This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in: Sara-Jane Bittner, Robert A. Leist, László Kopácsi, Omair S. Bhatti, Abdulrahman Mohamed Selim, Michael Barz, Daniel Sonntag: Interactive Multimodal Photobook Co-Creation in Virtual Reality. Companion Proceedings of the 30th International Conference on Intelligent User Interfaces, Association for Computing Machinery 2025. https://doi.org/10.1145/3708557.3716355

Interactive Multimodal Photobook Co-Creation in Virtual Reality

Sara-Jane Bittner*
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Oldenburg, Germany
sara-jane.bittner@dfki.de

Robert Andreas Leist*
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
robert.leist@dfki.de

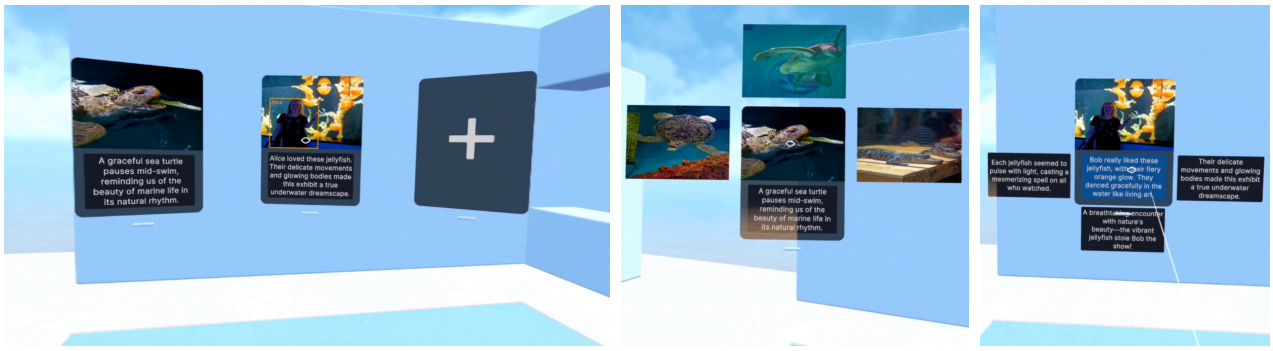
László Kopácsi*
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
laszlo.kopacsi@dfki.de

Omair Shahzad Bhatti
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
omair_shahzad.bhatti@dfki.de

Abdulrahman Mohamed Selim
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
abdulrahman.mohamed@dfki.de

Michael Barz
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
Applied Artificial Intelligence
University of Oldenburg
Oldenburg, Germany
michael.barz@dfki.de

Daniel Sonntag
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
Applied Artificial Intelligence
University of Oldenburg
Oldenburg, Germany
daniel.sonntag@dfki.de



(a) Page overview **(b) Image selection** **(c) Caption selection**

Figure 1: Screenshots of our VR setup showing the chapter overview in 1a, and the actions for image selection in 1b and caption selection in 1c triggered upon selection using gaze or controller input.

Abstract

The integration of Multimodal-Multisensor Interface (MMI) technologies into Virtual Reality (VR) enables users to engage with computational systems in a natural and immersive way. However,

*These authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
IUI Companion '25, March 24–27, 2025, Cagliari, Italy

© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1409-2/2025/03
https://doi.org/10.1145/3708557.3716355

these technologies remain underexplored when applied to deep learning (DL) systems in VR. This paper introduces a VR-based system designed to evaluate how users interact with DL models in virtual environments using MMI technologies, demonstrated through a photobook co-creation use case. The system facilitates human-AI collaboration (co-creation) by allowing users to work with DL models to create photobooks and supports incremental model learning based on user behaviour (Interactive DL) to produce personalised outputs. The tool features a Unity VR frontend that incorporates speech, gaze, and controller inputs. It has a modular backend architecture that allows seamless integration and testing of different DL models. This tool serves as a testbed for exploring MMI in immersive VR environments for both IDL and co-creation.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **Interaction techniques**; **Virtual reality**; • **Computing methodologies** → *Machine learning*.

