SOAP.AI: A Collaborative Tool for Documenting Human Behavior in Videos through Multimodal Generative AI

Qingxiao Zheng
University at Buffalo, SUNY;
University of Illinois
Urbana-Champaign, USA
qingxiao@buffalo.edu

Parisa Rabbani
University of Illinois
Urbana-Champaign, USA
rabbani8@illinois.edu

Yu-Rou Lin
National Yang Ming Chiao Tung
University, China
wn4doe.cs09@nycu.edu.tw

Daan Mansour
University of Illinois
Urbana-Champaign, USA
dmanso2@illinois.edu

Yun Huang
University of Illinois
Urbana-Champaign, USA
yunhuang@illinois.edu

ABSTRACT

Large Multimodal Models offer new opportunities for analyzing human activities and social behavior in fields requiring expert knowledge. Their in-context learning and adaptive abilities make customization possible for experts without coding skills. This paper introduces *SOAP.AI*, a collaborative tool facilitating experts to analyze human behaviors using AI. *SOAP.AI* is designed to foster a sense of ownership during human-AI collaboration, encouraging task modifications and evaluations to meet diverse goals. For instance, teaching AI to recognize behavioral nuances in autistic individuals could enhance AI's inclusion and value alignment. Our demonstration will engage CSCW researchers and HCI practitioners to discuss the design of collaborative AI systems for behavioral insights generation in various settings, such as medical settings, sports, social media, education, home care, and more.

CCS CONCEPTS

• **Human-centered computing → Systems and tools for interaction design**; *Collaborative and social computing systems and tools*; **Interactive systems and tools**; • **Information systems → Multimedia information systems**.

KEYWORDS

Behavior analysis, vision-language models, generative AI, collaborative work, videos

ACM Reference Format:
Qingxiao Zheng, Parisa Rabbani, Yu-Rou Lin, Daan Mansour, and Yun Huang. 2024. SOAP.AI: A Collaborative Tool for Documenting Human Behavior in Videos through Multimodal Generative AI. In *Companion of the 2024 Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24), November 9–13, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3678884.3681819

CSCW Companion '24, November 9–13, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-1114-5/24/11
https://doi.org/10.1145/3678884.3681819

1 INTRODUCTION

Human social behaviors, enriched by complex social cues, e.g., body language, gestures, facial expressions, and eye contact, have driven research interest in human activity analysis and social signal processing [2] , which aims to automatically detect, interpret, and synthesize various social cues [42]. These insights are crucial for supporting targeted interventions [19]. For example, in special education, experts rely heavily on detailed analyses and documentation of clients' social behavior, such as engagement, emotional status, and verbal expressions during therapy or coaching sessions [5, 9]. Such documentation, often referred to as *"invisible workload"* [26, 35], is critical for tracking clients' progress. Effective documentation can provide comprehensive evaluations and further develop effective therapeutic strategies. However, it is challenging to finish documentations because it requires intensive manpower and takes a large amount of time [20, 21, 34, 40].

Machine-learning models in human social behavior analysis can detect basic nonverbal cues but often struggle to adjust to the unique traits of individual users, specific groups, or social contexts [1, 8]. For example, these models can identify a person touching their face by analyzing spatial relationships among body parts [7], and determine the cause of action—whether it's due to thinking, embarrassment, or imitation [17]. However, they require extensive retraining to adapt to new or changing contexts. This limitation hampers non-coding experts, like clinicians and coaches, from effectively using these tools and exploratory analysis on what and how to document using these models. It also deepens the divide between domain experts and technology in application development.[39].

