[PREPRINT] In: 2025 IEEE International Conference on Robotics and Automation (ICRA), Atlanta, GA, USA, 19 May - 23 May, 2025.

HRI-Free: Cognitive Robotic Simulation for Evaluating Embodied Social Attention Models

Fares Abawi[†] and Di Fu[‡]

***Abstract*— Scaling social robot studies is constrained due to the need for human interaction, making large participant recruitment impractical. Robotics simulators help mitigate this limitation but generally lack the realism to accurately simulate social cues. We introduce a cognitive robotic simulation scheme to evaluate social attention models in physical environments. By projecting ground-truth priority maps to a simulated environment, we can directly compare predicted maps using common saliency metrics. Using the iCub robot, we assess a dynamic scanpath model that predicts attention targets, simulating human scanpaths. Evaluations with the FindWho and MVVA datasets show strong correlations between robot-captured metrics and direct-streamed video metrics. Our results indicate robustness of the social attention model to noise and real-world conditions, suggesting its practical usability for predicting personalized scanpaths in real settings. This approach reduces the need for extensive human-robot interaction studies in the early stages of study design, enabling the scalability and reproducibility of social robot evaluations.**

I. INTRODUCTION

Scaling social robot studies is challenging since most depend on human perception arising from interactions with the robots. Recruiting a large number of human participants to conduct these studies is generally impractical in terms of time and resource investment. Robotics simulators have emerged as a solution to the scaling problem in the social robotics sphere. However, although the physics and aesthetic realism of robotics simulators have been advancing rapidly, the fidelity of social cue (socialness) simulation is still limited. Automating the testing of embodied social models, like social navigation, is made possible with large generative language and multimodal models. Such approaches rely on simulating social behavior in the form of abstract action primitives using the generative models [3]. However, the level of abstraction and quality of generated outputs could misrepresent real-world conditions under which robots operate, potentially leading to inaccurate and unsafe behavior [4].

Moreover, the spectrum of social cues displayed by humans during social interactions is broad and context-dependent. These cues include facial expressions, gaze direction, and others [5]. Social cues are especially relevant when evaluating social attention models. Social attention—saliency and scanpath prediction—models predict human gaze by integrating social cues. To address the limitation of social cue simulation, we present a cognitive robotic simulation scheme by matching the observations in the physical environment onto a simulated one. Using the iCub robot, we evaluate a dynamic scanpath model that predicts attention targets on the MVVA [2] and FindWho [1] datasets, simulating human-like scanpaths, as shown in Figure 1.

(a) MVVA Evaluation

(b) FindWho Evaluation

Fig. 1. The iCub robot executing the evaluation pipeline based on the **(a)** FindWho [1] dataset allowing eye for movements only, and **(b)** MVVA [2] dataset allowing for head and eye movements. The physical robot *(bottom right)* observes clips *(left)* and predicts a priority map *(top right)* according to the observer under test. The ground-truth priority map is projected to a monitor in simulation *(center right)*, after which the projected ground-truth and predicted maps are compared in terms of the NSS and AUCJ metrics.

We ensure that we maintain similarity to the human data collection setup while accommodating technical limitations. Additionally, we assume that the physical environment provides a means for allowing the robot to perceive the pre-recorded stimuli. Our approach, although tailored specifically for social attention models, can be applied to other social tasks with varying degrees of complexity, such as behavior mirroring [6] and social cueing [7].

II. RELATED WORK

In social robotics, developing appropriate embodied social attention models contributes significantly to enabling robots

[†]Fares Abawi is with the Department of Informatics, University of Hamburg, Hamburg, Germany. fares.abawi@ieee.org
[‡]Di Fu is with the School of Psychology, University of Surrey, Guildford, UK. d.fu@surrey.ac.uk

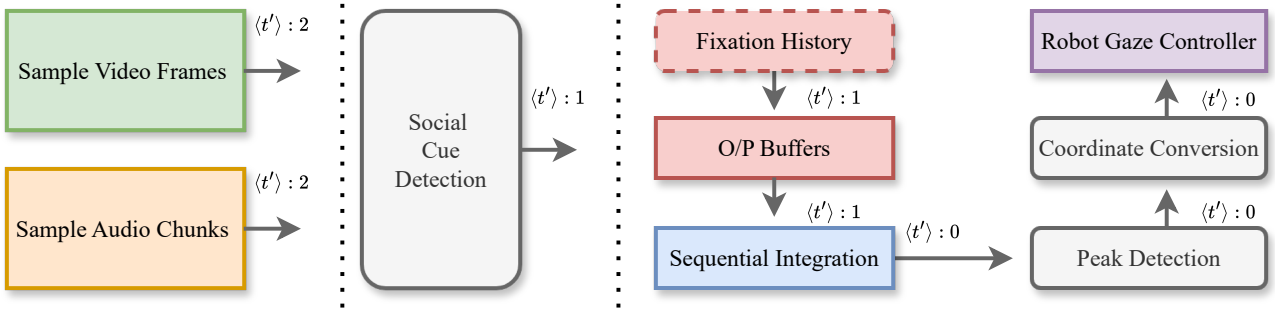


Fig. 2. The scanpath control pipeline for actuating the robot. Assuming the current timestep at $\langle t' \rangle : 0$, we show the output availability for each component at the relative timestep. Video and audio sampling are performed in parallel, blocking all other components to avoid interrupting the real-time capture. At any time point, the video and audio sampling is two timesteps ahead of the controller that actuates the robot in physical and simulated environments.