Keywords

Virtual Reality, Interaction Design, Co-Creation, Interactive Deep Learning, Photobook Creation

ACM Reference Format:

Sara-Jane Bittner, Robert Andreas Leist, László Kopácsi, Omair Shahzad Bhatti, Abdulrahman Mohamed Selim, Michael Barz, and Daniel Sonntag. 2025. Interactive Multimodal Photobook Co-Creation in Virtual Reality. In *30th International Conference on Intelligent User Interfaces Companion (IUI Companion '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3708557.3716355

1 Introduction & Related Work

Multimodal-Multisensor Interfaces (MMI) leverage human senses and behaviour observations, such as speech, gestures, eye movements, and touch, in user-system interaction [35, 46]. By integrating diverse input modalities, MMI can increase flexibility [49], user satisfaction, and task efficiency when interacting with a model [22]. In the context of Virtual Reality (VR), where increasingly realistic, immersive experiences are prioritised [13], researchers have explored a variety of multimodal techniques in virtual environments [27, 30, 34, 37]. For example, combining gaze and speech inputs can simplify virtual interfaces by creating shortcuts (e.g., issuing a spoken instruction rather than navigating through menus), thereby reducing interface complexity and improving overall usability. Such natural interaction methods offer many advantages in VR. Speech-based interfaces, for instance, lower the barriers to expressing intent [3] and enable flexible, natural language interactions [14]. Similarly, eye tracking offers hands-free, intuitive interactions by allowing users to engage with digital elements simply through gaze [16, 25, 28, 44]. The combination of immersiveness and advanced interaction techniques in VR presents unique opportunities for enhancing Interactive Deep Learning (IDL) and human-AI co-creation systems.

IDL systems mainly focus on incremental model updates driven by user feedback, actively involving users in the machine learning process [1]. Conversely, co-creation emphasises continuous, iterative collaboration between humans and Artificial Intelligence (AI), where users provide intuition, guidance, and adaptability, while AI generates diverse artefacts and explores possibilities [32, 38, 45, 48]. Integrating VR's immersive and natural interaction methods into IDL and co-creation processes has the potential to enhance the engagement of human-AI partnerships. However, IDL remains largely confined to 2D interfaces and conventional input methods such as mouse and keyboard [47].

In this paper, we introduce a system that supports speech, gaze, and controller-based interaction to enable multimodal IDL and co-creation in VR. The system is demonstrated through a photobook creation scenario, which serves as an ideal testbed due to its inclusion of various DL-related tasks such as image retrieval and selection, and caption generation [40, 42, 43]. Prior research has explored related tasks, including methods for user-guided image selection [31], iterative feedback to generate personalised captions [2], and multimodal approaches for identifying individuals within photobook collections [23]. However, existing tools like photobook.ai[1] primarily focus on 2D environments. While these tools provide automation, they often lack the flexibility and interactive depth required for effective co-creation [7, 19, 26]. By contrast, 3D environments in VR have demonstrated significant potential to enhance engagement and creativity. For example, transitioning 2D images into 3D exhibitions has been shown to increase audience interaction and immersion [10], while dynamic presentation methods, such as augmented informational text, have improved comprehension in virtual spaces [39]. Building on these insights, this work introduces a novel DL-supported co-creation system designed to personalise photobook creation based on real-time human feedback. By leveraging the immersive and natural interaction capabilities of VR, our system facilitates richer collaboration between users and AI, providing an intuitive, multimodal framework for creative and iterative tasks. Our future goal is to explore novel MMI and fusion methods for engaging with DL in VR while also investigating the human factors critical to fostering effective and meaningful human-AI collaboration in such immersive setups.

0

2 Photobook Creation System

We present a VR-based photobook creation system that leverages MMI—using speech, gaze, and controller inputs—to enable collaborative, human-in-the-loop generation of personalised photobooks with DL-supported features like chapter creation, image retrieval, and contextualised captioning.

