSTS). Regions in the ventral stream include the face-selective fusiform face area (FFA) and place-selective parahippocampal place area (PPA).

To adapt the dataset for our study, we conducted additional labeling of the entities needed to construct scene graphs. This included annotating bounding boxes for individuals' heads and key objects, as well as gaze directions. Key objects were defined as those at which at least one person gazed during the video. If a person moved out of the frame, was occluded, or gazed outside the visible area, their gaze direction was labeled as "none." Two videos (one from the training set and one from the test set) were excluded due to consistently containing fewer than two heads per frame. The 200 training videos were segmented into 230 shorter clips, and the 50 test videos into 51 clips, with segmentation occurring at points where the number of visible heads changed. To prevent data leakage, clips from the same video were assigned exclusively to either the training or test set.

For data encoding, we extracted the final hidden state representations of each clip from the model. Representations corresponding to clips from the same original video were then averaged before proceeding with the encoding analysis. We then followed the procedures outlined by Garcia et al. (2024). Model representations, behavior, and neural data were z-scored by fitting the transformation on the training set and applying it to both the training and test sets. We performed leave-one-out ridge regression to align model representations with behavior or neural responses. The alpha penalty for ridge regression was selected from a search space of seven values sampled from a log space ranging from 10^{-2} to 10^{5}.

As a point of comparison, we include the results from the DeiT3-L model (Touvron, Cord, & Jégou, 2022), identified by Garcia et al. (2024) as the best-performing vision model at matching human behavior on this dataset. The DeiT3-L was initially trained on ImageNet-22k and fine-tuned on ImageNet-1k. With 304.4 million parameters, it required 61.6 GFLOPs per input for computation.

PHASE Dataset

Finally, we followed Malik and Isik (2023) to evaluate SocialGNN, RNN Edge, RNN Node, and SIMPLE on the PHASE dataset (Figure 1b), which consists of 500 procedurally generated 2D animations depicting diverse social interactions. The dataset includes social events categorized into three relationship types: friendly, neutral, and adversarial.

Results

Edge Features Provide the Best Performance on Natural Video Classification

We first tested our models on the VACATION natural video dataset used in the original SocialGNN paper. We find that, as in the original study, SocialGNN significantly outperforms RNN Node; however, the new RNN Edge model significantly outperformed the SocialGNN in predicting social interactions (two-tailed paired permutation test, $p < 0.001$, $n = 10,000$

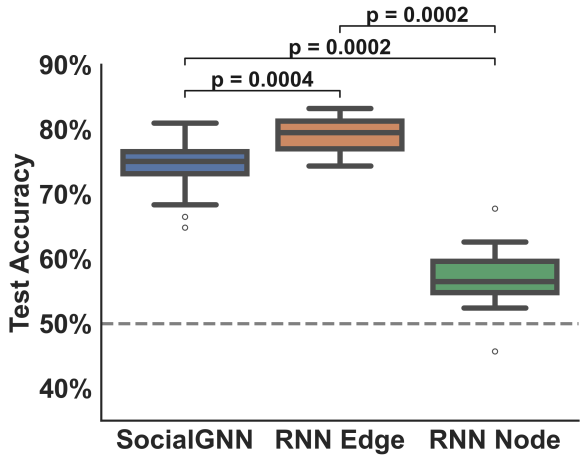


Figure 2: **Model accuracy on the VACATION dataset.** Test accuracies of different models on the VACATION dataset. Boxplots represent the test accuracies of predicting social versus nonsocial videos for each model type (SocialGNN, RNN Edge, and RNN Node). Each model type consists of 20 variations, trained on different bootstrapped data samples. Hollow dots are outliers. Pairwise comparisons between model accuracies are shown above the boxplots, with p-values indicating significance levels. The dashed line represents the chance accuracy of 50%.

resamples) (Figure 2). This result highlights the stronger predictive capability of using gaze direction (edge information) alone, suggesting that the simplified relational focus of the RNN Edge model captures the essential features needed for accurate social interaction prediction.