to display human-like social behaviors. These models aim to replicate human attention mechanisms, allowing robots to engage in more realistic and effective social interactions [8].

One of the main challenges is the collection of naturalistic human-robot interaction data. Traditionally, this has been accomplished through labor-intensive studies involving multiple participants. Human data are typically used to train the attention model [9] or serve as ground-truth for evaluating the model's performance [7], [10]. However, the scalability of such approaches is limited. When algorithms require modification, researchers often have to re-design and repeat the study with human participants, rendering the evaluation process inefficient and difficult to scale. To address this issue, our study draws on methods from social navigation [11], [12], using robotics simulators to generate synthetic data for training and evaluating social attention models. Contrary to such approaches where the simulator is used to generate synthetic data, our approach relies on replicating the same environment in digital and physical form. The simulator projects the sample ground-truth onto physically identical locations as they appear in the real environment, allowing for the evaluation of the predictions in the physical world.

Additionally, replicating how humans prioritize visual and auditory information during social interactions to predict gaze behaviors [10] remains a challenge. Fu et al. [7] develop an audiovisual saliency prediction model that can resolve modality incongruencies and attend selectively according to the task. This work provides the research basis of our current robotic simulation scheme.

Another challenge is detecting various social cues and integrating them during attention allocation [13]. Social cues such as facial expressions, gaze direction, and body language may impact social interaction differently [14]. In previous work, Abawi et al. [15] developed a personalized attention model that integrates several of such cues, enabling the prediction of individuals' scanpaths in social settings. We adopt this model in our current work, demonstrating the applicability of our cognitive robotic simulation framework for evaluating such social attention models in physical settings, without requiring humans to assess the model's performance.

III. Methods

We evaluate a dynamic scanpath prediction model [15] designed to infer priority maps, highlighting an individual observer's attention target. The peak of the priority map defines the gaze target. Actuating a robot to gaze toward that target, as well as the targets to follow in sequence, would effectively simulate human scanpaths. Our approach involves the projection of ground-truth priority maps to a simulated environment. The map is projected to a monitor within the simulator at a distance from the simulated robot, approximately equivalent to the distance of a real monitor from the physical robot. By controlling the gaze of the simulated robot to match the physical robot, the view of the ground-truth priority map resembles the view of the physical robot's predicted map. This allows us to compare the ground-truth to the predicted map using saliency metrics [16].

A. Cognitive Robotic Simulation Scheme

An overview of our robot control pipeline is shown in Figure 2. The pipeline defines the steps taken to evaluate our unified scanpath models in real-time. The components of the pipeline are detailed as follows:

1) Audio and Video Sampling: Initially, the videos are played on the physical monitor in one-second chunks and are paused until the pipeline repeats. During playback, the iCub robot facing the monitor captures a sequence of images at 10 FPS and audio at 16 kHz. The video playback is executed as a separate process that awaits a signal to resume. This signal is transmitted before the iCub begins capturing one-second chunks of audiovisual frames using its integrated sensors. The communication between the sampling and playback processes is handled by *Wrapyfi* [17]. This step is performed in a blocking manner to avoid interruptions to the capturing process.

2) Social Cue Detection: We utilize the social cue detectors proposed by Abawi *et al.* [15]. This includes the facial expression [18] and gaze estimation [19] cue modalities, along with the saliency prediction [20] modality. The cues are detected, transformed, and represented following the procedure detailed by Abawi *et al.* [21]. The cue detectors

extract the representations sequentially and maintain frames and chunks from previously sampled video and audio. As long as the same video is running, the frames are queued and processed by the detectors according to their context lengths. At the beginning of a video, the frames collected are not sufficient to cover the context length of all detection models. For instance, the DAVE saliency prediction model requires 16 video frames, however, our samplers return 10 frames only. The remaining 6 frames would be padded with the last acquired frame and shifted as more samples are collected. At every timestep, the detected representations are propagated to the output buffers as single 2D representations per modality.

3) Fixation History: The fixation history is the sequence of fixations that precede the one being predicted by the sequential integration model [15] in the form of a priority map. The fixation history serves the purpose of providing context to our model, in order to inform it on the observer priority map to be predicted. Moreover, the next fixation depends on the previous fixation positions. Without representing the previous scanpath—sequence of fixations—the predictions would be arbitrary. The fixation history module extracts the ground-truth priority map for a given timestep t' and propagates it to the output buffers.