2.1 Architecture

The system employs a modular architecture, as shown in Fig. 2, comprising a VR frontend and a Python-based DL backend. The frontend, developed with Unity 2022.3.37f1[2], operates on the Vive XR Elite headset[3]. Gaze and controller-based interactions are implemented using the XR Interaction Toolkit v3.0.3[4], while speech

[1] https://photobook.ai/

[2] https://unity.com/releases/editor/whats-new/2022.3.37

[3] https://www.vive.com/de/product/vive-xr-elite/overview/ (Accessed January 07, 2024)

[4] https://docs.unity3d.com/Packages/com.unity.xr.interaction.toolkit@3.0/manual/index.html (Accessed January 07, 2024)

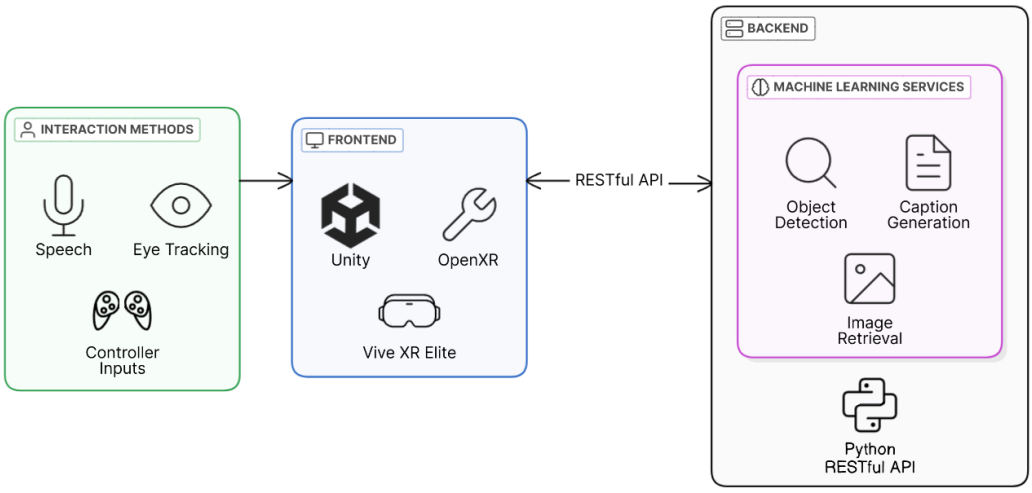


Figure 2: Overview of the architecture. The user can interact with the frontend via multiple modalities. The frontend and backend communicate to enable the co-creation process.

recognition is supported by the huggingface library[5]. These modalities are combined to enable seamless multimodal interactions within the virtual environment. The VR frontend communicates with the backend via a RESTful interface managed by a Python-based server. Session management and data persistence are handled within the backend layer. Computationally intensive tasks, such as object detection and caption generation, are delegated to specialised machine learning services connected through gRPC[6]. These services are modular, allowing for replacement or updates to models without affecting the user interface. Furthermore, they can be distributed and executed locally or in the cloud. By isolating machine learning components, the architecture supports incremental updates and ensures system stability without requiring major reconfigurations.

2.2 VR Frontend

The VR frontend facilitates co-creation through two phases: *initialisation* and *creation*. During the initialisation phase, users provide images and voice commands to generate photobook content. In the creation phase, users iteratively refine outputs—such as chapters, images, and captions—using multimodal inputs (i.e., gaze, speech, and controller interactions) to align the photobook content with their expectations dynamically.

The VR interface comprises three main stages: *Photobook Overview*, *Chapter Overview*, and *Page Overview*. In each stage, content is displayed as interactive floating cards arranged circularly around the user, as shown in Fig. 1a. In the Photobook Overview, users can view existing photobooks or create new ones using voice commands or by selecting an empty card marked with a plus sign via gaze or controller input. Once a photobook is selected, the system transitions to the Chapter Overview, where users can create or revise chapters by providing descriptions and specifying the number of pages. Selecting a chapter advances the system to the Page Overview, as shown in Fig. 1a, where users interact with image-caption cards, each representing chapter pages. Users can open a radial menu to explore replacement options for images or captions through MMI techniques, as shown in Fig. 1b and Fig. 1c, respectively.