Relational Models Encode Behavior and Neural Social Information

In the behavioral encoding task, both SocialGNN and the RNN Edge model performed well on the "agent facing" and "communicating" behavioral dimensions (Figure 3). We conducted a similar two-tailed paired permutation test ($n = 10,000$ resamples) to compare the performance of SocialGNN, RNN Edge, and RNN Node models across various behavioral features. Notably, SocialGNN and RNN Edge encoded significantly more information in the "agent facing" ($p = 0.0002$) and "communicating" ($p = 0.0002$) behavioral ratings than the control model, RNN Node. Interestingly, SocialGNN and RNN Edge even outperformed the best vision model in predicting these features. In the "agent facing" rating, SocialGNN significantly outperformed RNN Edge ($p = 0.0062$), while RNN Edge showed a slight but non-significant advantage over SocialGNN in the "communicating" rating ($p = 0.4232$).

In contrast, features more focused on spatial and object properties were predicted equally well or better by the RNN Node model. For instance, RNN Node significantly outperformed both RNN Edge and SocialGNN in "spatial expanse" ($p = 0.0002$). RNN Node also showed a significant advantage over SocialGNN in "object directed" ($p = 0.0086$). In "interagent distance", RNN Node is significantly outperforming RNN Edge ($p = 0.0142$) and SocialGNN ($p = 0.0014$). This suggests that edge-based representations are particularly advantageous for encoding social relational behaviors, whereas node-based representations may be more effective for spatial and object-related features.

In the neural encoding analysis, our two-tailed paired permutation test ($n = 10,000$ resamples) showed that both the SocialGNN and RNN Edge encoding scores outperform those of RNN Node in the lateral stream, specifically the right hemisphere of pSTS and aSTS (pSTS_rh and aSTS_rh) (Figure 4). The strongest difference in pSTS_rh and aSTS_rh showed that both SocialGNN ($p = 0.0002$ for both hemispheres) and RNN Edge ($p = 0.0002$ for both hemispheres) exhibited highly significant advantages over RNN Node. In the left hemisphere pSTS (pSTS_lh), SocialGNN again showed a significant improvement over RNN Node ($p = 0.0088$), but RNN Edge did not ($p = 0.2394$). In the left hemisphere, aSTS (aSTS_lh), SocialGNN significantly outperformed RNN Node ($p = 0.0308$), as did RNN Edge ($p = 0.0434$). There were no significant differences between SocialGNN and RNN Edge in these regions ($p > 0.7$ in all cases). In contrast, the RNN Node either matched or outperformed SocialGNN and RNN Node in the ventral stream. For example, the RNN Node is most significantly better in the left hemisphere PPA (PPA_lh) than SocialGNN ($p = 0.0002$) and RNN Edge ($p = 0.0002$). In the early visual area, RNN Edge is significantly better than the RNN Node ($p = 0.0030$) in the left hemisphere of EVC (EVC_lh). RNN Node is slightly better than the RNN Edge in the right hemisphere of EVC (EVC_rh) but not significant ($p = 0.0516$).

To further investigate whether SocialGNN and RNN Edge encode complementary information relevant to neural responses, we combined the activation patterns of SocialGNN and RNN Edge and used these concatenated activations to fit a ridge regression model, following the same procedures described earlier, to predict behavioral and neural responses. However, our analysis revealed that the combined model representations did not outperform the better-performing model (either SocialGNN or RNN Edge) in any case (data not shown). This suggests that the two models encode largely overlapping information.

Generalization of Relational Cues to Physical Contact in PHASE

We next turned to the PHASE dataset. Our optimized SocialGNN and the original generative inverse planning model perform similarly at the level of human agreement, substantially better than RNN Node. The human agreement was calculated as the averaged ratio of the human ratings equal to the mode of the human ratings across videos (Malik & Isik, 2023). As with the natural videos, we again find that a similar performance advantage can be seen with the even simpler RNN Edge model, suggesting that the importance of relational cues