4) Output Buffers: Output buffers represent all queues storing the latest state representations for each modality, agnostic to the input sampling mechanism. At every timestep t', each modality-specific buffer is queued with a single 2D feature map. The maximum size for all queues is governed by the context size of the sequential integration model.

5) Sequential Integration: We use the scanpath model developed in [15]. We employ two models trained with the FindWho [1] and MVVA [2] observer data. More specifically, we evaluate the unified integration model. The unified integration model is similar in structure to the sequential integration GASP [21] variant, additionally extended with the fixation history module. The Directed Attention Module (DAM) is trained on the fixation density maps of a group of observers, whereas the Late Attentive Recurrent Gated Multimodal Unit (LARGMU) is trained on the priority maps of all observers individually.

6) Peak Detection and Coordinate Conversion: The social cue detectors and saliency predictor extract features for the previously acquired audiovisual frames during the auditory and visual acquisition phase. Following the detection and generation of spatiotemporal maps, the unified scanpath model predicts a priority map $\hat{\mathbf{m}}^{\langle t \rangle} \colon \mathbb{Z}^2 \rightarrow [0,1]$ for a given frame. The peak is registered in pixel coordinates and remapped to scalar values within the range of $\in \{-1, 1\}$ in both x and y axes, such that:

$$\hat{\mathbf{p}}_{x,y} = -1 + \frac{2 \cdot \arg\max_{x,y} \hat{\mathbf{m}}(x,y)}{\hat{\mathbf{m}}_{X,Y}}, \tag{1}$$

where $\hat{\mathbf{p}}_{x,y}$ represents the peak location in the normalized range and $\hat{\mathbf{m}}_{X,Y}$ are the width and height of the predicted priority map in pixels. We actuate the robot to look toward the peak. For simplicity, we assume the camera view to be independent of its location relative to the playback monitor. For all experiments, we control the head and eye movements of the iCub, disregarding vergence effects, microsaccades, and fixation duration. The positions are expressed in Cartesian coordinates, assuming the monitor to be at a distance of $\sim \delta_z$ from the image plane. We scale $\hat{\mathbf{p}}_{x,y}$ by a factor of $\alpha_{x,y} = \{.35, .3\}$ to limit the viewing range of the eyes. We then convert the Cartesian to spherical coordinates:

$$\begin{aligned} \hat{\mathbf{p}}_{\phi} &= \operatorname{atan}\left(\frac{\alpha_y \cdot \hat{\mathbf{p}}_y}{\delta_z}\right), \\ \hat{\mathbf{p}}_{\theta} &= \operatorname{atan}\left(\frac{\alpha_x \cdot \hat{\mathbf{p}}_x}{\sqrt{{\delta_z}^2 + (\alpha_y \cdot \hat{\mathbf{p}}_y)^2}}\right), \end{aligned} \tag{2}$$

where $\hat{\mathbf{p}}_{\phi}$ and $\hat{\mathbf{p}}_{\theta}$ are the pitch and yaw angles respectively. These angles are used to actuate the eyes of the iCub such that they tilt $\sim 24^\circ$ and pan $\sim 27^\circ$ at most[1]. On extracting the output priority map from the scanpath prediction model, the peak of the priority map is registered as the target of gaze. The robot captures the images and audio from the environment, applies the scanpath prediction model to the captured stimuli, and directs the robot's gaze toward the peak. Simultaneously, the ground-truth priority map is projected to a monitor within a simulated environment as shown in Figure 3, and the peak of that map is detected relative to the monitor. Finally, the predicted priority map is evaluated against the simulator-projected ground-truth map.

7) Robot Gaze Controller: The iCub [22] robot is used in all experiments for evaluating performance on the MVVA [2] and FindWho [1] datasets as shown in Figure 1. The MVVA data collection procedure does not enforce fixing the head pose. To accommodate the influence of the head rotation, we utilize the *iKin* [23] library. More specifically, we aim to evaluate gaze shifts by relying on the iCub robot's vestibulo-ocular reflex functionality to compensate for the head movements resulting from fixating on a target location. The integration of such an effect is necessary due to its impact on stimuli capture as well as the fixations following the current at any given timestep. For the FindWho evaluation trials, the head pose is fixed such that the iCub's line-of-sight is perpendicular to the monitor. We, therefore, control the eyes directly by specifying the target of gaze as the peak of the predicted priority map in the visible pixel space.