The VR system supports various MMI techniques, detailed in Table 1, which outline how gaze, speech, and controller inputs can be used independently or in combination. Table 2 highlights specific actions enabled by these techniques, such as object highlighting and selection, image and caption replacement, and content modification or addition. These different interaction techniques allow us to explore more natural and intuitive ways for users to interact with IDL within an immersive VR environment, laying the groundwork for future studies on human factors. For example, the backend detects instances within images, highlighting them with bounding boxes when the user dwells on an image. Users can correct mislabelled instances, such as changing "Bob" to "Alice," using voice commands. These corrections are processed by the backend to update bounding box labels and regenerate captions accordingly.

2.3 Evaluation Plan

The proposed personalised photobook creation scenario will serve as a testbed for studying MMI and human-AI co-creation in immersive environments. We plan to conduct a set of empirical user studies: (i) assessing the usability and effectiveness of different MMI techniques (as shown in Table 1); and (ii) comparing the VR interface with a conventional 2D interface (i.e., utilising mouse and keyboard). In both studies complementary objective and subjective data will be measured: The objective data collection will include metrics to evaluate efficiency and effectiveness (e.g., completion time and error rate) and, implicit human behaviour (e.g., interaction logs and multimodal data like eye and facial tracking [24, 33]). Subjective data will be measured via self-reported questionnaires and

[5]https://huggingface.co/ (Accessed January 07, 2024)
[6]https://grpc.io/ (Accessed January 07, 2024)

Table 1: Interaction techniques categorized by their Primary and Secondary modalities.

Secondary \ Primary	Gaze	Speech	Controller
Gaze	gaze-based pointing and selection	contextual voice commands	gaze-aware controller-based selection
Speech	gaze-based selection + voice commands	voice commands	controller-based selection + voice commands
Controller	gaze-based pointing + controller-based selection	contextual voice commands	controller-based pointing and selection

Table 2: Actions supported by the VR frontend.

Action	Gaze	Speech	Controller
Pointing and Selecting	● □		● □
Image/Caption Replacement	○ ● □	○ △	△ ● □
Add new image/caption	○ ● □	○ ● △	△ ● □
Add new photobook/chapter	○	○ △	△
Modify existing content	○	○ △	△

Legend:
●: Unimodal (single modality) interaction
○: Multimodal interaction involving **Gaze + Speech**
△: Multimodal interaction involving **Speech + Controller**
□: Multimodal interaction involving **Gaze + Controller**

qualitative methods. Self-reported questionnairs will include human factors which have a significant impact on the successful use of user interfaces and tools,such as usability (e.g., the System Usability Scale [9]) and mental load (e.g., NASA-TLX [21]), and factors relevant to AI-supported co-creation, such as creativity support (e.g., Creativity Support Index [11]) and user experience (e.g., User Experience Questionnaire [36, 41]). In comparison, qualitative methods will provide insights to ground the objective findings with actionable implications from a human perspective (e.g., semi-structured interview [6], think-aloud protocol [36] and the reflective thematic analysis as formalisation method [8]). We aim to investigate human factors critical to fostering effective and meaningful human-AI collaboration in immersive setups.

2.4 Limitations

While the system includes basic error-handling mechanisms, such as allowing users to cancel voice commands in case of transcription errors and enabling corrections to images, captions, and instance labels (e.g., updating bounding boxes and regenerating captions), it lacks more advanced features like undo and redo functionality. Such capabilities could improve user control and flexibility during iterative co-creation tasks. Additionally, the system currently supports only image-based content. Although this aligns with existing tools like photobook.ai[7], VR environments offer the potential for integrating alternative media types, such as videos or RGBD, wide-angle, or 360°content, which could leverage the spatial affordances of immersive 3D spaces to potentially enhance user engagement and creativity [20]. The system also has limitations when onboarding new users. While it supports intuitive interactions through multimodal inputs, a guided "getting started" phase—such as written/oral instructions supported by visual aids or an avatar-based assistant [12]—could help users familiarize themselves with the system more effectively. Furthermore, adaptive help based on detecting the user's state, e.g. user disagreement with the system output [4, 5] could address user challenges dynamically.