Existing tools for behavior analysis in social interactions primarily focus on individual or dyadic interactions, such as Bedmutha et al.'s ConverSense [6], which tracks audio-based social signals during patient-provider interactions, and Patel et al.'s system [36] that captures nonverbal cues for real-time clinical feedback to enhance empathic, patient-centered care. Similarly, Arakawa and Yakura [4] use multimodal signals of gaze to identify anomalies in coach-coachee interactions. Research also extends to group dynamics, with studies like Willenbrock and Hung [28] on team cohesion and Samrose et al.'s MeetingCoach [37], which provides meeting dashboards of transcripts and behavioral cues. However, existing solutions are limited by the lack of flexibility because they only support one-shot scene analysis (e.g., looking for moments when speech speed is high) without allowing for more customized analysis. Also, most of them are visualization-based, using timelines to highlight moments of specific behaviors, which fails to support comprehensive documentation or inspire deeper analysis.

The recent advent of generative AI (GenAI), particularly Large Multimodal Models (LMMs) like GPT-4o, LLaVA, and Gemini [23, 30] may address the above limitations in social behavior analysis. These models excel at interpreting complex verbal and non-verbal behaviors and analyze text, video, and audio data to produce detailed behavioral descriptions [18, 23, 29, 30]. Unlike non-GenAI models, LMMs rapidly adapt to new contexts with minimal data, due to advanced in-context learning capabilities, and can modify behaviors "on the fly" after deployment [15, 44]. This adaptability enhances flexibility in social behavior analysis, and the models' text-based outputs can provide inspirational insights. However, challenges persist, particularly in multi-modal information-seeking [14], such as accurately interpreting domain-specific social cues and addressing algorithmic biases [3, 32]. This highlights a clear need for tools that support domain experts in customizing AI analyses to document social behaviors in videos more effectively.

Our demo aims to facilitate collaboration among experts, supported by AI, in analyzing human behaviors in videos across contexts that heavily depend on domain knowledge. This will be particularly beneficial to CSCW researchers and practitioners looking to integrate multimodal data with AI to extract deeper behavioral insights in diverse fields [11, 12, 22, 24, 31, 41, 45, 46]. Additionally, it encourages reflection on the responsible use of AI, particularly addressing concerns like algorithmic bias and value alignment.

2 SOAP CONCEPT AND DESIGN OVERVIEW

SOAP.AI is a collaborative tool that enables experts to work with AI in documenting human behaviors using multimodal AI. "SOAP" is a documentation method widely used by the education domain [10]. The design was informed by 17 interviews with special education experts who routinely analyze and document human behaviors across group sessions of coaching and therapy. We identified a core concern raised by experts: AI's *fluidity.* Fluidity in AI entails dynamically adapting task definitions and evaluations to diverse contexts and evolving user goals. It involves customizing assessments for specific environments, understanding context-dependent task definitions, and creating new performance benchmarks based on expert standards. For instance, AI should recognize that rocking behavior in autistic children often represents sensory needs rather than disengagement.

To address the fluidity concern, a key design of SOAP.AI is its support for developing end-user ownership in human-AI collaborative documentation. Recent HCI research has explored the influence of ownership during human-AI collaboration, particularly in collaborative writing tasks [16, 25, 33]. Ownership—of both tangible objects and intangible entities such as ideas—often develops through acts of creation and control [16]. It is formed via three major paths, control, self-investment, and developing an intimate knowledge of the owner's target [13]. Although a sense of ownership has been shown to enhance collaboration outcomes among team members [27, 38, 43], its impact on human-AI collaboration remains under-examined, which motivates us to adopt it in our design context.

2.1 Interface

Figure 1 illustrates how SOAP.AI operates, with a step-by-step guide as follows: In the Video Panel: (A) Users upload a video for analysis. In the Chatbot Panel: (B) The AI detects individuals in the video automatically. Users can first name these detected individuals by interacting with the chatbot. Further, users interact with the chatbot to query about the video content, receiving insights generated from analyzed human behaviors. The Behavior Collaboration Panel (C-H), inspired by the concept of self-investment in ownership theory, involves: (C) Users select predefined behaviors from the global behavior set. (D) Selected or newly added behaviors are incorporated into the user's local behavior set. (E-H) Users customize behaviors in the behavior manager by: (E) Editing the behavior name, (F) Defining the behavior, (H) Marking behavior occurrences with timestamps, aided by AI-generated video descriptions. Once customization is complete, users can test the AI's understanding by querying the chatbot, or they can delete, save, and share these local behaviors with the global set to enhance collaboration. Additional features include: (I) Interactions are supported by both transcripts and visual graphs, enhancing user engagement and understanding. (J) Based on conversation history, the AI displays contextual information, such as noting a child's abnormal behavior during a therapy session due to rainy weather, which informs the behavior analysis notes generation.