B. Experimental Design

1) Mapping Prediction to Ground-Truth Gaze: Videos displayed on a monitor would naturally require a different ground-truth mapping methodology to streamed video comparisons. To circumvent the ambiguity in mapping fixation positions, we project the ground-truth density map onto a monitor within the iCub simulator. The iCub robot was chosen since it is capable of moving both its head and eyes, with cameras attached to its pupils and microphones mounted on both its ears. This structure resembles the anatomy of a human, enabling us to assimilate the human data collection setup as closely as possible. In Figure 1 we

[1]The iCub can tilt and pan its eyes in ranges of $\in \{-40^\circ, 40^\circ\}$ and $\in \{-45^\circ, 45^\circ\}$ respectively.

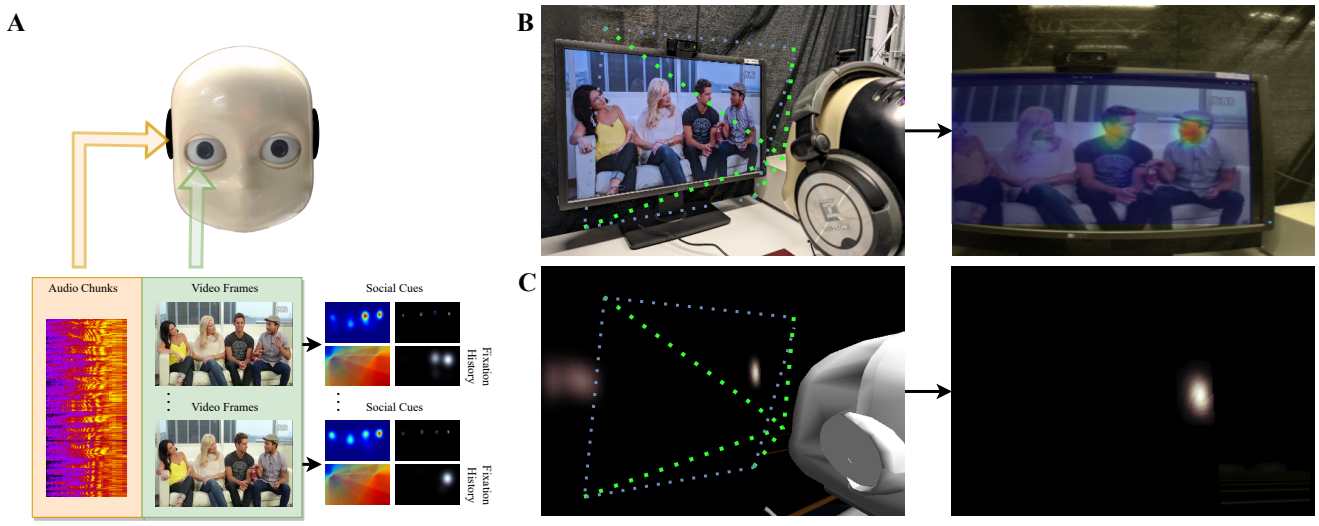


Fig. 3. The videos are played back in segments on a monitor facing the iCub robot. **(A)** Audio chunks and video frames are captured through the iCub's sensors. The social cue and saliency features are represented as 2D maps and propagated to the unified scanpath model, which predicts a priority map. The **(B)** external view of the physical environment *(left)* and the region of capture *(right)* upon which the priority map is inferred, and **(C)** captured camera view from the simulated environment *(right)*, in which the ground-truth priority map is projected on a virtual monitor followed by color-correction and evaluation against the inferred map are shown.

show the social attention model employed on the physical iCub robot. Knowing the robot's distance from the monitor within the physical environment, we mirror the head and eye movements of the physical robot within the simulator, providing an approximate position of the intended fixation.

We adjust the ground-truth priority maps to match the size of the monitor in the physical environment from the perspective of the observer. Given the distance from the monitor δ_z during the data collection phase, we can approximate the width and height of the projected ground-truth map by repositioning the simulated monitor at a distance of δ_z from the robot. Next, the simulated monitor is resized to match the size of the physical one and the view from the robot's camera captured. Finally, the simulated capture is compared to the predicted priority map from the physical environment.

2) Experimental Setup: In this study, we utilized the pretrained unified scanpath model with the best performance [15]: The integration architecture (DAM + LARGMU, context size $T' = 10$), yielded the best results for a majority of the experiments on both the MVVA [2] and FindWho [1] datasets. Given the procedural differences in the collection of the MVVA and FindWho datasets, we considered the properties shown in Table I. However, accounting for the robot's visual field and camera resolution, we did not fully align our setup with those properties. For evaluating the MVVA dataset, the corresponding integration model was deployed on the robot. We placed the robot at a distance of ~ 30 cm from a 23-inch monitor. For the FindWho dataset evaluation, we moved the robot further from the monitor to a distance of ~ 35 cm, and deployed the integration variant of our scanpath prediction model, trained on the FindWho dataset. In alignment with the datasets' collection protocols, we set the robot to move its eyes only when evaluating the FindWho dataset. As for the MVVA dataset

TABLE I

EXPERIMENTAL SETUP AND DATASET PROPERTIES.