In terms of accessibility, the system provides features that may benefit users with auditory or motor impairments (e.g. gaze-based interaction and speech input [17, 18]). However, it is not well-suited for users with visual impairments. While haptic feedback via controllers and adjustable diopters in the headset may assist users with corrected vision, VR systems, in general, are not optimized for visually impaired users [15, 29]. Addressing these accessibility gaps would require further research and design considerations.

[7]https://photobook.ai/

3 Conclusion

This paper introduced a VR-based photobook co-creation tool that combines MMI techniques with DL models to enable intuitive and immersive human-AI collaboration. The system supports iterative human-in-the-loop workflows for tasks such as image retrieval, caption generation, and content refinement using different unimodal and multimodal combinations of speech, gaze, and controller inputs. The modular backend architecture allows seamless integration and testing of different DL models while maintaining system functionality. This setup enables natural and intuitive user engagement with IDL and co-creation, offering a rich testbed for exploring MMI in immersive VR environments in future user studies. Future work will involve conducting user studies to assess different factors of MMI in immersive VR environments, integrating error-handling methods and alternative media types.

Acknowledgments

This work was funded by the European Union under grant number 101093079 (MASTER) and the German Federal Ministry of Education and Research (BMBF) under grant numbers 01IW23002 (No-IDLE) and 01IW24006 (NoIDLEChatGPT), as well as by the Endowed Chair of Applied AI at the University of Oldenburg.

References

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (Dec. 2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513 Number: 4.

[2] Aliki Anagnostopoulou, Mareike Hartmann, and Daniel Sonntag. 2023. Putting Humans in the Image Captioning Loop. arXiv:2306.03476 [cs.CL] https://arxiv.org/abs/2306.03476

[3] Jillian Aurisanoevl, Abhinav Kumarnl, Ava Khairi, Redal Ava, Jason Leighl, Barbara Dieugenionl, Andrew Johnsonevl, and Alberto Gonzalez Martinez. 2015. "Show me data." Observational study of a conversational interface in visual data exploration. https://doi.org/10.13140/RG.2.2.34456.90889

[4] Omair Bhatti, Michael Barz, and Daniel Sonntag. 2022. Leveraging Implicit Gaze-Based User Feedback for Interactive Machine Learning. In *KI 2022: Advances in Artificial Intelligence*, Ralph Bergmann, Lukas Malburg, Stephanie C. Rodermund, and Ingo J. Timm (Eds.). Springer International Publishing, Cham, 9–16.

[5] Omair Shahzad Bhatti, Harshinee Sriram, Abdulrahman Mohamed Selim, Cristina Conati, Michael Barz, and Daniel Sonntag. 2024. Detecting when Users Disagree with Generated Captions. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI Companion '24)*. Association for Computing Machinery, New York, NY, USA, 195–203. https://doi.org/10.1145/3686215.3688382

[6] Ann E. Blandford. 2013. Semi-structured qualitative studies. In *The Encyclopedia of Human-Computer Interaction*. Interaction Design Foundation, Denmark. https://discovery.ucl.ac.uk/id/eprint/1436174/

[7] Oliver Bown and Andrew R Brown. 2018. Interaction design for metacreative systems. *New Directions in Third Wave Human-Computer Interaction: Volume 1-Technologies* 1 (2018), 67–87.

[8] Virginia Braun, Victoria Clarke, Nikki Hayfield, Louise Davey, and Elizabeth Jenkinson. 2022. Doing Reflexive Thematic Analysis. In *Supporting Research in Counselling and Psychotherapy : Qualitative, Quantitative, and Mixed Methods Research*, Sofie Bager-Charleson and Alistair McBeath (Eds.). Springer International Publishing, Cham, 19–38. https://doi.org/10.1007/978-3-031-13942-0_2

[9] John Brooke. 1986. System usability scale (SUS): a quick-and-dirty method of system evaluation user information. *Reading, UK: Digital equipment co ltd* 43 (1986), 1–7.