2.2 Implementation

We implemented Soap as a web-based tool shown in figure 2, consisting of (A) a customizable front-end interface with a chatbot and feature visualizations, (B) a back-end server for video processing and AI models, and (C) a programmatic framework for analyzing and querying videos using GPT-4v. Users upload their videos, from which OpenAI Whisper extracts transcripts and timestamps. A Python script extracts video frames at a rate of 2 frames per second, and these frames, along with the transcripts, are fed into GPT-4v for analysis. Moreover, we automatically perform prompt engineering from expert conversations with the chatbot to provide context-sensitive insights and refine the model's outputs. The analyzed data is stored in an SQL database, accessible through the chatbot, supporting further visual analysis and report generation.

3 PILOT, LIMITATION, AND FUTURE WORK

We engaged eight special education experts for pilot sessions with SOAP.AI where they uploaded videos and collaborated with the AI on documentation. The feedback was generally positive, with experts impressed by the AI's capabilities and eager to co-create documents, aligning with our design goals. However, concerns arose about the AI's output.

One major issue identified was the inaccurate automatic speech recognition (ASR) in scenarios involving children's utterances, or people using non-speaking devices like iPads for communication. This exposes a significant inclusivity limitation in our current design, emphasizing the need for improvements to better meet the diverse user needs. Additionally, the AI sometimes struggled to