Property	*MVVA* [2]	*FindWho* [1]
Distance to monitor	~ 55 cm	~ 60 cm
Monitor resolution	1280×720 px (16:9)	1280×720 px (16:9)
Monitor size	23-inch	23.8-inch
Video duration	10-30s	~ 20s
Frames per second	30	25
Audio channels	Stereo	Monaural
Head-pose	Free	Fixed
No. training videos	210 (70%)	46 (70%)
No. validation videos	30 (10%)	-
No. test videos	60 (20%)	19 (30%)
No. observers	34 (1 excl.)	39

evaluation, we used the *iKin* [23] library to direct the robot's gaze shift through head and eye movements. Both datasets were evaluated separately. The stimuli videos were replayed a number of times equivalent to the number of individual observers. For each observer, the fixation history consisted of the preceding ground-truth fixations on observing the specific video frames and audio chunk. The videos were played back at 1 s intervals and captured using the iCub robot's left camera. Audio was played back also for 1 s intervals through on-ear headphones, placed on the iCub robot's microphones.

During the physical evaluation, the social cue detectors and the GASP model were distributed among two NVIDIA GeForce GTX 970 GPUs with a total of 8GB VRAM and 32GB RAM. Experiments on the physical robot, evaluating all observers individually required ~ 13 hours in total for the FindWho dataset (39 observers, 19 videos), and ~ 42 hours for the MVVA dataset (34 observers, 60 videos).

Videos in the MVVA [2] and FindWho [1] datasets are played on a monitor facing the iCub robot [22]. These datasets are composed of social videos that were watched

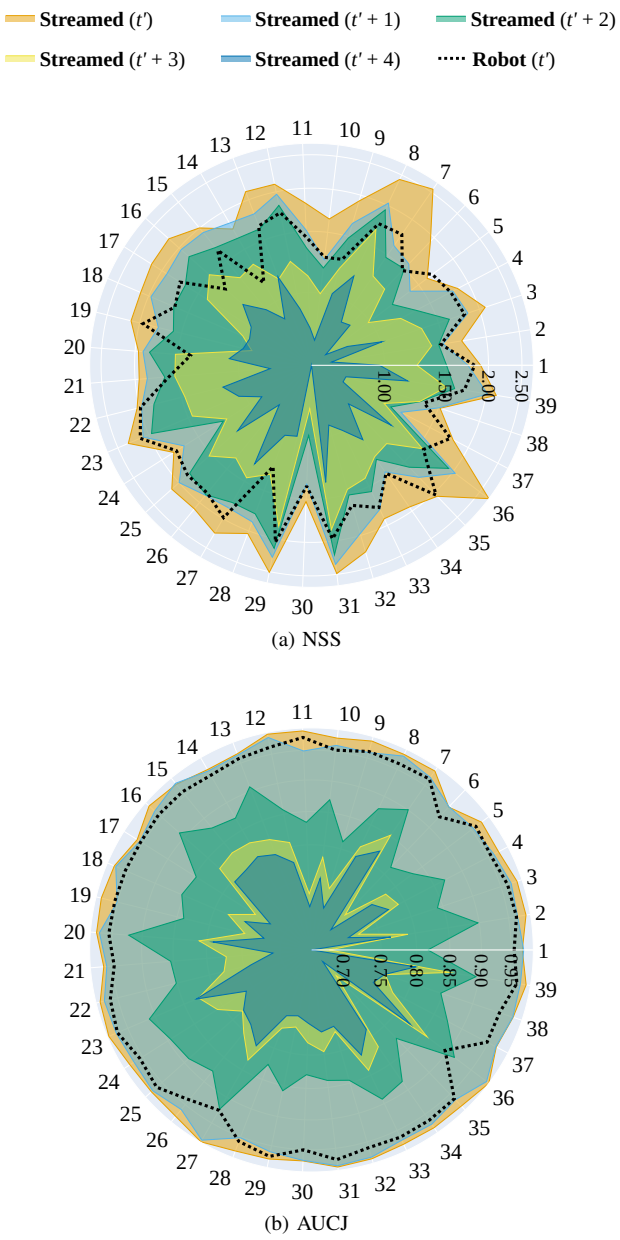
under the free-viewing condition [24, p. 26] by multiple human observers, whose eye movements were collected using an eye tracker. These datasets contain social videos, making them suitable for social attention model evaluation. Moreover, these datasets explicitly label the samples by the observer, allowing us to compare the model's performance across individuals. The robot captures those videos with its camera and microphones. Following capture, the social cue and saliency prediction modalities are executed, and their representations are generated. These representations are queued in the output buffer along with the fixation history—the preceding fixations of the observer under test. Concurrently, the sequential integration model operates on the representations of previous timesteps and predicts an individual observer's priority map. The predicted map is propagated to the peak detector. The peak coordinates are converted to yaw and pitch, which are then used to actuate the physical and simulated robot simultaneously using *YARP* [25]. The ground-truth priority map for the last video frame of a given context is channeled to the simulated monitor as shown in Figure 3. Finally, the metrics are computed, and the pipeline is looped until all videos in the evaluation set are completed.