[10] Jiaxun Cao, Qingyang He, Zhuo Wang, RAY LC, and Xin Tong. 2023. DreamVR: Curating an Interactive Exhibition in Social VR Through an Autobiographical Design Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 511, 18 pages. https://doi.org/10.1145/3544548.3581362

[11] Erin A. Carroll and Celine Latulipe. 2009. The creativity support index. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. Association for Computing Machinery, New York, NY, USA, 4009–4014. https://doi.org/10.1145/1520340.1520609

[12] Edwige Chauvergne, Martin Hachet, and Arnaud Prouzeau. 2023. User Onboarding in Virtual Reality: An Investigation of Current Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 711, 15 pages. https://doi.org/10.1145/3544548.3581211

[13] Maurice Chu and Bo Begole. 2010. Chapter 17 - Natural and Implicit Information-Seeking Cues in Responsive Technology. In *Human-Centric Interfaces for Ambient Intelligence*, Hamid Aghajan, Ramón López-Cózar Delgado, and Juan Carlos Augusto (Eds.). Academic Press, Oxford, 415–452. https://doi.org/10.1016/B978-0-12-374708-2.00017-6

[14] Kenneth Cox, Rebecca E Grinter, Stacie L Hibino, Lalita Jategaonkar Jagadeesan, and David Mantilla. 2001. A multi-modal natural language interface to an information visualization environment. *International Journal of Speech Technology* 4 (2001), 297–314.

[15] Chris Creed, Maadh Al-Kalbani, Arthur Theil, Sayan Sarcar, and Ian Williams. 2024. Inclusive AR/VR: accessibility barriers for immersive technologies. *Universal Access in the Information Society* 23, 1 (March 2024), 59–73. https://doi.org/10.1007/s10209-023-00969-0

[16] Andrew T. Duchowski. 2018. Gaze-based interaction: A 30 year retrospective. *Computers & Graphics* 73 (June 2018), 59–69. https://doi.org/10.1016/j.cag.2018.04.002

[17] John Dudley, Lulu Yin, Vanja Garaj, and Per Ola Kristensson. 2023. Inclusive Immersion: a review of efforts to improve accessibility in virtual reality, augmented reality and the metaverse. *Virtual Reality* 27, 4 (Dec. 2023), 2989–3020. https://doi.org/10.1007/s10055-023-00850-8

[18] Daniel Gepner, Jérôme Simonin, and Noëlle Carbonell. 2007. Gaze as a Supplementary Modality for Interacting with Ambient Intelligence Environments. In *Universal Access in Human-Computer Interaction. Ambient Interaction*, Constantine Stephanidis (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 848–857.

[19] Frederic Gmeiner, Kenneth Holstein, and Nikolas Martelaro. 2022. Team Learning as a Lens for Designing Human-AI Co-Creative Systems. https://doi.org/10.48550/arXiv.2207.02996 arXiv:2207.02996 [cs.HC]

[20] Laura Josephine Hagedorn, Alwin De Rooij, and Maryam Alimardani. 2023. Virtual Reality and Creativity: How do Immersive Environments Stimulate the Brain during Creative Idea Generation?. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 308, 7 pages. https://doi.org/10.1145/3544549.3585848

[21] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (Oct. 2006), 904–908. https://doi.org/10.1177/154193120605000909

[22] Ziyao He, Shiyuan Li, Yunpeng Song, and Zhongmin Cai. 2024. Towards Building Condition-Based Cross-Modality Intention-Aware Human-AI Cooperation under VR Environment. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 955, 13 pages. https://doi.org/10.1145/3613904.3642360

[23] Niels Henze and Susanne Boll. 2011. Who's that girl? handheld augmented reality for printed photo books. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction - Volume Part III* (Lisbon, Portugal) *(INTERACT'11)*. Springer-Verlag, Berlin, Heidelberg, 134–151.