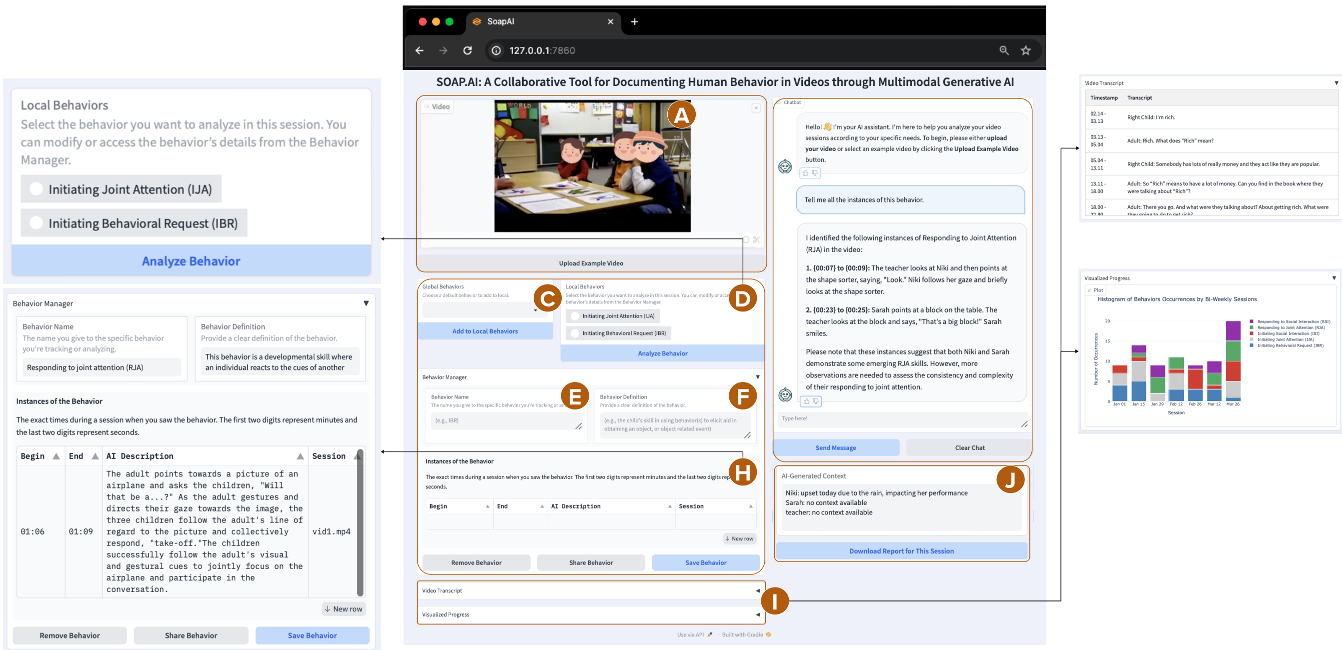

Figure 1: The core design of Soap.AI: the "control" path, where users select behaviors from collections and direct queries to the AI (C-D); the "self-invest" path, allowing users to input self-defined behavioral tasks for refined AI analysis (E-H); and the "knowledge" path, where the AI displays shared context (J), enabling users to verify its alignment with their expectations.

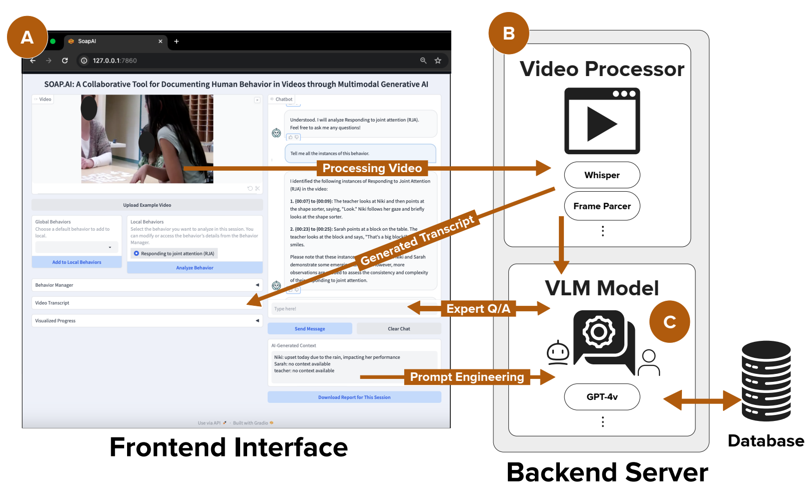

Figure 2: Architecture of Soap; (A) a customizable front-end interface with a chatbot and feature visualizations, (B) a back-end server for video processing and AI models, and (C) a programmatic framework for analyzing and querying videos using GPT-4v.

capture pauses in interactions, crucial for accurate behavioral analysis. Our team is actively refining this capability to ensure more accurate interaction assessments. Looking ahead, future versions of SOAP.AI will include features that enable experts to systematically evaluate and provide feedback on AI-generated prompts. These enhancements aim to ensure the tool evolves to effectively meet end user needs, reinforcing the commitment to making SOAP.AI a more inclusive and effective resource for professionals.

4 ACKNOWLEDGEMENT

This material is based upon work supported under the AI Research Institutes program by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award 2229873 - AI Institute for Transforming Education for Children with Speech and Language Processing Challenges. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

REFERENCES

[1] Abdulghafor, R., Abdelmohsen, A., Turaev, S., Ali, M. A., and Wani, S. An analysis of body language of patients using artificial intelligence. In *Healthcare* (2022), vol. 10, MDPI, p. 2504.

[2] Aggarwal, J. K., and Ryoo, M. S. Human activity analysis: A review. *Acm Computing Surveys (Csur) 43*, 3 (2011), 1–43.

[3] Arakawa, R., Maeda, K., and Yakura, H. Supporting experts with a multimodal machine-learning-based tool for human behavior analysis of conversational videos. *arXiv preprint arXiv:2402.11145* (2024).

[4] Arakawa, R., and Yakura, H. Rescue: A framework for real-time feedback on behavioral cues using multimodal anomaly detection. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–13.

[5] Association, A. S. L. H., et al. American speech-language-hearing association (asha).