IV. Results

Pearson correlation analyses were conducted to assess the alignment between robot-captured metrics and direct-streamed video metrics over one-step and multi-step-ahead time intervals. The robotic experiments were only conducted for one-step-ahead predictions. Multi-step-ahead predictions refer to evaluations extending over multiple future steps. This describes feeding the predicted priority map back into the fixation history module as future samples are collected, where every additional step into the future is denoted by $t' + N$. Here, t' refers to one-step-ahead prediction and N to the number of additional steps ahead.

We evaluated the FindWho [2] dataset on the robot, and measured its performance in terms of the NSS and AUCJ metrics against one-step-ahead and multi-step-ahead streamed video predictions as shown in Figure 4. For the NSS metric, moderate correlations were observed between the robot-captured and direct-streamed videos (r = 0.498), which decreased with the addition of the steps ahead (r = 0.442 at $t'+1$, r = 0.401 at $t'+2$, r = 0.279 at $t'+3$, and r = 0.336 at $t'+4$). In contrast, the AUCJ metric exhibited a weak initial correlation (r = 0.165) that turned negative for future predictions (r = -0.142 at $t'+1$ through r = -0.098 at $t'+4$), indicating a divergence in attention distribution metrics with step-ahead increments.

We evaluated the MVVA [2] dataset on the robot, and measured its performance in terms of the NSS and AUCJ metrics against one-step-ahead and multi-step-ahead streamed video predictions as shown in Figure 5. For the NSS metric, strong correlations were observed between the robot-captured and direct-streamed videos, starting at r = 0.76 for one-step-ahead, with a gradual decrease through the steps ahead (r = 0.74 at $t'+1$, r = 0.68 at $t'+2$, and r = 0.63 at $t'+3$). For



(a) NSS

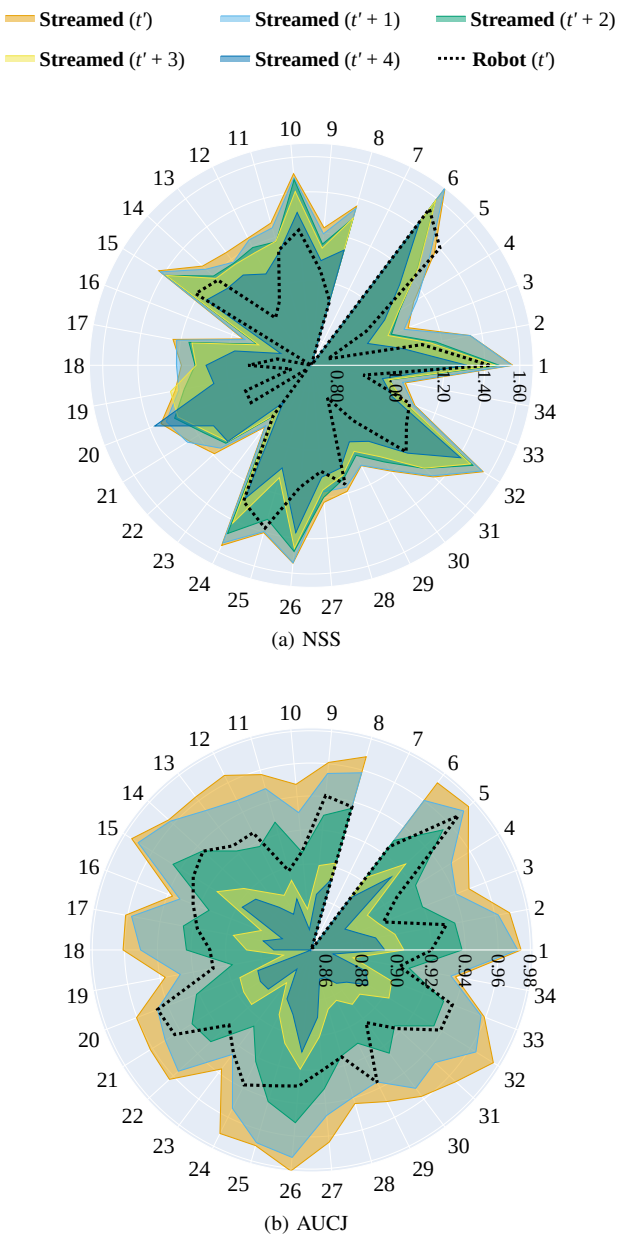
(b) AUCJ

Fig. 4. Robot one-step-ahead predictions using the Unified late integration model (DAM + LARGMU, context size $T' = 10$) trained on the FindWho [1] dataset, compared to the streamed multi-step-ahead predictions in terms of the **(a)** NSS and **(b)** AUCJ metrics. The angular axis indicates the observer identifier, whereas the radial axis shows the metric score.

the AUCJ metric, a moderate initial correlation (r = 0.48) was observed, which gradually increased for future predictions (r = 0.52 at $t'+1$, r = 0.57 at $t'+2$, and r = 0.59, $t'+3$, and r = 0.59 at $t'+4$), suggesting a strengthening of alignment in visual attention metrics with step-ahead increments.