[24] M. Sazzad Hussain, Rafael A. Calvo, and Fang Chen. 2013. Automatic Cognitive Load Detection from Face, Physiology, Task Performance and Fusion During Affective Interference. *Interacting with Computers* 26, 3 (06 2013), 256–268. https://doi.org/10.1093/iwc/iwt032 arXiv:https://academic.oup.com/iwc/article-pdf/26/3/256/2054117/iwt032.pdf

[25] Robert J. K. Jacob. 1990. What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90)*. Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/97243.97246

[26] Anna Kantosalo, Prashanth Thattai Ravikumar, Kazjon Grace, and Tapio Takala. 2020. Modalities, Styles and Strategies: An Interaction Framework for Human-Computer Co-Creativity.. In *ICCC*. International Conference on Computational Creativity, Online, 57–64.

[27] László Kopácsi, Albert Klimenko, Michael Barz, and Daniel Sonntag. 2024. Exploring Gaze-Based Menu Navigation in Virtual Environments. In *Proceedings of the 2024 ACM Symposium on Spatial User Interaction* (Trier, Germany) *(SUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 40, 2 pages. https://doi.org/10.1145/3677386.3688887

[28] László Kopácsi, Tobias Sebastian Schneider, Chiara Karr, Michael Barz, and Daniel Sonntag. 2024. GazeLock: Gaze- and Lock Pattern-Based Authentication. In *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology* (Trier, Germany) *(VRST '24)*. Association for Computing Machinery, New York, NY, USA, Article 94, 2 pages. https://doi.org/10.1145/3641825.3689520

[29] Tianhang Liu, Pooyan Fazli, and Heejin Jeong. 2024. Artificial Intelligence in Virtual Reality for Blind and Low Vision Individuals: Literature Review. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 68, 1 (2024), 1333–1338. https://doi.org/10.1177/10711813241266832 arXiv:https://doi.org/10.1177/10711813241266832

[30] Mathias N. Lystbæk, Peter Rosenberg, Ken Pfeuffer, Jens Emil Grønbæk, and Hans Gellersen. 2022. Gaze-Hand Alignment: Combining Eye Gaze and Mid-Air Pointing for Interacting with Menus in Augmented Reality. *Proceedings of the ACM on Human-Computer Interaction* 6, ETRA (May 2022), 1–18. https://doi.org/10.1145/3530886

[31] Matthias Maszuhn, Larbi Abdenebaoui, and Susanne Boll. 2021. A User-Centered Approach for Recognizing Convenience Images in Personal Photo Collections. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, Lille, France, 1–4. https://doi.org/10.1109/CBMI50038.2021.9461908

[32] Jack McGuire, David De Cremer, and Tim Van de Cruys. 2024. Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence. *Scientific Reports* 14, 1 (Aug. 2024), 18525. https://doi.org/10.1038/s41598-024-69423-2 Publisher: Nature Publishing Group.

[33] Abdulrahman Mohamed Selim, Michael Barz, Omair Shahzad Bhatti, Hasan Md Tusfiqur Alam, and Daniel Sonntag. 2024. A review of machine learning in scanpath analysis for passive gaze-based interaction. *Frontiers in Artificial Intelligence* 7 (2024), 1–28. https://doi.org/10.3389/frai.2024.1391745

[34] Aunnoy Mutasim, Anil Ufuk Batmaz, Moaaz Hudhud Mughrabi, and Wolfgang Stuerzlinger. 2022. Performance Analysis of Saccades for Primary and Confirmatory Target Selection. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology* (Tsukuba, Japan) *(VRST '22)*. Association for Computing Machinery, New York, NY, USA, Article 18, 12 pages. https://doi.org/10.1145/3562939.3565619

[35] Sharon Oviatt, Björn Schuller, Philip R. Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger (Eds.). 2017. *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations - Volume 1*. Vol. 14. Association for Computing Machinery and Morgan & Claypool. https://doi.org/10.1145/3015783