[6] Bedmutha, M. S., Tsedenbal, A., Tobar, K., Borsotto, S., Sladek, K. R., Singh, D., Casanova-Perez, R., Bascom, E., Wood, B., Sabin, J., et al. Conversense: An automated approach to assess patient-provider interactions using social signals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–22.

[7] Beyan, C., Bustreo, M., Shahid, M., Bailo, G. L., Carissimi, N., and Del Bue, A. Analysis of face-touching behavior in large scale social interaction dataset. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (2020), pp. 24–32.

[8] Beyan, C., Vinciarelli, A., and Bue, A. D. Co-located human–human interaction analysis using nonverbal cues: A survey. *ACM Computing Surveys 56*, 5 (2023), 1–41.

[9] Binns, A. V., Cunningham, B. J., Andres, A., and Oram Cardy, J. Current practices, supports, and challenges in speech-language pathology service provision for autistic preschoolers. *Autism & Developmental Language Impairments 7* (2022), 23969415221120768.

[10] Cameron, S., and Turtle-Song, I. Learning to write case notes using the soap format. *Journal of Counseling & Development 80*, 3 (2002), 286–292.

[11] Chen, C., Arakawa, Y., Watanabe, K., and Ishimaru, S. Quantitative evaluation system for online meetings based on multimodal microbehavior analysis. *Sensors & Materials 34* (2022).

[12] Chen, X., Li, S., Liu, S., Fowler, R., and Wang, X. Meetscript: Designing transcript-based interactions to support active participation in group video meetings. *Proceedings of the ACM on Human-Computer Interaction 7*, CSCW2 (2023), 1–32.

[13] Dawkins, S., Tian, A. W., Newman, A., and Martin, A. Psychological ownership: A review and research agenda. *Journal of Organizational Behavior 38*, 2 (2017), 163–183.

[14] Deldjoo, Y., Trippas, J. R., and Zamani, H. Towards multi-modal conversational information seeking. In *Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval* (2021), pp. 1577–1587.

[15] Dos Santos Melicio, B. C., Xiang, L., Dillon, E., Soorya, L., Chetouani, M., Sarkany, A., Kun, P., Fenech, K., and Lorincz, A. Composite ai for behavior analysis in social interactions. In *Companion Publication of the 25th International Conference on Multimodal Interaction* (2023), pp. 389–397.

[16] Draxler, F., Werner, A., Lehmann, F., Hoppe, M., Schmidt, A., Buschek, D., and Welsch, R. The ai ghostwriter effect: When users do not perceive ownership of ai-generated text but self-declare as authors. *ACM Transactions on Computer-Human Interaction 31*, 2 (2024), 1–40.

[17] Feese, S., Arnrich, B., Tröster, G., Meyer, B., and Jonas, K. Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion. In *2012 International conference on privacy, security, risk and trust and 2012 international confernece on social computing* (2012), IEEE, pp. 520–525.

[18] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., and Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion 91* (2023), 424–444.

[19] Gatica-Perez, D. Analyzing group interactions in conversations: a review. In *2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems* (2006), IEEE, pp. 41–46.

[20] Graebe, L. C., and Haines, K. B. Documentation and monitoring of patient progress. *Mosby's Review Questions for the Speech-Language Pathology PRAXIS Examination E-Book* (2009), 252.

[21] Hegde, M. N., and Kuyujian, K. *Clinical methods and practicum in speech-language pathology.* Plural Publishing, 2019.

[22] Jackson, M., Anderson, A. H., McEwan, R., and Mullin, J. Impact of video frame rate on communicative behaviour in two and four party groups. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (2000), pp. 11–20.

[23] Jain, J., Yang, J., and Shi, H. VCoder: Versatile Vision Encoders for Multimodal Large Language Models. In *IEEE Conference on Computer Vision and Pattern Recognition* (2024).

[24] Jin, Q., Sodhi, I., Chen, A., and Yarosh, S. Interaction forms of collaborative vr video learning: An exploratory study.