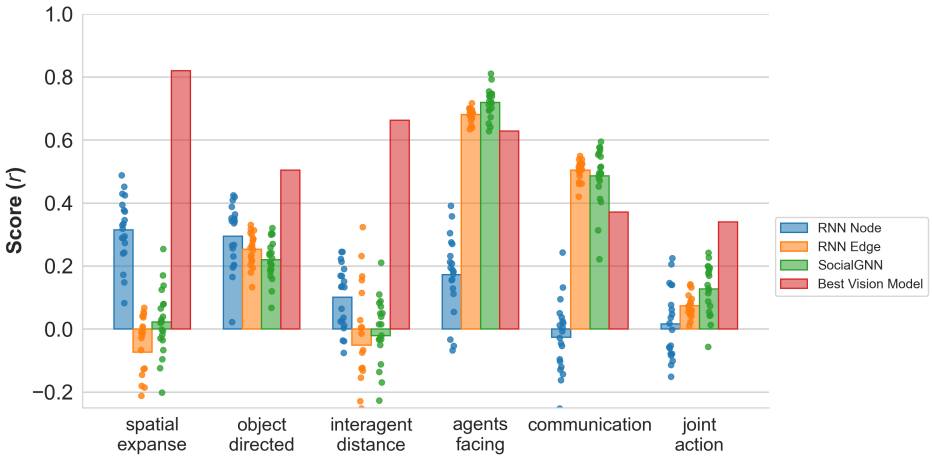


Figure 3: **Behavioral encoding scores.** The figure illustrates the behavioral encoding scores of four models: RNN Node, RNN Edge, SocialGNN, and the best vision model, DeiT3-L (Touvron et al., 2022), reported in Garcia et al. (2024). For RNN Node, RNN Edge, and SocialGNN, each dot represents a model trained using the different bootstrapped train-test splits, with 20 bootstraps per model type, and the bars denote the average performance. For the best vision model, DeiT3-L, the bar is the encoding score from this single model.

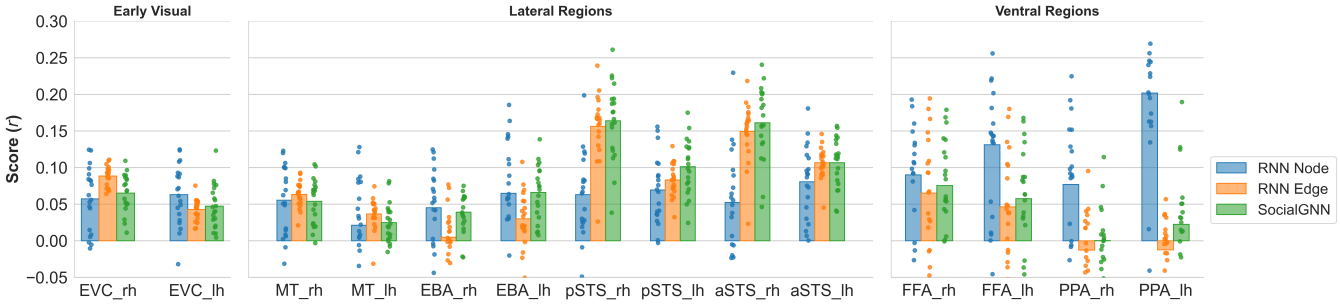


Figure 4: **Neural encoding scores.** The neural encoding scores of the three models: RNN Node, RNN Edge, and SocialGNN. Each dot represents a model trained using different bootstrapped train-test splits, with 20 bootstraps per model type. The bar for each model denotes the average performance.

generalizes beyond gaze direction to physical contact (Figure 5). This is particularly noteworthy given that the performance of these models reaches the level of human agreement, and the inverse planning model instantiates a physical simulation of the agent's goals and the physical world, requiring many orders of magnitude more memory and runtime than either neural network model (Malik & Isik, 2023).

Discussion

Our findings highlight the strengths and limitations of different neural network models in predicting and encoding social interactions across natural and animated datasets and provide valuable insight into the strategies humans use to recognize this information. Across three diverse datasets, we found that models designed to capture relational information aligned more closely with human social judgments than models lacking this edge information. Relatively simple models like SocialGNN and RNN Edge performed on par with generative inverse planning models for animated shapes and even exceeded the performance of SOTA vision models for natural videos. Furthermore, RNN Edge performed as well or better than SocialGNN across datasets, indicating that simplified relational representations are sufficient to explain complex human social judgments, and the importance of relational information generalizes across different stimuli and visual cues (gaze direction and physical contact).