[36] Fabiana Frata Furlan Peres, Fatima Nunes, João Marcelo Teixeira, Claudio R. M. Maurício, Katiani P. Conceição, and Leina Yoshida. 2024. Methods for Evaluating Immersive 3D Virtual Environments: a Systematic Literature Review. In *Symposium on Virtual and Augmented Reality*. ACM, Manaus Brazil, 140–151. https://doi.org/10.1145/3691573.3691595

[37] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. 2017. Gaze + pinch interaction in virtual reality. In *Proceedings of the 5th Symposium on Spatial User Interaction* (Brighton, United Kingdom) *(SUI '17)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/3131277.3132180

[38] Jeba Rezwana and Mary Lou Maher. 2023. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Transactions on Computer-Human Interaction* 30, 5 (Oct. 2023), 1–28. https://doi.org/10.1145/3519026

[39] Rufat Rzayev, Polina Ugnivenko, Sarah Graf, Valentin Schwind, and Niels Henze. 2021. Reading in VR: The Effect of Text Presentation Type and Location. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 531, 10 pages. https://doi.org/10.1145/3411764.3445606

[40] Philipp Sandhaus, Sabine Thieme, and Susanne Boll. 2008. Processes of photo book production. *Multimedia Systems* 14 (2008), 351–357.

[41] Andrej Somrak, Matevž Pogačnik, and Jože Guna. 2021. Suitability and comparison of questionnaires assessing virtual reality-induced symptoms and effects and user experience in virtual environments. *Sensors* 21, 4 (2021), 1185. https://www.mdpi.com/1424-8220/21/4/1185 Publisher: MDPI.

[42] Daniel Sonntag, Michael Barz, and Thiago Gouvea. 2024. *A look under the hood of the Interactive Deep Learning Enterprise (No-IDLE).* Technical Report. German Research Center for AI.

[43] Daniel Sonntag, Thiago Gouvea, Michael Barz, Aliki Anagnostopoulou, Siting Liang, Sara-Jane Bittner, and Franziska Scheurer. 2024. *The Interactive Deep Learning Enterprise (No-IDLE) meets ChatGPT.* Technical Report. German Research Center for AI.

[44] India Starker and Richard A. Bolt. 1990. A gaze-responsive self-disclosing display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90).* Association for Computing Machinery, New York, NY, USA, 3–10. https://doi.org/10.1145/97243.97245

[45] Maurizio Talamo. 2023. The Digital Revolution and the Art of Co-creation. In *Technological Imagination in the Green and Digital Transition*, Eugenio Arbizzani, Eliana Cangelli, Carola Clemente, Fabrizio Cumo, Francesca Giofrè, Anna Maria Giovenale, Massimo Palme, and Spartaco Paris (Eds.). Springer International Publishing, Cham, 27–35. https://doi.org/10.1007/978-3-031-29515-7_4

[46] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern recognition letters* 36 (2014), 189–195.

[47] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24).* Association for Computing Machinery, New York, NY, USA, Article 185, 21 pages. https://doi.org/10.1145/3613904.3642803

[48] Zhuohao Wu, Danwen Ji, Kaiwen Yu, Xianxu Zeng, Dingming Wu, and Mohammad Shidujaman. 2021. AI Creativity and the Human-AI Co-creation Model. In *Human-Computer Interaction. Theory, Methods and Tools*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 171–190. https://doi.org/10.1007/978-3-030-78462-1_13

[49] Jackie (Junrui) Yang, Yingtian Shi, Yuhan Zhang, Karina Li, Daniel Wan Rosli, Anisha Jain, Shuning Zhang, Tianshi Li, James A. Landay, and Monica S. Lam. 2024. ReactGenie: A Development Framework for Complex Multimodal Interactions Using Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24).* Association for Computing Machinery, New York, NY, USA, Article 483, 23 pages. https://doi.org/10.1145/3613904.3642517