[25] Joshi, N., and Vogel, D. Writing with ai lowers psychological ownership, but longer prompts can help. *arXiv preprint arXiv:2404.03108* (2024).

[26] Kotilainen, N., and Takala, M. "so much invisible work"–the role of special education teachers in finnish lower secondary schools. *Scandinavian Journal of Educational Research* (2024), 1–15.

[27] Lee, H., Yang, S.-B., and Koo, C. Exploring the effect of airbnb hosts' attachment and psychological ownership in the sharing economy. *Tourism Management 70* (2019), 284–294.

[28] Lehmann-Willenbrock, N., and Hung, H. A multimodal social signal processing approach to team interactions. *Organizational Research Methods* (2023), 10944281231202741.

[29] Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J., et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision 16*, 1-2 (2024), 1–214.

[30] Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *IEEE Conference on Computer Vision and Pattern Recognition* (2024).

[31] Lu, X., Chen, Y., and Epstein, D. A. A model of socially sustained self-tracking for food and diet. *Proceedings of the ACM on Human-Computer Interaction 5* (10 2021), 451.

[32] Maaz, M., Rasheed, H., Khan, S., and Khan, F. S. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* (2023).

[33] Mieczkowski, H. N. *AI-Mediated Communication: Examining Agency, Ownership, Expertise, and Roles of AI Systems.* Stanford University, 2022.

[34] Moore, B. J. Documentation issues. *Professional Issues in Speech-Language Pathology and Audiology* (2019), 401.

[35] Paloniemi, A., Pulkkinen, J., Kärnä, E., and Björn, P. M. The work of special education teachers in the tiered support system: The finnish case. *Scandinavian Journal of Educational Research 67*, 1 (2023), 35–50.

[36] Patel, R. A., Hartzler, A., Czerwinski, M. P., Pratt, W., Back, A. L., and Roseway, A. Leveraging visual feedback from social signal processing to enhance clinicians' nonverbal skills. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 2013, pp. 421–426.

[37] Samrose, S., McDuff, D., Sim, R., Suh, J., Rowan, K., Hernandez, J., Rintel, S., Moynihan, K., and Czerwinski, M. Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–13.

[38] Seo, J. Motives and role of psychological ownership in ar workspaces for remote collaboration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–5.

[39] Shih, J. Y., Mohanty, V., Katsis, Y., and Subramonyam, H. Leveraging large language models to enhance domain expert inclusion in data science workflows. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–11.

[40] Staltari, C. F., Baft-Neff, A., Marra, L. J., and Rentschler, G. J. Supervision: formative feedback for clinical documentation in a university speech-language pathology program. *Perspectives on Administration and Supervision 20*, 3 (2010), 117–123.

[41] Villagrán, I. Exploring the effects of applying learning analytics for teaching procedural skills in health sciences education. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (2021), pp. 299–302.

[42] Vinciarelli, A., Pantic, M., and Bourlard, H. Social signal processing: Survey of an emerging domain. *Image and vision computing 27*, 12 (2009), 1743–1759.

[43] Wang, Q., Battocchi, A., Graziola, I., Pianesi, F., Tomasini, D., Zancanaro, M., and Nass, C. The role of psychological ownership and ownership markers in collaborative working environment. In *Proceedings of the 8th international conference on Multimodal interfaces* (2006), pp. 225–232.

[44] Whitehead, R., Nguyen, A., and Järvelä, S. The generative multimodal analysis (gma) methodology for studying socially shared regulation in collaborative learning. In *The International Conference on Learning Analytics & Knowledge (LAK24)* (2024).

[45] Yoon, C. D., Terol, A. K., Meadan, H., and Lee, J. D. Gaze behaviors and social communication skills of young autistic children: A scoping review. *Review Journal of Autism and Developmental Disorders* (2024), 1–15.

[46] Zheng, Q., Xu, S., Wang, L., Tang, Y., Salvi, R. C., Freeman, G., and Huang, Y. Understanding safety risks and safety design in social vr environments. *Proceedings of the ACM on Human-Computer Interaction 7*, CSCW1 (2023), 1–37.