While the RNN Edge model performs well across various tasks, the SocialGNN's performance may be constrained by its reliance on edge information and the limited utility of node data in these settings. This limitation arises from two factors: extracting node information is more complex than edge information, leading the model to rely heavily on the latter. The considerable overlap between node and edge features may result in redundancy that hinders the GNN's ability to learn and generalize to novel patterns. To enhance SocialGNN's performance, future efforts could replace standard VGG19 node features with distinct, theoretically grounded node features that complement edge information. For example, incorporat-

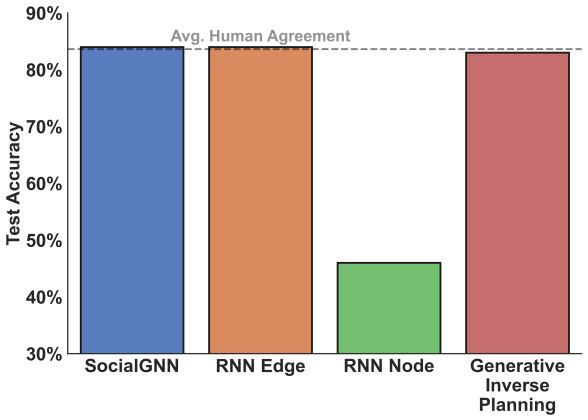


Figure 5: **Model accuracy on the PHASE dataset.** The bar graph of test accuracy of models predicting social interaction types (friendly, neutral, adversarial) in the PHASE generalization set. Four models (SocialGNN, RNN Edge, RNN node, and Generative Inverse Planning) were trained on the entire main dataset and evaluated on the generalization set. The dashed horizontal line is the average human agreement level.

ing diverse social cues as node features such as body pose, facial expressions, and dialogue analysis could provide unique multimodal inputs (Lee et al., 2024).

In the behavioral encoding analysis, the high scores for agent-facing and communicating suggest a strong reliance on gaze direction to recognize these social features. In contrast, metrics tied to human-object interactions (like "object-directed") or spatial relations (like "spacial expanse" and "interagent distance") exhibited worse performance. Thus, these behavior judgments may capture a distinct dimension of social interaction that depends on cues like body poses and standing positions rather than gaze. Incorporating this information in node features of a GNN as described above may improve SocialGNN's performance across the range of human social judgments and highlight the promise of SocialGNN (over the simple RNN Edge model) for providing a comprehensive model of human social judgments.

In the neural encoding analysis, SocialGNN and RNN Edge demonstrated a clear advantage in their capacity to encode activity in the right hemisphere of STS. This region is known to play a critical role in processing dynamic social information, such as interpreting the actions and interactions of others. The lateralization effect is consistent with the neural predictivity of our models, indicating that the SocialGNN and RNN Edge are capturing information relevant to social brain function. In contrast, SocialGNN and RNN Edge models do not show a very clear advantage in early visual regions and no advantage in the ventral stream, suggesting that these models are particularly adept at encoding higher-level relational social cognitive processes, rather than capturing lower-level features in the visual hierarchy.

The PHASE experiment demonstrated that interactions depicted in simple 2D displays can also be effectively encoded through relational information. However, models like SocialGNN and RNN Edge, which rely on bottom-up processing, may still struggle with interaction types that require understanding the intentions driving physical interactions. For instance, physical contact patterns alone are likely insufficient to understand the unintentional touches. Integrating these approaches with top-down strategies, such as generative inverse planning models that account for underlying intentions, could potentially enhance their ability to classify these social dynamics. These complementary methods, by combining relational cues with an understanding of the goals driving interactions, may further improve both classification accuracy and their applicability in encoding relational dynamics.

In conclusion, our study underscores the importance of relational information, such as gaze direction and physical contact, in understanding and predicting human social behaviors and neural responses and highlights lightweight strategies to improve SOTA AI models. Both the SocialGNN and RNN Edge models demonstrate strong capabilities in encoding social interactions, with the RNN Edge model excelling due to its simplified, targeted approach to relational cues. Future work incorporating diverse social cues, such as body poses, facial expressions, and dialogue, alongside a balance of bottom-up and top-down processes, could enhance human-aligned AI and provide deeper insights into the mechanisms underlying human social cognition.