V. Discussion

Expectedly, the robot scored lower than the streamed video evaluation in terms of the NSS and AUCJ metric scores. We found that the observer scores were correlated for streamed and robot videos, when evaluating on the MVVA dataset. The correlation was weaker on the FindWho dataset, even



(a) NSS

(b) AUCJ

Fig. 5. Robot one-step-ahead predictions using the Unified late integration model (DAM + LARGMU, context size $T' = 10$) trained on the MVVA [2] dataset, compared to the streamed multi-step-ahead predictions in terms of the **(a)** NSS and **(b)** AUCJ metrics. The angular axis indicates the observer identifier, whereas the radial axis shows the metric score.

tending towards negative values for the AUCJ correlation, as the number of steps ahead were increased. The AUCJ metric is sensitive to false-positive predictions. When training and evaluating on a relatively small dataset, the mean scores are higher, however, the variance is significantly larger. The unified model trained on the FindWho dataset observed fewer universal attention patterns, biasing the model more toward the scanpaths of the individual (personalized attention). This resulted in fewer erroneous predictions as evident from the higher NSS score. However, given the small size of the dataset, some observers' scanpaths were not learned sufficiently, while others were more similar to the average among the group of observers, and therefore, predicted accurately.

As for the MVVA dataset, which is approximately three times as large as the FindWho dataset, the unified model was exposed to more universal attention patterns. We saw that the robot's predicted gaze was robust to noise, as the scores of all participants were highly correlated with one-step-ahead and multi-step-ahead predictions, both in terms of NSS and AUCJ. Moreover, the larger size of the dataset meant that the evaluation was a better representative of the model's performance compared to the smaller FindWho dataset.

We conclude that the unified model is robust to noise since the input arriving from the robot's sensors differed to a large degree from the streamed videos. The lighting effects, distractors, and lower resolution were not very detrimental to the robot's performance, suggesting that our social attention model can be used in real-world settings to predict personalized scanpaths. We presume, however, that increasing the number of steps ahead during the robot evaluation would reduce its performance further. This reduction in performance was observed for the streamed videos under multi-step-ahead evaluation, suggesting a similar pattern for the robot as well. Our cognitive robotic simulation approach makes it possible to improve our social attention models and evaluate their performances on a physical robot or possibly several robotic platforms without needing to conduct HRI studies. This has the advantage of enabling reproducible experiments and scaling experiments beyond what is possible through HRI.

The scanpath prediction model in this study is subject to limitations related to intrinsic factors like camera resolution, microphone sensitivity, and external factors such as lighting, motion blur, and background noise, which can affect prediction accuracy in real-world settings. Additionally, the structure of the data acquisition and execution pipelines influences model performance, with delays in data processing or robot response times potentially hindering real-time execution. Despite these challenges, the approach holds significant potential for future applications. It can be adapted to predict personalized scanpaths for enhanced human-robot interaction in areas like education and healthcare. Furthermore, by refining the simulation framework and incorporating more sensory modalities, the model could be extended to simulate complex social scenarios, enabling scalable and reproducible evaluations without extensive human involvement.

VI. CONCLUSION

Our method addressed the limitations of traditional social robot studies, which often rely on human participants and are difficult to scale. By applying cognitive robotic simulation, we were able to conduct reproducible and scalable evaluations, reducing the need for HRI studies in the initial phases. This approach allows for the refinement and validation of social attention models in controlled settings, ensuring their suitability for real-world applications.

ACKNOWLEDGMENT

We thank Prof. Dr. Stefan Wermter (University of Hamburg) for providing the equipment used in this study.

REFERENCES

[1] M. Xu, Y. Liu, R. Hu, and F. He, "Find Who to Look At: Turning from Action to Saliency," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4529–4544, 2018.

[2] Y. Liu, M. Qiao, M. Xu, B. Li, W. Hu, and A. Borji, "Learning to Predict Salient Faces: A Novel Visual-Audio Saliency Model," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 12365. Springer, 2020, pp. 413–429.

[3] S. R. Marpally, P. Goyal, and H. Soh, "Towards Automated Scenario Testing of Social Navigation Algorithms," *Unsolved Problems in Social Robot Navigation Workshop at RSS2024*, 2024, https://unsolvedsocialnav.org/papers/Marpally.pdf.