References

Allison, T., Puce, A., & McCarthy, G. (2000, July). Social perception from visual cues: role of the STS region. *Trends Cogn Sci*, *4*(7), 267–278. doi: 10.1016/s1364-6613(00)01501-1

Bolotta, S., & Dumas, G. (2022, May). Social Neuro AI: Social Interaction as the "Dark Matter" of AI. *Front. Comput. Sci.*, *4*. (Publisher: Frontiers) doi: 10.3389/fcomp.2022.846440

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015, November). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cereb Cortex*, *25*(11), 4596–4609. doi: 10.1093/cercor/bhv111

Deen, B., Saxe, R., & Kanwisher, N. (2020, November). Processing communicative facial and vocal cues in the superior temporal sulcus. *NeuroImage*, *221*, 117191. doi: 10.1016/j.neuroimage.2020.117191

Fan, L., Wang, W., Huang, S., Tang, X., & Zhu, S.-C. (2019). Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 5724–5733).

Fox, E. (2005, January). The role of visual processes in modulating social interactions. *Visual Cognition*, *12*(1), 1–11. (Publisher: Routledge

_eprint: https://doi.org/10.1080/13506280444000067) doi: 10.1080/13506280444000067

Garcia, K., McMahon, E., Conwell, C., Bonner, M. F., & Isik, L. (2024, June). *Modeling dynamic social vision highlights gaps between deep learning and humans.* OSF. doi: 10.31234/osf.io/4mpd9

Hafri, A., & Firestone, C. (2021, June). The Perception of Relations. *Trends in Cognitive Sciences*, *25*(6), 475–492. doi: 10.1016/j.tics.2021.01.006

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017, October). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1714471114

Kreifelts, B., Ethofer, T., Shiozawa, T., Grodd, W., & Wildgruber, D. (2009, December). Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice- and face-sensitive regions in the superior temporal sulcus. *Neuropsychologia*, *47*(14), 3059–3066.

Lee, S., Li, M., Lai, B., Jia, W., Ryan, F., Cao, X., … Rehg, J. M. (2024, September). *Towards Social AI: A Survey on Understanding Social Interactions.* arXiv. (arXiv:2409.15316 [cs]) doi: 10.48550/arXiv.2409.15316

Malik, M., & Isik, L. (2023, November). Relational visual representations underlie human social interaction recognition. *Nat Commun*, *14*(1), 7317. (Publisher: Nature Publishing Group) doi: 10.1038/s41467-023-43156-8

McMahon, E., Bonner, M. F., & Isik, L. (2023, December). Hierarchical organization of social action features along the lateral visual pathway. *Curr Biol*, *33*(23), 5035–5047.e8. doi: 10.1016/j.cub.2023.10.015

McMahon, E., & Isik, L. (2023, December). Seeing social interactions. *Trends Cogn Sci*, *27*(12), 1165–1179. doi: 10.1016/j.tics.2023.09.001

Netanyahu, A., Shu, T., Katz, B., Barbu, A., & Tenenbaum, J. B. (2021, May). PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(1), 845–853. (Number: 1) doi: 10.1609/aaai.v35i1.16167

Papeo, L. (2020). Twos in human visual perception. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *132*, 473–478. (Place: France Publisher: Elsevier Masson SAS) doi: 10.1016/j.cortex.2020.06.005

Salatiello, A., Hovaidi-Ardestani, M., & Giese, M. A. (2021). A Dynamical Generative Model of Social Interactions. *Front Neurorobot*, *15*, 648527. doi: 10.3389/fnbot.2021.648527

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., … Ullman, T. D. (2021, July). *AGENT: A Benchmark for Core Psychological Reasoning.* arXiv. (arXiv:2102.12321 [cs]) doi: 10.48550/arXiv.2102.12321

Touvron, H., Cord, M., & Jégou, H. (2022, April). *DeiT III: Revenge of the ViT.* arXiv. (arXiv:2204.07118 [cs]) doi: 10.48550/arXiv.2204.07118