[4] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh, "Core Challenges of Social Robot Navigation: A Survey," *Journal of Human-Robot Interaction*, vol. 12, no. 3, 2023.

[5] S. Wang and R. Adolphs, *Social Saliency*. Springer, 2017, pp. 171–193.

[6] D. Fu, F. Abawi, P. Allgeuer, and S. Wermter, "Human Impression of Humanoid Robots Mirroring Social Cues," in *Companion of the ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion)*. ACM, 2024, pp. 458–462.

[7] D. Fu, F. Abawi, H. Carneiro, M. Kerzel, Z. Chen, E. Strahl, X. Liu, and S. Wermter, "A Trained Humanoid Robot can Perform Human-Like Crossmodal Social Attention and Conflict Resolution," *International Journal of Social Robotics*, vol. 15, pp. 1325–1340, 2023.

[8] Y. Fang, J. M. Pérez-Molerón, L. Merino, S.-L. Yeh, S. Nishina, and R. Gomez, "Enhancing Social Robot's Direct Gaze Expression through Vestibulo-Ocular Movements," *Advanced Robotics*, pp. 1–13, 2024.

[9] G. I. Parisi, P. Barros, D. Fu, S. Magg, H. Wu, X. Liu, and S. Wermter, "A Neurorobotic Experiment for Crossmodal Conflict Resolution in Complex Environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2330–2335.

[10] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan, "A Multimodal Saliency Model for Videos with High Audio-Visual Correspondence," *IEEE Transactions on Image Processing*, vol. 29, pp. 3805–3819, 2020.

[11] R. Möller, A. Furnari, S. Battiato, A. Härmä, and G. M. Farinella, "A Survey on Human-aware Robot Navigation," *Robotics and Autonomous Systems*, vol. 145, p. 103837, 2021.

[12] L. J. Manso, P. Nuñez, L. V. Calderita, D. R. Faria, and P. Bachiller, "Socnav1: A Dataset to Benchmark and Learn Social Navigation Conventions," *Data*, vol. 5, no. 1, p. 7, 2020.

[13] S. Vinanzi, A. Cangelosi, and C. Goerick, "The Role of Social Cues for Goal Disambiguation in Human-Robot Cooperation," in *2020 29th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, 2020, pp. 971–977.

[14] J. Kajopoulos, G. Cheng, K. Kise, H. J. Müller, and A. Wykowska, "Focusing on the Face or Getting Distracted by Social Signals? The Effect of Distracting Gestures on Attentional Focus in Natural Interaction," *Psychological Research*, vol. 85, pp. 491–502, 2021.

[15] F. Abawi, D. Fu, and S. Wermter, "Unified Dynamic Scanpath Predictors Outperform Individually Trained Neural Models," *arXiv Preprint*, vol. abs/2405.02929, 2024.

[16] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What Do Different Evaluation Metrics Tell Us About Saliency Models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2019.

[17] F. Abawi, P. Allgeuer, D. Fu, and S. Wermter, "Wrapyfi: A Python Wrapper for Integrating Robots, Sensors, and Applications Across Multiple Middleware," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 2024, pp. 860–864. [Online]. Available: https://wrapyfi.readthedocs.io

[18] H. Siqueira, S. Magg, and S. Wermter, "Efficient Facial Feature Learning with Wide Ensemble-Based Convolutional Neural Networks," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, 2020, pp. 5800–5809.

[19] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically Unconstrained Gaze Estimation in the Wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 6912–6921.

[20] H. R. Tavakoli, A. Borji, J. Kannala, and E. Rahtu, "Deep Audio-Visual Saliency: Baseline Model and Data," in *ACM Symposium on Eye Tracking Research and Applications (ETRA)*, ser. ETRA '20 Short Papers. ACM, 2020, pp. 1–5.

[21] F. Abawi, T. Weber, and S. Wermter, "GASP: Gated Attention for Saliency Prediction," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI Organization, 2021, pp. 584–591.

[22] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. Von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano, "The iCub Humanoid Robot: An Open-Systems Platform for Research in Cognitive Development," *Neural Networks*, vol. 23, no. 8-9, pp. 1125–1134, 2010.

[23] A. Roncone, U. Pattacini, G. Metta, and L. Natale, "A Cartesian 6-DoF Gaze Controller for Humanoid Robots," in *Robotics: Science and Systems (RSS)*, vol. 2016, 2016.

[24] R. P. Van Gompel, *Eye Movements: A Window on Mind and Brain*. Elsevier, 2007.

[25] G. Metta, P. Fitzpatrick, and L. Natale, "YARP: Yet Another Robot Platform," *International Journal of Advanced Robotic Systems*, vol. 3, no. 1, p. 8, 2006. [Online]. Available: https://www.yarp.it